



NATIONAL OPEN UNIVERSITY OF NIGERIA

STATISTICS FORECONOMIST II

ECO 254

FACULTY OF SOCIAL SCIENCES

COURSE GUIDE

Course Developer:

Samuel Olumuyiwa Olusanya

Economics Department, National Open University of Nigeria

COURSE CONTENT EDITOR

Dr Segun Quadri Oyebo Hassan

Senior Research Fellow

Econometrics Consultant Research Firm

CONTENT

Introduction

Course Content

Course Aims

Course Objectives

Working through This Course

Course Materials

Study Units

Textbooks and References

Assignment File

Presentation Schedule

Assessment

Tutor-Marked Assignment (TMAs)

Final Examination and Grading

Course Marking Scheme

Course Overview

How to Get the Most from This Course

Tutors and Tutorials

Summary

Introduction

Welcome to ECO: 254STATISTICS FOR ECONOMIST II.

ECO 254: Statistics for Economist II is a three-credit and one-semester undergraduate course for Economics student. The course is made up of nineteen units spread across fifteen lectures weeks. This course guide gives you an insight to Statistics for economist II in a broader way and how to study the make use and apply statistical analysis in economics. It tells you about the course materials and how you can work your way through these materials. It suggests some general guidelines for the amount of time required of you oneach

unit in order to achieve the course aims and objectives successfully. Answers to your tutor marked assignments (TMAs) are therein already.

Course Content

This course is basically on Statistics for economist because as you are aspiring to become an economist, you must be able to apply statistical techniques to economics problems. The topics covered include probability distribution; hypotheses testing; sampling theory; t test, f test and chi-square analysis and simple regression analysis and its application.

Course Aims

The aims of this course is to give you in-depth understanding of the economics as regards

- Fundamental concept and calculation of probability distribution
- To familiarize students with the knowledge of hypotheses testing
- To stimulate student's knowledge on sampling theory
- To make the students to understand the statistical calculation of t test, f test and chi-square analysis.
- To expose the students to analysis of simple linear regression analysis
- To ensure that the students know how to apply simple linear regression to economics situations.
- To make the students to be to interpret simple linear regression analysis result.

Course Objectives

To achieve the aims of this course, there are overall objectives which the course is out to achieve though, there are set out objectives for each unit. The unit objectives are included at the beginning of a unit; you should read them before you start working through the unit. 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. This is to assist the students in accomplishing the tasks entailed in this course. In this way, you can be sure you have done

what was required of you by the unit. The objectives serves as study guides, such that student could know if he is able to grab the knowledge of each unit through the sets of objectives in each one. At the end of the course period, the students are expected to be able to:

- Understand the meaning of probability distribution and also to understand the application of probability as well as to understand different types of probability distribution and analyse probability problem.
- Teach students the meaning and application of continuous random variables such as continuous probability distribution and Distribution functions for Continuous Random Variables.
- Distinguish and explain the various higher probability distributions and their applications.
- Know the meaning of hypotheses testing and to understand type 1 and type 2 errors and able to calculate one-tailed and two-tailed tests. However, it is also to calculate various hypothesis testing and be able to apply test to economics problems.
- Differentiate between one-tailed and two-tailed tests and to also understand the procedures for carrying out test of Hypothesis.
- Know the procedures for carrying out the test of hypothesis and to know how to calculate statistical test for mean of a single population as well as to know how to calculate the interval estimation for mean of a single population.
- Understand the test of difference between two means of independent samples and confidence intervals for Difference between two means.
- Know how to calculate the mean of the population of different score of t statistic and variance of the difference of scores.
- Know the meaning of sampling and also to understand random samples as well as to understand random population parameters.
- Analyse the population parameter and to also understand sample statistics.

- Understand the meaning of sampling Error and to also understand sampling distribution as well as sampling distribution of means.
- Estimate the mean and variance from a population and also to calculate the sampling distribution of proportion as well as to calculate the sampling distribution of Differences and sums.
- Analyse the frequency distribution and frequency polygon and to know the relative frequency distributions as well as to understand the calculation of ungrouped data for frequency distributions.
- Understand the history of t statistics and to know the unpaired and paired two sample t tests as well as to understand the t test formula.
- Know the various examples of f tests Statistics and understand the formulae as well as analyse the f tests Statistics.
- Know the examples of chi-square distribution and to understand the application of chi-square analysis.
- Know the uses of regression analysis and to understand the three conceptualizations of regression analysis as well as understand regression models.
- Know the uses of simple linear regression and understand linear regression analysis.
- Understand what a simple regression analysis can deduce from its application and how to interpret the parameter using to test for t and f-tests respectively.

Working Through The Course

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

Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, you will be required to submit assignments for assessment purposes. At the end of the course there is a final examination. This course should take about 15weeks to complete and some components of the course are outlined under the course material subsection.

Course Material

The major component of the course, What you have to do and how you should allocate your time to each unit in order to complete the course successfully on time are listed follows:

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

Study Unit

There are 19 units in this course which should be studied carefully and diligently.

MODULE 1 PROBABILITY DISTRIBUTION

Unit 1 Analysis of Probability Distribution

Unit 2 Continuous Random Variables

Unit 3 Other Probability Distributions

MODULE 2 HYPOTHESES TESTING

Unit 1 Meaning of Hypothesis

Unit 2 The Criterion of Significance

Unit 3 Statistical Test for Hypothesis

Unit 4 Testing Differences between Two Means

Unit 5 Testing Difference between Matched Samples

MODULE 3 SAMPLING THEORY

Unit 1 Population and Sample

Unit 2 Population Parameters

Unit 3 Sampling Parameters

Unit 4 Calculation of Sampling Distribution and Estimators for Mean Variance

Unit 5 Frequency Distribution

MODULE 4 T-TEST, F-TEST AND CHI SQUARE ANALYSIS

Unit 1 T-test

Unit 2 F-test

Unit 3 Chi-Square Test

MODULE 5 SIMPLE LINEAR REGRESSION ANALYSIS AND ITS APPLICATION

Unit 1 Meaning of Regression Analysis

Unit 2 Simple Linear Regression Analysis

Unit 3 Application of Simple Linear Regression Analysis

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

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

Textbook and References

For further reading and more detailed information about the course, the following materials are recommended:

Adedayo, O.A (2000). *Understanding Statistics*, JAS Publisher Akoka, Lagos

Armstrong, J. S. (2012). Illusion in Regress Analysis, *International Journal of forecasting (forthcoming)* pg 28.

Ademisan, T.Y. (2011) *Introduction to Statistics, a contemporary issue*,

1st edition, Mill world Publication limited.

Chian, C. L. (2003). *Statistical Methods of Analysis*, World scientific,

ISBN pg 981 – 938.

Daniel, A.P., & Yu, X. (1999). *Statistical Methods for categorical data Analysis*, Academic Press, inc. 1999.

David, A. F. (2005). *Statistical models, theory and practice*, Cambridge University Press, 2005.

Datel, R.O. (2013). *Statistics for Business and Management Studies*, 2nd edition, Merrigon Press Company limited.

Faraday, D.F. (2009). *Statistics for Business Management*, 1st edition, Junit Press limited, Lagos.

Gago, C.C. (2009). *Statistics for Economist*, 1st edition DALT Publication limited.

- Samuelson, H. (2012). *Introduction to Statistical for Economics*, Mill world Publication limited, 2nd edition.
- Lee, I.G. (2008). *Working through the Statistics*, a broader approach, 1st edition, Leepy Publication limited.
- Murray, S., & John, S., & Alu, S. (2001). *Probability and Statics*, Schaum's easy outlines, Macgraw Hill Publication Company, New York.
- Micks, J.J. (1997). *Business statistics for Managers*, 2nd edition, Migrawhill Publishing Comapny Limited, Berkshire England.
- Nelson, L. (1984). *The Shewhart control chat, test for Special courses*, Journal of QualityTechnology, 16, 237 – 23.
- Nievergelt, Y. (1994). *Total Least Square*, State of the art Regression in Numerical Analysis, 1st edition.
- Ojo, J. B (2011). *Statistics made easy*, melting point publication, Lagos.
- Olomeko, O. A. (2012). *Regression Analysis*, 1st edition, pg 88 Millworld Publication Limited, Lagos.
- Richard, M. O. (2001). *The Analysis of Understanding Statistics*, 2nd Edition, Millworld Publication Limited.
- Ramsey, J.J. (2011). *Statistics and Probability theory*, 2nd edition, Mill world Press limited.
- Sawilowsky, S. (2005). Misconceptions Leading to Choosing the t-Test over the Wilcoxon Mann Whitney U Test for Shift in Location Parameter, *Journal of Modern Applied Statistical Method* 4(2) 598-600.
- Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, pg 267-288, Jstor edition.
- Wesley, H.F. (2010) *Statistics and Economics, a broader approach*, 1st edition, Queror publication limited.

Wemimo, A. (2007). *Undersatnding the Concept of Statistics*, a contemporary approach, 1st edition, Merrinlyn Press limited.

Assignment File

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

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

Assignment 1 - All TMAs' question in Units 1 – 5 (Module 1 and 2)

Assignment 2 - All TMAs' question in Units 6 – 11 (Module 2 and 3)

Assignment 3 - All TMAs' question in Units 12 – 15 (Module 3 and 4)

Assignment 4 - All TMAs' question in Unit 16 – 19 (Module 4 and 5).

Presentation Schedule

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

Assessment

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

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

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

Tutor-Marked Assignments (TMAs)

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

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

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

Final Examination and Grading

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

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

Course Marking Scheme

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

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

Course Overview

The Table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, Statistics for Economist (ECO 254).

Units	Title of Work	Week's Activities	Assessment (end of unit)
	Course Guide		
Module 1 Probability Distribution			
1	Analysis of Probability Distribution	Week 1	Assignment 1
2	Continuous Random Variables	Week 1	Assignment 1
3	Other Probability Distributions	Week 2	Assignment 1
Module 2 Hypotheses Testing			
1	Meaning of Hypothesis	Week 2	Assignment 1
2	The Criterion of Significance	Week 3	Assignment 1
3	Statistical Test for Hypothesis	Week 3	Assignment 2
4	Testing Differences Between Two Means	Week 4	Assignment 2
5	Testing Difference Between Matched Samples	Week 4	Assignment 2

Module 3 Sampling Theory			
1	Population and Sample	Week 5	Assignment 2
2	Population Parameters	Week 6	Assignment 2
3	Sampling Parameters	Week 7	Assignment 2
4	Calculation of Sampling Distribution and Estimators for Mean Variance	Week 8	Assignment 3
5	Frequency Distribution	Week 9	Assignment 3
Module 4 T-test, F-test and Chi-square analysis			
1	T test	Week 10	Assignment 3
2	F test	Week 11	Assignment 3
3	Chi-square analysis	Week 12	Assignment 4
Module 5 Simple Linear Regression Analysis and its Application			
1	Meaning of Regression Analysis	Week 13	Assignment 4
2	Simple Linear Regression Analysis	Week 14	Assignment 4
3	Application of Simple Linear Regression Analysis	Week 15	Assignment 4
	Total	15 Weeks	

How To Get The Most From This Course

In distance learning the study units replace the university lecturer. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace and at a time and place that suit you best.

Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your books or other material, and when to embark on discussion

with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provides exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit.

You should use these objectives to guide your study. When you have finished the unit you must go back and check whether you have achieved the objectives. If you make a habit of doing this you will significantly improve your chances of passing the course and getting the best grade.

The main body of the unit guides you through the required reading from other sources. This will usually be either from your set books or from a readings section. Some units require you to undertake practical overview of historical events. You will be directed when you need to embark on discussion and guided through the tasks you must do.

The purpose of the practical overview of some certain historical economic issues are in twofold. First, it will enhance your understanding of the material in the unit. Second, it will give you practical experience and skills to evaluate economic arguments, and understand the roles of history in guiding current economic policies and debates outside your studies. In any event, most of the critical thinking skills you will develop during studying are applicable in normal working practice, so it is important that you encounter them during your studies.

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

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

1. Read this Course Guide thoroughly.
2. Organize a study schedule. Refer to the 'Course overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your diary or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working through each 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 it to 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

The course, Statistics for Economist II (ECO 254), expose you to the analysis of probability distribution and you will also be introduced to continuous random variables and other higher probability distributions. This course also gives you insight into hypotheses testing in other to know some of the statistical test of hypothesis, testing the differences between two means and matched samples with the analysis of testing for level of significance. Thereafter it shall enlighten you about sampling theory which gives an insight to population parameter and analysis of sampling parameters. More so, the calculation of sampling distribution and estimators for mean variance was also examined to enable you understand more about sampling theory as well as the techniques of frequency distribution. Furthermore the course also enables you to know how to calculate t test, f test and chi-square analysis which you can use in different economics applications. Finally, the use of simple regression analysis in statistical test was also examined and exposes you to know how to apply it to economics problems.

On successful completion of the course, you would have developed critical thinking skills with the material necessary for efficient and effective discussion on macroeconomic issues: national income analysis, monetary issue, government expenditure and macroeconomics in open economy. However, to gain a lot from the course please try to apply anything you learn in the course to term papers writing in other economic development courses. We wish you success with the course and hope that you will find it fascinating and handy.

MODULE 1 PROBABILITY DISTRIBUTION

- Unit 1 Analysis of Probability Distribution
- Unit 2 Continuous Random Variables
- Unit 3 Other Probability Distributions

UNIT 1 ANALYSIS OF PROBABILITY DISTRIBUTION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Meaning of probability distribution
 - 3.2 Discrete Probability Distribution
 - 3.3 Continuous Probability Distribution
 - 3.4 Properties of probability Distributions
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Reading

1.0 INTRODUCTION

In probability and statistics, a probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference. Examples are found in experiments whose sample is non-numerical, where the distribution would be categorical distribution; experiments whose spaces is encoded by discrete random Variables, where the distribution can be specified by a probability mass function and experiments with sample spaces encoded by continuous random variables, where the distribution can be specified by a probability density function.

2.0 OBJECTIVES

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

- understand the meaning of probability distribution
- understand the application of probability
- understand different types of probability distribution
- Analyse probability problem.

3.0 MAIN CONTENT

3.1 Meaning of probability Distribution

To define probability distribution for simplest cases, you need to distinguish discrete and continuous random variables. In the discrete case, you can easily assign a probability to each possible value: for example, when throwing a die, each of the six values 1 to 6 has the probability $1/6$. In contrast, when a random variable takes value from a continuous, probabilities are non zero only if they refer to finite intervals: in quality control, one might demand that the probability of a 500g package containing between 490g and 510g should be no less than 98%.

If the random variable is real-valued (or more generally, if a total order is defined for its possible values), the cumulative distribution function (CDF) gives the probability that the random variable is no larger than a given value; in the real – valued case, the CDF is the integral of the probability density function provided that this function exists.

Let's consider the following three problems:

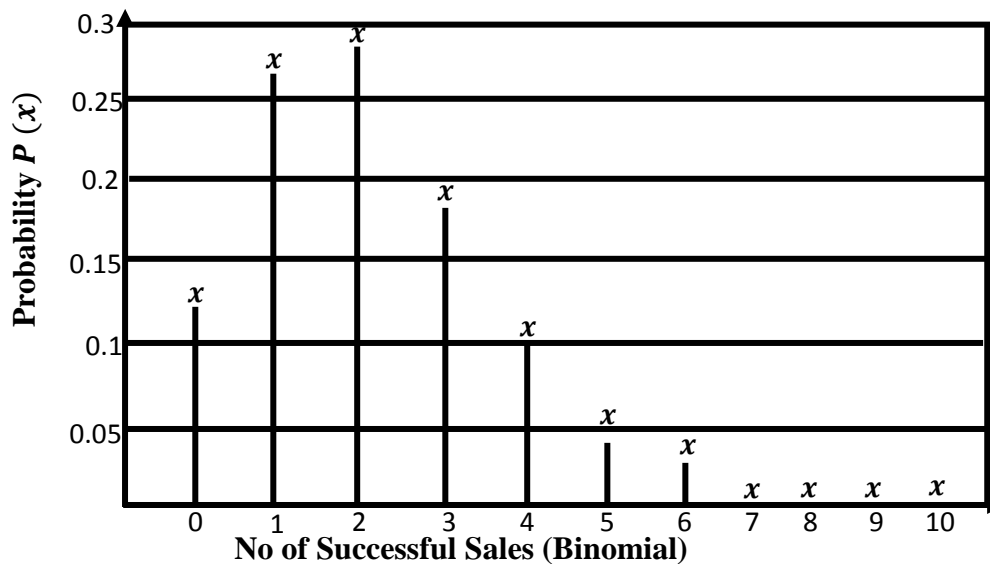
- Problem I: A sales person is given 25 addresses to call on each day. The household at each address has responded to a mail shot expressing an interest in having a sales person call to discuss the product. The salesperson's experience is that a sale is made at 1 in 10 households. What is the probability that 5 sales will be made in a given day?
- Problem II: A type of photocopier has a paper jam on average once every 3000 copies. What is the probability there will be more than two jams in a 3000 copy run?
- Problem III: The (arithmetic) mean life (continuous play on a specific cassette player) of a make and type of battery is 22 hours with a standard deviation of 0.11 hour. What is the probability a battery will last no more than 21 hours?

In problem I, the random variable of interest is the number of Sales per day. We refer to it as a random variable since the outcome in any one day is uncertain and may be considered to be dependent on chance. A random variable is a quantity resulting from a random experiment that by chance can assume different values. The experiment is calling each day on 25 addresses trying to sell the product. However, the different values are the number of sales each day.

In problem II, the different values are the number of sales each day. The ‘number of jams in 3000 copies’ and ‘life of a battery’ in problem III are also random variables. Associated with each possible value of the random variable in problem I is a probability, and the concern of the problem is with the probability that the random variable, denoted by x , takes the value of 5 that is $P(x=5)$. A probability distribution (also called a frequency function) maps out how probabilities are distributed over all possible values of the random variable.

However, probability distribution might be drawn up for the ‘number of jams in 3000 copies’ with associated probabilities for each value of x (though these will become very small for even relatively low values of x). Similarly, for problem III looking at table 1 below, we can conclude that the probabilities of a battery lasting for specific (non-overlapping and exhaustive) ranges of x (e.g. $0 \leq x < 1$) hour; $1 \leq x < 2$ (hours, etc). It should be noted that being able to evaluate the probability that a random variable takes a particular value (or range of values) is useful for planning purposes.

Table 1 Probability Distribution: Number of Successful Sales (Binomial)



In our examples, it helps the salesperson to evaluate the expected variability of bonuses and the manufacturer of copiers and batteries the performance of these products. However, in order to derive these variable distributions, it might seem we need to observe sales performance for a long period to identify the relative frequency (probability) with which different sales figures (value of x) emerge. The problem with this is, first by the time the data are collected they might be out of date. And, second, it is not helpful for ‘what if?’ questions: for example, if the overall proportion of successes (problem I) increased to 1 in 8, how would the distribution change? These difficulties can be overcome by expressing the distribution in a mathematical form which we can build into more complicated models of business-related behaviour.

Moreover, the magic of probability distributions is that, given limited information (for example in problem I on the overall proportion of successes and number of contact addresses), the probability of x taking any value or range of values can be generated. What we can do here is to try and identify which of a number of ‘Standard’ probability distribution might describe the problem situation, then identify the particular information or parameters that the particular distribution requires to generate the individual probabilities.

SELF ASSESSMENT EXERCISE

Briefly discuss the use of probability distribution in solving analytical problems.

3.2 Discrete Probability Distribution

A discrete probability distribution shall be understood as a probability distribution characterized by a probability mass function. Thus, the distribution of a random variable x is discrete and x is called a discrete random variable if:

$$\sum_u \Pr(x = u) = 1$$

as u runs through the set of all possible values of x . It follows that such a random variation can assume only a finite or countable infinite number of values. For the number of potential values to be countable infinite even though their probabilities sum to 1 requires that the probabilities decline to zero fast enough. For example, if

$\Pr(x=n) = \frac{1}{2^n}$ for $n = 1, 2, \dots$, we have the sum of probabilities $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1$.

In cases more frequently considered, this set of possible values is a topologically discrete set in the sense that all its points are isolated points. But there are discrete random variables for which this countable set is dense on the real line (for example, a distribution over rational numbers). Among the most well known discrete probability distributions that are used for statistical modeling are the ‘Poisson distribution’, ‘the Bernoulli distribution’, the binomial distribution’, the geometric distribution, and ‘the negative binomial distribution’. In addition, the discrete uniform distribution is commonly used in computer programs that make equal – probability random selections between a numbers of choices.

3.3. Continuous Probability Distribution

A continuous probability distribution is a probability distribution that has a probability density function. Mathematics also call such a distribution absolutely continuous, since its cumulative distribution function is absolutely continuous with respect to the Lebesgue measure λ (Lebesgue measure is the standard way of assigning a measure to a subsets of an n-dimensional volume). If the distribution of x is continuous, then x is called a continuous random variable. There are many examples of continuous probability distribution; normal, uniform, chi-squared and others.

Intuitively, a continuous random variable is the one which can take a continuous range of values – as opposed to a discrete distribution, where the set of possible values for the random variable is at most countable. While for a discrete distribution an event with probability zero is impossible (e.g. rolling 3½ on a standard die is impossible, and has probability zero), this is not so in the case of a continuous random variable. For example, if one measure the width of an oak leaf, the result of 3½cm is possible, however it has probability zero because there are uncountably many other potential values even between 3cm and 4cm. Each of these individual outcomes has probability zero, yet the probability that the outcome will fall into the interval (3cm, 4cm) is non zero. This apparent paradox is resolved by the fact that the probability that x attains some value within an infinite set, such as an interval, cannot be found by naively adding the probabilities for individual values. Formally, each value has an infinitesimally small probability, which statistically is equivalent to zero.

Formally, if x is a continuous random variable, then it has a probability density function $f(x)$ and therefore its probability of falling into a given interval, say (a, b) is given by the integral

$$Pr [a \leq x \leq b] = \int_a^b f(x)dx$$

In particular, the probability of x to take any single value a (that is $a \leq x \leq a$) is zero, because an integral with coinciding upper and lower limits is always equal to zero. However, the definition states that a continuous probability distribution must possess a density, or equivalently, its cumulative distribution function should be absolutely continuous. This requirement is stronger than simple continuity of the cumulative distribution function, and there is a special distribution, and singular distributions which are neither continuous nor discrete nor a mixture of those. An example is given by the Cantor distribution.

Such singular distributions however are never encountered in practice. But it should be noted that some authors use the term “continuous distribution” to denote the distribution cumulative distribution function. Thus, their definition includes both the (absolutely) continuous and singular distributions.

$$E(x) \left\{ \begin{array}{l} \sum_{-\infty}^{+\infty} xp(x) \text{ discrete random variable} \\ \int_{-\infty}^{+\infty} xf(x)dx \text{ continuous random variable} \end{array} \right.$$

It should be noted that expected value may be considered as averages in that the probabilities serve as relative frequencies in the formula for the mean i.e.

$$\bar{x} = \Sigma fx / \Sigma f$$

$$\Sigma \left(\frac{f}{\Sigma f} \right) x$$

Where $f/\Sigma f$ is the relative frequency or probability: it is a weighted average where the weights are the probabilities, giving relatively more emphasis to outcomes of the random variable with relatively high probabilities. We also denote $E[x]$ by the symbol μ (pronounced ‘mu’).

This is used as the mean of the population of all values of a variable. We use it here because the probability distribution covers all outcomes and $E(x)$ is, as just noted, an arithmetic mean.

The variance of a probability distribution can also be considered in a similar manner and provides a measure of the dispersion of the values of a random variable. The variance is denoted by the symbol σ^2 (pronounced sigma squared), therefore the lower case version of the Greek letter for s) and is defined by follows:

$$\sigma^2 \left\{ \begin{array}{l} \sum_{-\infty}^{+\infty} (x - \mu)^2 P(x) \text{ discrete random variables} \\ \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) \text{ continuous random variables} \end{array} \right.$$

The standard deviations are the square roots of these expressions. Table 1 above shows the expected value of the number of sales per day to be 2.0002 with a standard deviation $\sqrt{1.801} = 1.342$. The mean and standard deviations are interpreted along similar lines.

4.0 CONCLUSION

Probability distributions is used for so many purpose such as measurement of different possible outcome, a random experiment, survey, procedure of statistical inferences and each probability distribution is applied to a particular situation and analysis; for example, probability that an event will occur plus probability that an event will not occur is equal to 1, mathematically $\Pr(\text{an event})$

will occur) + Pr (an event will not occur) = 1. Therefore, probability distribution has a variety of application in different life situation.

5.0 SUMMARY

Probability distributions show the distributions of probabilities associated with values or ranges of a random variable. For discrete random variables the probabilities sum to 1, and for continuous random variables the equivalent of this is the area under the probability density function being equal to 1. The distributions naturally have means and variances that are calculated as shown above. For particular situations, we can judge the form the probability distribution is likely to take, and then consider some of those forms.

6.0 TUTOR MARKED ASSIGNMENT

1. The probability that it will rain in Lagos is $\frac{3}{4}$. Find the probability that it will not rain.
2. Differentiate with mathematical example between continuous and Discrete Probability distribution.
3. Discuss with detailed examples of properties of probability distribution.
4. Explain in detail the term 'probability distribution'.

7.0 REFERENCES

- Micks (1997) Business statistics, 2nd edition, Migrawhill Publishing Comapny Limited, Berkshire England.
- Nelson; L (1984) The Shewhart control chat – Test for Special courses, Journal of Quality Technology, 16, 237 – 23.

UNIT 2 CONTINUOUS RANDOM VARIABLES

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Continuous probability distribution
 - 3.2 Distribution functions for continuous random variables
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Reading

1.0 INTRODUCTION

A continuous random variable is a random variable where the data can take infinitely many values. For example a random variable measuring the time taken for something to be done is continuous since there are an infinite number of possible times that can be taken. A non-discrete random variable m is said to be absolutely continuous, and it can also be called simply continuous, if its distribution function can be denoted as:

$$F(m) = P(M \leq m) = \int_{-\infty}^m f(u) du \quad (1)$$

Where the function $f(x)$ has the properties:

1. $f(m) \geq 0$
2. $\int_{-\infty}^{\infty} f(m) dm = 1$

2.0 OBJECTIVES

The objective of this unit is to teach students the meaning and application of continuous random variables such as continuous probability distribution and Distribution functions for Continuous Random Variables.

3.0 MAIN CONTENT

3.1 Continuous Probability Distribution

Base on the above analysis in introduction, it follows that if m is a continuous random variable, then the probability that M takes on any one particular value is zero, where – as the internal probability that M lies between two different values, say q and p is given by

$$P(q < M < p) = \int_p^q f(m)dm \text{_____} (2)$$

Example 1

If an individual were selected at random from a large group of adult females, the probability that has height M is precisely 68 inches (that is, 68.00 inches) would be zero. However, there is a probability that is greater than zero that M is between 67.000 inches and 68.000 inches.

Solution

A function $f(m)$ that satisfies the above requirements is called a probability function or probability distribution for a continuous random variables, but it is more often called a probability density function or simply density function. However, any function $f(m)$ satisfying the two properties above with automatically be a density function and required probabilities can be obtained from equation (2) above.

Example 2

Find the constant b such that the function

$$f(x) = \begin{cases} bx^2 & 0 < x < 3 \\ 0 & \text{otherwise is the density} \end{cases}$$

Function and then find $P(1 < x < 2)$

Solution

Note that if $b \geq 0$, then property 1 (equation 1) is satisfied. So $f(x)$ must satisfy property 2 (equation 2) in order for it to be a density function.

Therefore:

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^3 bx^2 dx = \frac{bx^3}{3} \Big|_0^3 = 9b$$

And since this must equal 1, we have $b = \frac{1}{9}$, and therefore our density function is:

$$f(x) = \begin{cases} \frac{1}{9}bx^2 & 0 < x < 3 \\ 0 & \text{otherwise.} \end{cases}$$

Next,

$$P(1 < x < 2) = \int_1^2 \frac{1}{9}bx^2 dx = \frac{x^3}{27} \Big|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}.$$

SELF ASSESSMENT EXERCISE

Find the constant J such that the function

$$f(x) = \begin{cases} Jx^2 & 0 < x < 4 \\ 0 & \text{otherwise} \end{cases} \text{ is a density function, and then find } P(2 < x < 3).$$

3.2 Distribution Functions for Continuous Random Variables

The cumulative distribution function or distribution function for a random variable is defined by

$$f(x) = P(x \leq x) \text{_____} (3)$$

Where x is any real number, i.e. $-\infty < x < \infty$, so,

$$f(x) = \int_{-\infty}^x f(x) dx \text{_____} (4)$$

Example 3

Find the distribution function for:

$$f(x) = \begin{cases} bx^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

is the density function and then find $P(1 < x < 2)$.

Solution

$$F(x) = \int_{-\infty}^x f(x) dx = \int_0^x \frac{1}{9}x^2 dx = \frac{x^3}{27} \text{ where } x \leq 3.$$

We can see that there is a nice relationship between the distribution function and the density function. For us to see this relationship, we can consider the probability that a random variable x takes on a value, x and a value fairly close to x ,

say $x + \Delta x$ is given by

$$P(x \leq x \leq x + \Delta x) = \int_x^{x+\Delta x} f(x) dx \quad (5)$$

so that if Δx is small, we have approximately

$$P(x \leq x \leq x + \Delta x) \approx f(x)\Delta x \quad (6)$$

We also can see from (1) above in differentiating both sides that:

$$\frac{dF(x)}{dx} = f(x) \quad (7)$$

at all points where $f(x)$ is continuous, i.e. the derivative of the distribution function is the density function.

SELF ASSESSMENT EXERCISE

Find the distribution function for:

$$f(x) = \begin{cases} kx^2 & 0 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

is a density function, and then find $P(2 < x < 6)$.

4.0 CONCLUSION

In this unit, we can conclude the continuous random variables take any value in an interval. They are used to model physical characteristics such as time, length, position, etc.

5.0 SUMMARY

In summary continuous random variable can be in continuous probability distribution and distribution functions for continuous random variables. Therefore, I believe that at the end of this unit you must have understand and know how to analyse/ calculate continuous random variables.

6.0 TUTOR-MARKED ASSIGNMENT

1. Differentiate between continuous random variables and continuous probability distribution.

2. Given

$$f(x) = \begin{cases} cx^2 & 0 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

is the density function and then find $P(4 < x < 8)$.

7.0 REFERENCES/ FURTHER READINGS

Murray, R. S & John S & Srinivasan R-A (2001) *Probability and Statistics*, Schaum's outline series, Mcgraw – Hill, USA.

Ojo, J. B (2011) *Statistics made easy*, melting point publication, Lagos.

UNIT 3 OTHER PROBABILITY DISTRIBUTIONS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 The Multinomial distributions
 - 3.2 The Hyper geometric distribution
 - 3.3 The uniform distributions
 - 3.4 Cauchy, Gamma and Bata Distribution
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Reading

1.0 INTRODUCTION

Other probability distribution is not the same as the other probability distributions we treated in Unit 1 and Unit 2 above. However, this unit explains the other higher level of probability distribution such as multinomial distribution, the hyper geometric distribution, the Cauchy distribution, Gamma distribution and Beta distribution respectively.

2.0 OBJECTIVES

At the end of this unit you must be able to distinguish and explain the various higher probability distributions and their applications.

3.0 MAIN CONCEPT

3.1 The Multinomial Distribution

Let us make some hypothetical examples such that: an event A_1, A_2, \dots, A_k are mutually exclusive, and can occur with respective probabilities P_1, P_2, \dots, P_k where $P_1 + P_2 + \dots + P_k = 1$. If X_1, X_2, \dots, X_k are the random variables respectively, given the number of times that A_1, A_2, \dots, A_k occur in a total of n trials, so that $X_1 + X_2 + \dots + X_k = n$, then $P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} P_1^{n_1} P_2^{n_2} \dots P_k^{n_k}$ (1)

where $n_1 + n_2 + \dots + n_k = n$, is the joint probability function for the random variables x_1, x_2, \dots, x_k

However, the distribution that are analysed above are subject to binomial distribution and it is called multinomial distribution because the above equation

we have is the general term in the multinomial expansion of $P(P_1 + P_2 + \dots + P_k)^n$.

SELF ASSESSMENT EXERCISE

Explain the term ‘the multinomial distribution’

3.2 The Hyper geometric Distribution

This distribution’ like the binomial and Poisson distribution, is a probability distribution that makes use of discrete variables. It makes use of combinatorial analysis.

Suppose the N members of a sample can be divided into two dichotomous groups 1 and 2 with group 1 having R members and group 2, N – R members. Suppose further that n numbers are to be selected, without replacement and at random from the N members in such a way that h of the members come from group 1 then the probability of selecting these h members is given by:

$$P(x = h) = \frac{{}^R C_n \times {}^{N-R} C_{n-h}}{{}^N C_n}$$

that is,

$$\frac{\binom{R}{h} \binom{N-R}{n-h}}{\binom{N}{n}}$$

Where h takes integral values of 0, 1, 2, 3, n. The above distribution is known as the hyper geometric probability distribution.

Example 1

Three out of the 9 finalists in an African American beauty competition are Nigerians. If two winners are to be selected, find the probability that:

- (a) at least one of them would be a Nigerian.
- (b) only one of them would be a Nigerian.

Solution

According to the hyper geometric distribution:

$$P(x = h) = \frac{{}^R C_n \times {}^{N-R} C_{n-h}}{{}^N C_n}$$

- (a) $N = 9, n = 2, R = 3.$

$P(\text{at least one of them would be a Nigerian}) = 1 - P(\text{None of them is a Nigerian})$

For $P(\text{None is a Nigerian}) h = 0.$

$$\therefore P(h = 0) = \frac{{}^3 C_0 {}^{9-3} C_{2-0}}{{}^9 C_2} = \frac{{}^3 C_0 \times {}^6 C_2}{{}^9 C_2} = \frac{1 \times 15}{36} = \frac{15}{36} = 0.42$$

$P(\text{at least one of them would be a Nigerian}) = 1 - 0.42 = 0.58$

- (b) $P(\text{Only one would be a Nigerian}), \text{ in this case } h = 1$

$$P(h = 1) = \frac{{}^3 C_1 \times {}^{9-3} C_{2-1}}{{}^9 C_2} = \frac{{}^3 C_1 \times {}^6 C_1}{{}^9 C_2} = \frac{3 \times 6}{36} = 0.5$$

SELF ASSESSMENT EXERCISES

A company short-lists 12 applicants for an interview out of which 4 are men. If there are only five vacant posts to be filled, find the probability that the list of successful applicant:

- (a) Contains 3 men
- (b) Contains at least 2 men
- (c) Contains only 1 man
- (d) Contains all women

3.3 The uniform distribution

In statistics, this is a type of probability distribution in which all outcomes are equally likely. A deck of cards has a uniform distribution because the likelihood of drawing a heart, a club, a diamond or a spade is equally likely. A coin also has a uniform distribution because the probability of getting either heads or tails in a coin toss is the same. However, uniform distribution can also be seen as a statistical distribution in which every possible outcome has an equal chance, or likelihood, of occurring (1 out of the total number of outcomes). For example,

imagine a man standing on a street corner handing a N500 to a lucky passersby. If it were completely random, then every person that walked by would have an equal chance of getting the N500. This is an example of a uniform probability distribution. It's uniform because everyone has an equal chance (probability percent is equal to 1 divided by the number of people walking by). If the man favored tall people or dark-haired people, and was more likely to give them the money instead of others, well, that would not be uniform, because some would have a higher probability of getting a dollar than others.

A random variables x is said to be uniformly distributed in $a \leq x \leq b$ if its density function is

$$f(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and the distribution is called a uniform distribution. The distribution function is given by:

$$f(x) = P(x \leq x) = \begin{cases} 0 & x < a \\ (n-a)/(b-a) & a \leq x < b \\ 1 & x \geq b \end{cases} \quad (2)$$

The mean and variance are, respectively

$$\mu = \frac{1}{2}(a+b) \quad \sigma^2 = \frac{1}{12}(b-a)^2 \quad (3)$$

Examples

Calculate the mean and variance of a uniform distribution given that

$$(a+b) = 4/128 \text{ and } b = 6.40 \text{ while } a = 2.10.$$

Solution

$$\mu = 1/2(a+b) \Rightarrow \mu = \frac{1}{2} \times \frac{4}{128} = \frac{4}{2 \times 128} = \frac{4}{256} = 0.016$$

$$\sigma^2 = \frac{1}{2}(b-a)^2$$

While $b = 6.40$ and $a = 2.10$

$$\sigma^2 = \frac{1}{2}(6.40 - 2.10) \Rightarrow \sigma^2 = \frac{1}{2} = (4.3) = \frac{4.3}{2} = 2.15$$

SELF ASSESSMENT EXERCISES

Calculate the mean and variance of a uniform distribution given that
 $a = 4.42$ and $b = 8.22$ respectively.

3.4 Cauchy, Gamma and Beta Distribution

3.4.1 Cauchy Distribution

The Cauchy distribution, named after Augustin Cauchy, is a continuous probability distribution. It is also known, especially among physicists, as the Lorentz distribution (after Hendrik Lorentz), Cauchy–Lorentz distribution, Lorentz(ian) function, or Breit–Wigner distribution.

The Cauchy distribution is often used in statistics as the canonical example of a "pathological" distribution since both its mean and its variance are undefined. (But see the section Explanation of undefined moments below.) The Cauchy distribution does not have finite moments of order greater than or equal to one; only fractional absolute moments exist. The Cauchy distribution has no moment generating function.

A random variable X is said to be Cauchy distributed or it can be said that it has Cauchy distribution, if the density function of the variable X is;

$$f(x) = \frac{a}{\pi(x^2 + a^2)} \quad a > 0, -\infty < x < \infty \quad (4)$$

The density function is symmetrical about $x = 0$, so that its median is zero. But it should be noted that the mean and variance do not exist.

3.4.2 Gamma Distribution

The gamma distribution is a two-parameter family of continuous probability distributions. The common exponential distribution and chi-squared distribution are special cases of the gamma distribution. There are three different parametrizations in common use:

1. With a shape parameter k and a scale parameter θ .
2. With a shape parameter $\alpha = k$ and an inverse scale parameter $\beta = 1/\theta$, called a rate parameter.
3. With a shape parameter k and a mean parameter $\mu = k/\beta$.

In each of these three forms, both parameters are positive real numbers.

A random variable X is said to have the gamma distribution, or it can be said that is gamma distribution, if the density function is:

$$f(x) = \begin{cases} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ \frac{\beta^\alpha \Gamma(\alpha)}{\Gamma(\alpha)} & x \leq 0 \end{cases} (\alpha, \beta > 0) \text{_____} (5)$$

where $\Gamma(\alpha)$ is the gamma function and the mean and the variance are given as follows;

$$\mu = \alpha\beta \quad \sigma^2 = \alpha\beta^2 \text{_____} (6)$$

3.4.3 Beta Distribution

The beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ parametrized by two positive shape parameters, denoted by α and β , that appear as exponents of the random variable and control the shape of the distribution.

The beta distribution has been applied to model the behavior of random variables limited to intervals of finite length in a wide variety of disciplines. For example, it has been used as a statistical description of allele frequencies in population genetics; time allocation in project management / control systems; sunshine data; variability of soil properties; proportions of the minerals in rocks in stratigraphy; and heterogeneity in the probability of HIV transmission.

In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions. For example, the beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success such as the probability that a space vehicle will successfully complete a specified mission. The beta distribution is a suitable model for the random behavior of percentages and proportions.

The usual formulation of the beta distribution is also known as the beta distribution of the first kind, whereas *beta distribution of the second kind* is an alternative name for the beta prime distribution.

However, a random variable is said have the beta distribution, or to be beta distributed as it may be called in another way, if the density function is:

$$f(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} & 0 < x < 1 \\ 0 & \text{otherwise } (\alpha, \beta > 0) \end{cases} \text{_____} (7)$$

where $B(\alpha, \beta)$ is the beta function. However, in view of the relationship between the beta and gamma functions, the beta distribution can also be defined by the density function below;

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad 0 < x < 1 \quad (8)$$

where α, β are positive. Therefore we can then specify the mean and variance as follows;

$$\mu = \frac{\alpha}{\alpha + \beta}, \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (9)$$

for $\alpha > 1, \beta > 1$, there is a unique mode at the value

$$x_{mode} = \frac{\alpha - 1}{\alpha + \beta - 2} \quad (10)$$

SELF ASSESSMENT EXERCISES

Differentiate with example using equation and parameters between Cauchy, Gamma and Beta distribution.

4.0 CONCLUSION

Base on the analysis that we have discussed in this unit, we can conclude that other probability distribution better explain the use of mean, median and variance of different parameters.

5.0 SUMMARY

Other probability distribution has enlightened us in this unit, on how we can use it to calculate/analyse different parameters such as alpha (α), beta (β), gamma (Γ) and even the measure of central tendency such as mean, median and mode plus the variance.

6.0 TUTOR MARKED ASSIGNMENT

1. Briefly discuss the importance of higher order probability distribution in statistics.
2. Write short note on the following with detailed examples:
 - (a) Multinomial distribution

- (b) Hyper geometric distribution
- (c) Cauchy distribution
- (d) Gamma distribution
- (e) Beta distribution

7.0 REFERENCES/FURTHER READING

Adedayo, O.A (2000) Understanding Statistics, JAS Publisher Akoka, Lagos

Murray, R.S & John S. & Srinivasan, R.A (2001) Probability and Statistics, Schaum's utline Series, McGraw-Hill, USA.

MODULE 2 HYPOTHESES TESTING

Unit 1 Meaning of Hypothesis

Unit 2 The Criterion of Significance

Unit 3 Statistical Test for Hypothesis

Unit 4 Testing Differences between Two Means

Unit 5 Testing Difference between Matched Samples

UNIT 1 MEANING OF HYPOTHESIS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Contents
 - 3.1 Statistical Hypothesis
 - 3.2 Type 1 and Type 2 Errors
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assessment
- 7.0 References/Further Readings

1.0 INTRODUCTION

One of the uses of Statistics is to make a decisive decision. However, in some cases the decisions may be on a population based on results obtained from selected samples of the said given population. For example, a doctor may want to know which of the malaria drugs work faster on a patient, a lecturer may wish to take decision on strategies to use in teaching his/her students. Hypotheses are formulated and different types of statistical tests are carried out in order to make a reasonable decision but some errors are committed in the quest for taking a good decision.

2.0 OBJECTIVES

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

- know the meaning of hypotheses testing
- understand type 1 and type 2 errors
- be able to calculate one-tailed and two-tailed tests
- be able to calculate various hypothesis testing
- be able to apply hypotheses testing to economics problems.

3.0 MAIN CONTENTS

3.1 Statistical Hypothesis

You may be wondering that why do we have to take a step in going forward to take a good statistical hypothesis. However, when taking statistical decision, the first step is to make an assumptions or guesses about the population you want to study. These assumptions are called and known as hypotheses. Usually, an hypothesis is formulated with the basic idea of nullifying the hypotheses and rendering the hypotheses insignificant. Therefore, any assumptions made with the sole purpose of rendering the statistical hypothesis insignificant is called a null hypothesis. For example, when a lecturer wants to know the best method of teaching, he will need to take or formulate a null hypothesis that there is no difference in teaching methodology. Null hypothesis are usually denoted with the symbol(H_0).

However, apart from the null hypothesis, one needs to formulate another hypothesis which is quite different from the null hypotheses and this hypothesis is known as the alternative hypothesis and it is denoted as(H_1).

The hypothesis may be directional or non-directional, but if one takes about “no difference” then we have non directional hypothesis but if the aim is to conclude that one item is “better” or “less” than the other then we have directional hypothesis. Let us take some examples to analyse this issues:

1. **Null Hypothesis(H_0)** ∴Product P and product Q are equally popular.
Alternative Hypothesis(H_1) ∴Product P is more popular than product Q-
Then we can say that we have a case of directional hypothesis because of the word “more”. So an example of directional type we can have $H_1: \mu_1 < \mu_2$.
2. **Hypothesis(H_0)**: There is no difference in their mean scores of male and female students in Noun. ($\mu_1 = \mu_2$).
Alternative Hypothesis(H_1) ∴There is a difference in their mean scores, ($\mu_1 \neq \mu_2$). So this is a non directional hypothesis. More so, after setting up the null and alternative hypotheses statistical test are carried out to justify whether to reject the null hypothesis or the alternative hypothesis using a level of significance (1% or 5% level of significance).

SELF ASSESSMENT EXERCISE

Differentiate between Null hypothesis and Alternative hypothesis with detailed examples.

3.2 Type 1 and Type 2 Errors

There are two types of errors in hypothesis testing, it is called type 1 and type 2 errors.

However, type 1 error occurs when/if an hypothesis (Null hypothesis) is rejected when it should be accepted and this occurs when the hypothesis value falls within acceptance region falls within the rejection region while type 2 error is the reverse that is one accepts the hypothesis when it should be rejected. In our day to day business activities, type 1 error is known as producer's risk while type 2 error is known as consumer's risk. For example, if a murderer is taken to court and the judge frees him, he has committed a type 1 error when the null hypothesis that he is guilty is being tested. Another practical example we can look at in this unit is the care of a study director in NOUN who is supposed to raise alarm when a student presents forged certificates, he must therefore decide between:

H_0 : the certificate is genuine

and H_1 : the certificate is a forged one and the student should be expelled.

A false alarm by the officer will indicate that he is rejecting a true hypothesis and it's therefore committing a type 1 error.

However, a missed alarm implies that he is accepting that the certificate is genuine when he should reject it and thus committing a type 2 error. In test of hypothesis, it is desirable that the rule of decision is taken in such a way that the two errors usually lead to an increase in the other error. In some instances, one type of error may be more serious than then other. For example, if the null hypothesis states that there is a dangerous level of outbreak in an environment, committing type 2 error (accepting what is not true) and a compromise should be reached in favour of the more hazardous or serious error, but the best ways of reducing both errors is to increase the sample but is not possible in all cases.

It should be noted that the probability (or risk) or committing type 1 error on a true null hypothesis is denoted by the Greek letter alpha (α) and it's called α -risk but the probability of committing a type 2 error is denoted by the Greek letter beta (β)and it's called beta risk. The probability of correctly rejecting (H_0)when it is false is called the power of the statistical test and it's denoted by $1 - \beta$ while the probability of correctly accepting H_0 is equal to $1 - \alpha$

SELF ASSESSMENT EXERCISE

Differentiate between the following terms:

- i. Type 1 and type 2 errors
- ii. Null hypothesis and alternative hypothesis.

4.0 CONCLUSION

Hypothesis testing has been seen as the way forward to justify the truth and false of a research situation or problem. So at this end, I believe you must have gained enough on how we formulate hypothesis.

5.0 SUMMARY

This unit has vividly taken a look at the meaning of hypothesis and how it is used in solving some research problems. However, type 1 and type 2 errors were also examined to know whether we have accepted or rejected what we do not need to accept or reject.

6.0 TUTOR MARKED ASSESSMENT

1. Write short note on the following:
 - (a) Null hypothesis and alternative hypothesis
 - (b) Type 1 and type 2 errors
 - (c) Non directional and Directional Hypothesis.
2. The probability that an Ade is taking a statistical risk is given as $\frac{1}{2}$ and probability that he is not taking a statistical risk is $\frac{2}{3}$. Show how a null and alternative hypothesis can be specify with Ade's decision.

7.0 REFERENCES/FURTHER READINGS

Adedayo, O. A (2000), Understanding Statistics, JAS publisher Akoka, Lagos.

UNIT 2 THE CRITERION OF SIGNIFICANCE

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 One-Tailed and Two-Tailed Tests
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In test of hypothesis, the maximum probability of risking a type 1 error is known as the level of significance and the probability is usually decided upon before data collection. You should note that the numerical value of the decision rule is called criterion of significance or level of significance. But the most common levels used in hypothesis testing are 0.05 and 0.01. If we make use of the alpha level of 0.05, we are 95% confident that a right decision has been made; that is, an average of 5 out of a 100 would be the times we commit a type 1 error and incorrectly reject the null hypothesis. However, when we use the 0.01 level, we are 99% confident that a right decision has been made. The selection of the criterion of significance depends on the type of errors which the investigator considers to be more serious. For example, an hypothesis that a specified level of an outbreak is safe should be tested at a high significant level (e.g. 99%) because not rejecting the hypothesis would be very dangerous than rejecting it should the assertion prove to be false.

2.0 OBJECTIVES

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

- Differentiate between one-tailed and two-tailed tests.
- Understand the procedures for carrying out test of Hypothesis.

3.0 MAIN CONTENT

3.1 One-tailed and Two-tailed tests

1. One-tailed test

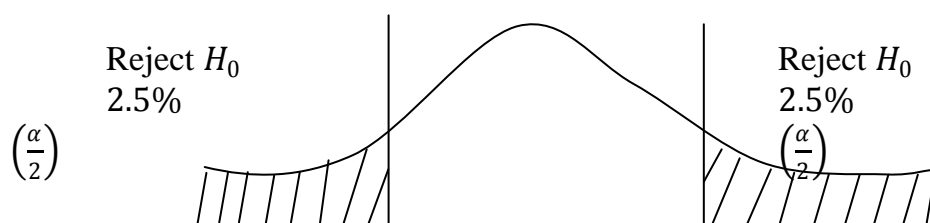
A statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample that is being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis. The one-tailed test gets its name from testing the area under one of the tails (sides) of a normal distribution, although the test can be used in other non-normal distributions as well.

2. Two-tailed test

A statistical test in which the critical area of a distribution is two sided and tests whether a sample is either greater than or less than a certain range of values. If the sample that is being tested falls into either of the critical areas, the alternative hypothesis will be accepted instead of the null hypothesis. The two-tailed test gets its name from testing the area under both of the tails (sides) of a normal distribution, although the test can be used in other non-normal distributions.

The normal curve is one of the most popular models used in statistical tests of hypothesis. For a non directional hypothesis, a two-tailed test is used when finding the critical region.

If the 0.05 level of significance is to be used in a two-tailed test, the 0.05 level is shared between the two ends of the tails giving 0.025 or 2½%. The rejection areas in both tails are displayed below:

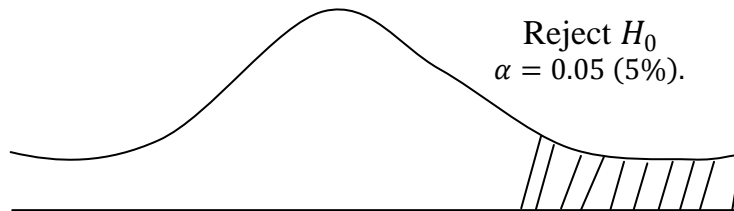


Two – tailed.

The cut off scores beyond which H_0 should be rejected can be estimated from the normal table if the normal curve model is used. If the computed value is less than obtained (known as critical value and read off from a table) we do not reject the null hypothesis while if it is more we reject the null hypothesis. Moreover, the two-tailed we are interested in deviant (extreme) values of the statistics.

But it should be noted that if the focus of interest is on one side of the mean, as in the case with directional hypothesis, a one-tailed test is used. The critical region in this case is on one side of the curve and so the rejection area is one sided.

For the 0.05 level, we obtain the diagram below:



SELF ASSESSMENT EXERCISE

Differentiate between one-tailed and two tailed tests.

4.0 CONCLUSION

This unit has vividly looks at one and two tailed test. However, the unit analyses that in statistical significance testing, a one-tailed test or two-tailed test are alternative ways of computing the statistical significance of a data set in terms of a test statistic, depending on whether only one direction is considered extreme (and unlikely) or both directions are considered equally likely.

5.0 SUMMARY

One and two tailed test has be discussed extremely in this unit and their graphs was also examined. Therefore we can conclude that one-tailed tests are used for asymmetric distributions that have a single tail, such as the chi-squared distribution, which are common in measuring goodness-of-fit, or for one side of a distribution that has two tails, such as the normal distribution, which is common in estimating location; this corresponds to specifying a direction. Two-tailed tests are only applicable when there are two tails, such as in the normal distribution, and correspond to considering either direction significant.

6.0 TUTOR MARKED ASSIGNMENT

1. Distinguish between One tailed test and two test.
2. Explain the decision rule in accepting and rejecting null hypothesis using the one tailed and two tailed test.

7.0 REFERENCES/FURTHER READINGS

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

Gago, C.C. (2009). *Statistics for Economist*, 1st edition DALT Publication limited.

UNIT 3 STATISTICAL TEST FOR HYPOTHESIS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Procedures for Carrying Out Tests for Hypothesis
 - 3.1.1 Statistical Test for Mean of a Single Population when Population Variance is Known
 - 3.2 Statistical Test of a Single Population when the Population Variance is known
 - 3.3 Interval Estimation for Mean of a Single Population
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit, we shall try to explain and show the different calculation of carrying out test of hypothesis. However, the procedures of carrying the test out are a basic root in getting to calculate different analysis of sample in a population.

2.0 OBJECTIVES

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

- know the procedures for carrying out the test of hypothesis.
- know how to calculate statistical test for mean of a single population
- know how to calculate the interval estimation for mean of a single population.

3.0 MAIN CONTENT

3.1 Procedures for Carrying Out Tests for Hypothesis

The following are the general steps to take when testing hypothesis. Most of the time, the distributors involved are the normal and t-distribution.

- (a) State the null hypothesis(H_0) and the alternative hypothesis (H_1).
- (b) State the criterion level of significance given.
- (c) Calculate the mean and standard deviation of the given population or their estimates, if not given.

- (d) Compute the appropriate statistics which could be standard z or t value using the appropriate formulas and obtain the calculated value.
- (e) Determine the tabulated or critical value corresponding to the given level of significance. Care must be taken about whether the test is a two-tailed or one-tailed type when determining the critical values.
- (f) If the calculated value is less than the tabulated value (i.e. falls within the accepted region), we accept the null hypothesis. If the calculated statistic is more than the tabulated value (i.e. lies in the rejection area), we reject H_0 and make our conclusion.

3.1.1 Statistical Test for Mean of a Single Population when Population Variance is Known

In this situation, the population for which inferences is to be made is assumed to be normally distributed with mean(μ) and variance σ^2 . The test statistic will be the z-test.

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Where \bar{x} = sample mean, n = sample size, μ = the hypothesized value of the population mean, σ = population standard deviation.

Note that $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean when σ is known.

Exercise 1

The mean of 25 samples selected from a population of mean, μ and variance 100 is 52. Test the hypothesis $H_0: \mu = 49$ vs $H_1: \mu > 49$ at 0.05 level of significance.

Solution

Using the formula $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

Since is known $\sigma = \sqrt{100} = 10$

Therefore, $Z = \frac{52 - 49}{\frac{10}{\sqrt{25}}} = \frac{3}{\frac{10}{5}} = 1.5$

This is a directional test and the critical value is one tailed at = 0.05. From the table, the tabulated value (1.65) is greater than the computed value. Therefore, we do not reject the hypothesis and we retain H_0 .

3.2 Statistical Test of a Single Population when the Population Variance is known

In this case, the student t statistics is used. The formular is given as follows:

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

And the number of degree of freedom is n-1 and the population is assumed to be normal.

Example 2

A midwife claims that the mean weights of babies delivered at her maternity clinic is 3.5kg. A statistician takes a sample of 10 babies and obtains the following weights:

2.8, 2.5, 3.2, 3.5, 3.7, 2.7, 4.0, 4.5, 3.9, 3.6. Test the midwife's claim at 0.05 level of significance.

Solution

$H_0: \mu = 3.5$ versus $H_1: \mu \neq 3.5$.

The first step here is to find the mean of the sample and the standard deviation using formula $\bar{x} = \frac{\sum x}{N}$

We then obtain mean = 3.44, using the formula for standard deviation $\sqrt{\frac{\sum(x-\bar{x})^2}{N-1}}$ since it is a sample, we then obtain $\sigma = 0.60$.

Therefore $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

We then have: $\frac{3.44-3.5}{\frac{0.6}{\sqrt{10}}} = \frac{-0.06}{\frac{0.6}{\sqrt{10}}} = 0.32$.

This is a two tailed test with d.f = 10 - 1 = 9. However, from the table, the t-statistic value for 0.05 at 9 degree of freedom is 2.26. Since the calculated value 0.32 is less than the computed value of 2.26, we do not reject the midwife's claim.

SELF ASSESSMENT EXERCISES

The mean life span of electric bulbs produced by a company has in the past been 1100 hours with a standard deviation of 140 hours. A sample of 13 electric bulbs recently taken showed a mean life span of 1020 hours. Using

criterion levels of 0.05 and 0.01 respectively, tests the following hypothesis: (a)

$$H_0: \mu_1 = 100 \text{ vs } H_1: \mu \neq 1100$$

$$(b) H_0: \mu_1 = 1100 \text{ vs } H_1: \mu_0 < 1100.$$

3.3 Interval Estimation for Mean of a Single Population

Another technique that can be employed with respect to the issue of rejecting or accepting H_0 is interval estimation. However, this process involves estimating an interval which is known as confidence interval, within which the population mean is likely to fall. H_0 is rejected if computation reveals that the value of the population mean (μ) assumed under the null hypothesis falls outside the interval.

Therefore the use of interval estimation enables us to have an estimate of how sample mean deviates from the population mean.

In other words, interval estimation enables us to know how much samples are likely to vary from the population due to a sampling error and one can then be sure or confident that the population mean falls within the confident interval. But it should be noted that the conference limit are the end points of the confidence interval. More so, one advantage of its use is that we can test several different hypothetical values of the population mean, μ , without making extra computational effort. We do not need to compute different t values if we have to test samples for different values of μ .

Example 3

$$\text{Suppose } N = 25, \bar{x} = 20.4, s = 8.0, \text{ then } S_x = \frac{s}{\sqrt{n}} = \frac{8.0}{\sqrt{25}} = \frac{8.0}{5} = 1.6$$

At $\alpha = 0.05$, we talk of 95% confidence interval. If we wish to find the 95% confidence interval, we need to obtain the value of t. Since $N = 25$, the number of degrees of freedom = $N - 1 = 25 - 1 = 24$ and the tabulated t-value is 2.15, for two-tailed test. The interval is thus given by $\bar{x} - t_{s_{\bar{x}}} \leq \mu \leq \bar{x} + t_{s_{\bar{x}}}$.

Where $\bar{x} = 20.4, t = 2.15, s_{\bar{x}} = 1.6$,
confidence interval is

$$20.4 - (2.15 \times 1.6) \leq \mu \leq 20.4 + (2.15 \times 1.6)$$

that is $20.4 - 3.44 \leq \mu \leq 20.4 + 3.44$

$$16.96 \leq \mu \leq 23.84.$$

The confidence limits are 16.96 and 23.84.

At $\alpha = 0.01$ that is at the 99% confidence interval for the population means, μ .

$$\bar{x} - t_s \bar{x} \leq \mu \leq \bar{x} + t_s \bar{x}$$

This becomes:

$$20.4 - (2.98 \times 1.6) \leq \mu \leq 20.4 + (2.98 \times 1.6)$$

$$20.4 - 4.768 \leq \mu \leq 20.4 + 4.768$$

$$15.632 \leq \mu \leq 25.168$$

The confidence limits are 15.632 and 25.168.

Furthermore, you must note the following:

- (a) The critical value of t used in the computation of the confidence intervals depends on N and the degree of freedom.
- (b) It also depends on whether the hypothesis is directional or non-directional.
- (c) If the population variance is known, use the critical values of z (1.96 for 95% confidence interval and 2.58 for 99% interval).

SELF ASSESSMENT EXERCISE

Given $\bar{x} = 30.6$, $t = 4.25$, $S_{\bar{x}} = 1.8$, calculate the confidence interval at 95% and 99% level of significance.

4.0 CONCLUSION

This unit has look at the statistical test for hypothesis, and we can conclude that hypothesis testing is a process by which an analyst tests a statistical hypothesis. The methodology employed by the analyst depends on the nature of the data used, and the goals of the analysis. The goal is to either accept or reject the null hypothesis.

5.0 SUMMARY

In this unit we have examine various aspect of statistical test for hypothesis such as procedure for carrying out tests for hypothesis, statistical test for mean of a single population when population variance is known and unknown and interval estimation for mean of a single population. Therefore, it is at this point that I believe that you must have learnt a lot from the unit.

6.0 TUTOR MARKED ASSIGNMENT

1. The mean yield of wheat harvested on a treated plot of size 1 was 5.2 litres with a standard deviation of 3 while that of an untreated plot of size 16 was 4.7 with a standard deviation of 3.5. Compute

- (i) the pooled variance
- (ii) the 90% confidence interval for the difference in the means.
- 2. Differentiate between Statistical test for mean of a single population when population variance is known and when is unknown.
- 3. Explain the term “interval estimation for mean of a single population.

7.0 REFERENCES/FURTHER READINGS

Samuelson, H. (2012). *Introduction to Statistical for Economics*, Mill world Publication limited, 2nd edition.

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

UNIT 4 TESTING DIFFERENCES BETWEEN TWO MEANS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Testing Difference between Two Means of Independent Samples
 - 3.2 Confidence intervals for Differences between Two means
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Reading

1.0 INTRODUCTION

Based on our discussion in Unit 3, we can then go on to discuss the test of difference between two means. However, in this unit, we shall proceed to carry out series calculation on how to do a good calculation of difference of two mean.

Therefore, we now go on to test differences between two means that come from:

- (a) Independent samples and
- (b) Dependent matched samples.

2.0 OBJECTIVES

At the end of this unit, you should be able to understand the test of difference between two means of independent samples, confidence intervals for Difference between two means.

3.0 MAIN CONTENT

3.1 Testing Difference between Two Means of Independent Samples

If you recall in our last discuss in unit 3 where we discussed more on the techniques of drawing inferences between two or more population. However, in some cases at times, there may be need to draw inferences about differences between two or more populations. But you should note that in an experimental research, the scientist may have two groups, an experimental group and a control group, and may wish to test if there is any difference between the two groups. For example, a chemist may wish to check if two types of solutions have different degree of acidity, the agriculturist may wish to test the effect of

fertilizer on crop yield and compare yields from a plot treated with the fertilizer and another plot treated without the fertilizer. The decision taken will be based on the test of hypothesis known as the students test given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{\sqrt{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}}{N_1 + N_2 - 2}} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$$

Where the degrees of freedom is $N_1 + N_2 - 2$. Note that $S_{\bar{x}_1 - \bar{x}_2}$ is the standard error of the mean difference and \bar{x}_1 and \bar{x}_2 are the respective sample means of the two groups S_1 and S_2 are the standard deviation, N_1 and N_2 are the sample size of the two groups. The above formula is used when the population variance is known. But however, if the population variances are known, the standard score is used:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Where σ_1^2 and σ_2^2 are the two percent or population variances. However, in some test books, they make use of

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2} \right) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

But the formula above is the same as the one given earlier above if our N is very large. However, we still make use of the one we specified earlier above and it should be noted that:

$$\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} \text{ is the pooled sample variance.}$$

Example 1

In a statistics examination for secondary students, the 22 females used in the study has a mean score of 81 and a variance of 12 while the 20 males used has a mean score of 78 and a variance of 10. Do you think gender have an effect on the score of these secondary students at $\alpha = 0.05$ and $\alpha = 0.01$?

Solution

$H_0: \mu_1 = \mu_2$ and there is no difference versus $H_1: \mu_1 \neq \mu_2$

Let $N_1 = 22, N_2 = 20, S_1^2 = 12, S_2^2 = 10, \bar{x}_1 = 81, \bar{x}_2 = 78,$

Using the formular:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(N_1-1)S_1^2 + (N_2-1)S_2^2}} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$$

$$t = \frac{81-78}{\sqrt{\left(\frac{21 \times 12 \times 19 \times 10}{22+20-2}\right)\left(\frac{1}{22} + \frac{1}{20}\right)}} = \frac{3}{\left(\sqrt{\frac{252+190}{40}}\right)\left(\frac{1}{22} + \frac{1}{20}\right)}$$
$$= \frac{3}{\frac{\sqrt{442}}{40} \times \frac{42}{440}} = 2.92$$

Degree of freedom = $(22 + 20) - 2 = 40.$

From the t-table, using two-tailed test, (since it is a non-directional hypothesis) tabulated value for 40 degrees freedom is 2.02. Since the calculated value is greater than the tabulated value, we conclude that there is a difference in performance of males and females, with females scoring higher. At 0.01 level the tabulated is 2.70 which is less than the calculated, so we still reject the hypothesis.

Example 2

Using brand P petrol for the mean number of kilometres covered by 22 similar kekomarwa were 52.5km with standard deviation of 7.0. Using brand Q petrol, the mean was 51km with standard deviation of 7.5. Using significance level of 0.05, is there any reason to belief that brand P is better than brand Q?

Solution

The word “better” implies that the hypothesis is directional. So

$$H_0: \mu_1 = \mu_2$$

and the difference is due to chance versus $H_1: \mu_1 > \mu_2$

However, since the number (N) is the same, the formula to be used is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{\sqrt{S_1^2 + S_2^2}}{N}}$$

$$\text{Therefore } t = \frac{52-5-51}{\frac{\sqrt{7^2+7.5^2}}{22}} = \frac{1.5}{\frac{\sqrt{49+56.52}}{22}} = \frac{1.5}{\frac{\sqrt{105.25}}{22}} = \frac{1.5}{2.187} = 0.69.$$

The degree of freedom = $22 + 22 - 2 = 42$

The table value for 40 degrees of freedom at α level of 0.05 is 2.021 and for 60 is 2.00.

However to obtain the one for 42, we interpret as thus:

$$\begin{aligned} \text{Value for 40} + \frac{\text{value for 60} - \text{value for 40}}{60 - 40} \times (42 - 40) \\ = 2.021 + \left(\frac{2.000 - 2.021}{20} \right) \times 2 \end{aligned}$$

However, the calculated value is less than tabulated value and so we do not reject the null hypothesis. We can conclude that there is no difference in the two brands of petrol. But you should note that once a t-value of less than 1 is obtained, the null hypothesis is not rejected since the difference is not significant. If $\bar{x}_1 - \bar{x}_2$ is close to zero, we say the difference is small and this could be caused by random error or by poor sampling. In using t-test, we assume:

that the variable we are using is normally distributed and that our population variances are equal. ($\sigma_1^2 = \sigma_2^2$).

3.2. Hypothesis Test: Difference between Means

This explains how to conduct a hypothesis test for the difference between two means. The test procedure, called the **two-sample t-test**, is appropriate when the following conditions are met:

- The sampling method for each sample is simple random sampling.
- The samples are independent.
- Each population is at least 20 times larger than its respective sample.
- The sampling distribution is approximately normal, which is generally the case if any of the following conditions apply.
 - The population distribution is normal.
 - The population data are symmetric, unimodal, without outliers, and the sample size is 15 or less.
 - The population data are slightly skewed, unimodal, without outliers, and the sample size is 16 to 40.
 - The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

3.2.1. State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The table below shows three sets of null and alternative hypotheses. Each makes a statement about the difference d between the mean of one population μ_1 and the mean of another population μ_2 . (In the table, the symbol \neq means "not equal to".)

3.2.2. Set Null hypothesis Alternative hypothesis Number of tails

1	$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \neq d$	2
2	$\mu_1 - \mu_2 > d$	$\mu_1 - \mu_2 < d$	1
3	$\mu_1 - \mu_2 < d$	$\mu_1 - \mu_2 > d$	1

The first set of hypotheses (Set 1) is an example of a two-tailed test, since an extreme value on either side of the sampling distribution would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are one-tailed tests, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

When the null hypothesis states that there is no difference between the two population means (i.e., $d = 0$), the null and alternative hypothesis are often stated in the following form.

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

3.2.3. Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the two-sample t-test to determine whether the difference between means found in the sample is significantly different from the hypothesized difference between means.

3.2.4. Analyze Sample Data

Using sample data, find the standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

- Standard error. Compute the standard error (SE) of the sampling distribution.

$$SE = \text{sqrt}[(s_1^2/n_1) + (s_2^2/n_2)]$$

where s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- Degrees of freedom. The degrees of freedom (DF) is:

$$DF = (s_1^2/n_1 + s_2^2/n_2)^2 / \{ [(s_1^2/n_1)^2 / (n_1 - 1)] + [(s_2^2/n_2)^2 / (n_2 - 1)] \}$$

If DF does not compute to an integer, round it off to the nearest whole number. Some texts suggest that the degrees of freedom can be approximated by the smaller of $n_1 - 1$ and $n_2 - 1$; but the above formula gives better results.

- Test statistic. The test statistic is a t statistic (t) defined by the following equation.

$$t = [(x_1 - x_2) - d] / SE$$

where x_1 is the mean of sample 1, x_2 is the mean of sample 2, d is the hypothesized difference between population means, and SE is the standard error.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the t Distribution Calculator to assess the probability associated with the t statistic, having the degrees of freedom computed above. (See sample problems at the end of this lesson for examples of how this is done.)

3.2.5. Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Example 3: Two-Tailed Test

Within a school district, students were randomly assigned to one of two Math teachers - Mr. Wilson Ohioze and Mr. Samuel Olusanya. After the assignment, Mr. Wilson Ohioze had 30 students, and Mr Samuel Olusanya had 25 students.

At the end of the year, each class took the same standardized test. Mr. Wilson Ohioze's students had an average test score of 78, with a standard deviation of 10; and Mr. Samuel Olusanyas' students had an average test score of 85, with a standard deviation of 15.

Test the hypothesis that Mr. Wilson Ohioze and Mr. Samuel Olusanya are equally effective teachers. Use a 0.10 level of significance. (Assume that student performance is approximately normal.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: $\mu_1 - \mu_2 = 0$

Alternative hypothesis: $\mu_1 - \mu_2 \neq 0$

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.10. Using sample data, we will conduct a two-sample t-test of the null hypothesis.
- **Analyze sample data.** Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{10^2}{30} + \frac{15^2}{25}} = \sqrt{3.33 + 9} = \sqrt{12.33} = 3.51$$

$$DF = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left\{ \left[\frac{s_1^2}{n_1} \right] / (n_1 - 1) + \left[\frac{s_2^2}{n_2} \right] / (n_2 - 1) \right\}}$$

$$DF = \frac{(10^2/30 + 15^2/25)^2}{\left\{ \left[\frac{10^2}{30} \right] / (29) + \left[\frac{15^2}{25} \right] / (24) \right\}}$$

$$DF = \frac{(3.33 + 9)^2}{\left\{ \left[\frac{(3.33)^2}{(29)} \right] + \left[\frac{(9)^2}{(24)} \right] \right\}} = \frac{152.03}{(0.382 + 3.375)} = \frac{152.03}{3.757} = 40.47$$

$$t = \frac{(x_1 - x_2) - d}{SE} = \frac{(78 - 85) - 0}{3.51} = -7/3.51 = -1.99$$

where s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, n_1 is the size of sample 1, n_2 is the size of sample 2, x_1 is the mean of sample 1, x_2 is the mean of sample 2, d is the hypothesized difference between the population means, and SE is the standard error.

Since we have a two-tailed test, the P-value is the probability that a t statistic having 40 degrees of freedom is more extreme than -1.99; that is, less than -1.99 or greater than 1.99.

We use the t Distribution Calculator to find $P(t < -1.99) = 0.027$, and $P(t > 1.99) = 0.027$. Thus, the P-value = $0.027 + 0.027 = 0.054$.

- **Interpret results.** Since the P-value (0.054) is less than the significance level (0.10), we cannot accept the null hypothesis.

Note: Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, the sample size was much smaller than the population size, and the samples were drawn from a normal population.

Example 4: One-Tailed Test

OgooluwaOlusanya Company has developed a new battery. The engineer in charge claims that the new battery will operate continuously for *at least* 7 minutes longer than the old battery.

To test the claim, the company selects a simple random sample of 100 new batteries and 100 old batteries. The old batteries run continuously for 190 minutes with a standard deviation of 20 minutes; the new batteries, 200 minutes with a standard deviation of 40 minutes.

Test the engineer's claim that the new batteries run at least 7 minutes longer than the old. Use a 0.05 level of significance. (Assume that there are no outliers in either sample.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

$$\begin{aligned} \text{Null hypothesis: } & \mu_1 - \mu_2 \geq 7 \\ \text{Alternative hypothesis: } & \mu_1 - \mu_2 < 7 \end{aligned}$$

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the mean difference between sample means is too small.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a two-sample t-test of the null hypothesis.
- **Analyze sample data.** Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t statistic test statistic (t).

$$\begin{aligned} \text{SE} &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ \text{SE} &= \sqrt{\frac{40^2}{100} + \frac{20^2}{100}} = \sqrt{16 + 4} = 4.472 \end{aligned}$$

$$\begin{aligned} \text{DF} &= \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left\{ \left[\frac{s_1^2}{n_1} \right]^2 / (n_1 - 1) \right\} + \left\{ \left[\frac{s_2^2}{n_2} \right]^2 / (n_2 - 1) \right\}} \\ \text{DF} &= \frac{(40^2/100 + 20^2/100)^2}{\left\{ \left[\frac{40^2}{100} \right]^2 / (99) \right\} + \left\{ \left[\frac{20^2}{100} \right]^2 / (99) \right\}} \\ \text{DF} &= \frac{(20)^2}{\left\{ \left[\frac{16}{99} \right] + \left[\frac{2}{99} \right] \right\}} = 400 / (2.586 + 0.162) = 145.56 \end{aligned}$$

$$t = \frac{(x_1 - x_2) - d}{\text{SE}} = \frac{(200 - 190) - 7}{4.472} = 3/4.472 = 0.67$$

where s_1 is the standard deviation of sample 1, s_2 is the standard deviation of sample 2, n_1 is the size of sample 1, n_2 is the size of sample 2, x_1 is the mean of sample 1, x_2 is the mean of sample 2, d is the hypothesized difference between population means, and SE is the standard error.

Here is the logic of the analysis: Given the alternative hypothesis ($\mu_1 - \mu_2 < 7$), we want to know whether the observed difference in sample means is small enough (i.e., sufficiently less than 7) to cause us to reject the null hypothesis.

The observed difference in sample means (10) produced a t statistic of 0.67. We use the t Distribution Calculator to find $P(t < 0.67) = 0.75$.

This means we would expect to find an observed difference in sample means of 10 or less in 75% of our samples, if the true difference were actually 7. Therefore, the P-value in this analysis is 0.75.

- **Interpret results.** Since the P-value (0.75) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Note: Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, the sample size was much smaller than the population size, and the sample size was large without outliers.

3.3 Confidence intervals for Differences between Two means

Confidence interval is used to determine all reasonably likely values of the difference between two population means. The formula becomes:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{S_{\bar{x}_1 - \bar{x}_2}} \right] \leq \mu_1 - \mu_2 \leq \left[(\bar{x}_1 - \bar{x}_2) + t_{S_{\bar{x}_1 - \bar{x}_2}} \right]$$

Where t is the tabulated value and the error term to be used is given as

$$S_{\bar{x}_1 - \bar{x}_2}$$

Example 5

In a statistics examination for Secondary School, 22 females used in the study has a mean score of 81 and a variance of 12 while the 20 males used had a mean score 78 and a variance of 10. Compute the confidence intervals at 0.05 and 0.01.

Solution

$\bar{x}_1 = 81, \bar{x}_2 = 78, \text{degree of freedom} = 40.$

With tabulated value of 2.02 for the 0.05 level. $S_{\bar{x}_1 - \bar{x}_2}$ has been calculated as 1.03.

Therefore, interval is

$$\begin{aligned} &= \left[(\bar{x}_1 - \bar{x}_2) - t_{S_{\bar{x}_1 - \bar{x}_2}} \right] \leq \mu_1 - \mu_2 \leq \left[\bar{x}_1 - \bar{x}_2 + t_{S_{\bar{x}_1 - \bar{x}_2}} \right] \\ &= [(81 - 78) - 2.02 \times 1.03] \leq \mu_1 - \mu_2 \leq [(81 - 78) + 2.02 \times 1.03] \\ &(3 - 2.08 \leq \mu_1 - \mu_2 \leq 3 + 2.08 \\ &0.92 \leq \mu_1 - \mu_2 \leq 3 + 2.08 \\ &\therefore 0.92 \leq \mu_1 - \mu_2 \leq 5.08 \end{aligned}$$

which is the 95% confidence interval at $\alpha = 0.01$ the 99% confidence interval can be found the same way. For degrees of freedom of 40, the tabulated value is 2.7. Therefore the interval is

$$[(81 - 78) - (2.7 \times 1.03)] \leq \mu_1 - \mu_2 \leq [(81 - 78) + 2.7 \times 1.03]$$

$$3 - 2.78 \leq \mu_1 - \mu_2 \leq (3 + 2.78)$$

$$0.22 \leq \mu_1 - \mu_2 \leq 5.78$$

4.0 CONCLUSION

In this unit, we can conclude that testing the difference of two means of independent samples make us to distinguish between the t calculated of the mean and the students t test.

5.0 SUMMARY

This unit, as vividly takes a look at the test of difference of two means, and its calculation and the confidence intervals for difference of two mean.

6.0 TUTOR MARKED ASSIGNMENT

1. Differentiate between independent sample of difference of two mean and confidence intervals for differences of two mean.
2. In a class, there are 44 females to be used for B.Sc Economic pilot study called proper which have a mean score of 88 and a variance of 16 while in the same class, there are 40 male and with mean score of 82 and a variance of 12. Does the sex of the students have effect on the score of the students at $\alpha = 0.05$ and $\alpha = 0.01$?

7.0 REFERENCES/FURTHER READING

- Datel, R.O. (2013). *Statistics for Business and Management Studies*, 2nd edition, Merrigon Press Company limited.
- Adedayo, O. A (2000), *Understanding Statistics*, JAS publisher Akoka, Lagos

UNIT 5 TESTING DIFFERENCE BETWEEN MATCHED SAMPLES

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Calculation of t Statistic and Variance of the Difference of Scores
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

The test carried out here involves when comparing between means using samples that are dependent. However, suppose we wish to test the hypothesis that in families with two children, the first born is usually more intelligent than the second born. So it should be noted here that the focus on interest is on the mean of two populations/first born and second born but this time the pairs are from the same families and are said to be dependent. The way of carrying out the test of hypothesis is to match pairs of the children with each pair comprising of first born and second born, administer IQ test and report the result in matched pairs are as follows:

2.0 OBJECTIVES

At the end of this unit, you should be able to know how to calculate the mean of the population of different score of t statistic, variance of the difference of scores.

3.0 MAIN CONTENT

3.1 Calculation of t statistic and variance of the difference of scores

Pair	1 st born IQ-Score (\bar{x}_1)	2 nd born IQScore (\bar{x}_2)
1	x_{11}	x_{21}
2	x_{12}	x_{22}
3		
N	x_{1N}	x_{2N}

Another example is a pre test/post test design applied to the same individual. For example, you may wish to find the effect of a drug on asthma patients. For each of the patients, measures are taken of their progress before and after the administration of the drug. Moreover, each person thus serves as his own control and matching is done on the fact that the measured variables come from the same person. For the pair of matched samples the null and alternative hypotheses are given as follows:

$$H_0: \mu_0 = 0 \text{ vs } H_1: \mu_0 \neq 0$$

That is H_0 states that mean of the population of different score is 0 while the alternative states that the mean of the difference is not 0.

Recall formular t statistic:

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

for a single t – test. However, $\bar{x} = \bar{\Delta}$ in this case since difference of mean is used and $\mu = 0$ by virtue of the null hypothesis.

So the statistics become:

$$t = \frac{\bar{\Delta} - 0}{s_{\Delta}/\sqrt{n}} = \frac{\bar{\Delta}}{s_{\Delta}/\sqrt{n}} = \frac{\bar{\Delta}}{\sqrt{s_{\Delta}^2/N}}$$

where $\bar{\Delta}$ is the mean of difference of means, s_{Δ}^2 is the variance of the difference of scores and it's given by

$$s_{\Delta}^2 = \Sigma \frac{(\Delta - \bar{\Delta})^2}{N-1}$$

and N is the number of samples.

Example 1

In an investigation to find out the effect of a diet on a patient's weight, 10 patients were weighed before and after the administration of the drugs. The obtained results are as follows:

Weight After	68	72	71	75	64	74	62	75	82	76
Weight Before	71	64	70	75	67	60	57	69	84	72

You are required to the hypothesis that the mean of difference of means is zero at 95% confidence level and 99% confidence level.

Solution

The hypothesis is $H_0: H_\Delta = 0$ versus $H_1: H_\Delta \neq 0$ with test statistic

$$t = \frac{\bar{\Delta}}{\frac{\sqrt{S_\Delta^2}}{N}}$$

The difference between means is as follows:

Weight After	Weight Before	$\Delta = x_1 - x_2$	$(\Delta - \bar{\Delta})$	$(\Delta - \bar{\Delta})^2$
68	71	-3	-5	25
72	64	+8	6	36
71	70	+1	-1	1
75	75	0	-2	4
64	67	-3	-5	25
74	60	+14	12	144
62	57	-5	-7	49
75	69	+6	4	16
82	84	-2	-4	16
76	72	+4	2	4
		+20		320

The sign of the difference must be retained

$$\bar{\Delta} = \frac{\Sigma \Delta}{N} = \frac{+20}{10} = +2.$$

The variance of the difference of scores

$$S_\Delta^2 = \frac{\Sigma (\Delta - \bar{\Delta})^2}{N-1} = \frac{320}{10-1} = \frac{320}{9} = 35.56.$$

$$t = \frac{\bar{\Delta}}{\frac{\sqrt{S_\Delta^2}}{N}} = \frac{2}{\frac{\sqrt{35.56}}{10}} = 1.06$$

Degrees of freedom = $N - 1 = 10 - 1 = 9$.

- (a) At 95% confidence level, $\alpha = 0.05$ and the tabulated value is 2.26. Since the calculated is smaller than the critical value, we do not reject H_0 and we conclude that the drugs technique has not make any remarkable change in weights of the patients. So we can be 95% confidence that there has been no remarkable change in the means of the weights.

- (b) At 99% confidence level $\alpha = 0.01$ and a value of 3.25 is obtained. The conclusion is the same as for 95% level. That is there is no remarkable change in weight due to administration of the drugs.

SELF ASSESSMENT EXERCISE

Briefly make a clear difference between mean of difference of mean of t statistic and the variance of the difference of scores.

4.0 CONCLUSION

The mean of different of means and the variance of the difference scores has shown that hypothesis can be tested at difference level of confidence that is 95% and 99% level of confidence. However, there is a difference between their calculation and conclusion using the t statistic and variance of mean formular.

5.0 SUMMARY

In this unit, we have tried to do justice to the estimation of the mean of difference of mean and its variance of difference score and how to calculate them using the 99% and 95% confidence level.

6.0 TUTOR MARKED ASSIGNMENT

1. In an investigation to find out the effect of a drug on patient's reaction to the drugs, 20 patients were examined before and after the administration of the drugs.

The following results were obtained:

Drugs After	45	46	63	72	57	66	76	88	84	71
Drugs Before	60	51	50	52	69	84	92	73	77	75

You are required to test the hypothesis that the mean of difference of means is zero at 95% confidence level and 99% confidence level.

2. An experiment was carried out on 20 students to test the effect of a teaching method on their performances in Economics. To this end, a pre test and pro test were administered to the 20 students using teaching methods as treatment. The results of analysis of the matched pair of pre and pro tests for each of the ten students were as follows:

$$\Sigma \Delta = 52, \Sigma (\Delta - \bar{\Delta})^2 = 420.5, N = 20$$

Is there any reason to believe that the teaching method has made any significant improvement in students' performance at
(a) 0.05 level? (b) 0.01 level?

7.0 REFERENCES/FURTHER READINGS

Wesley, H.F. (2010) *Statistics and Economics, a broader approach*, 1st edition, Queror publication limited.
Adedayo, O. A (2000), *Understanding Statistics*, JAS publisher Akoka, Lagos

MODULE 3 SAMPLING THEORY

Unit 1 Population and Sample

Unit 2 Population Parameters

Unit 3 Sampling Parameters

Unit 4 Calculation of Sampling Distribution and Estimators for Mean Variance

Unit 5 Frequency Distribution

UNIT 1 POPULATION AND SAMPLE

CONTENTS

- 1.0 Introduction
- 2.0 Objective
- 3.0 Main Content
 - 3.1 Meaning of Sampling
 - 3.2 Random Samples and Random Numbers
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In statistics, data is collected from a carefully selected sample from the population. However, we are concerned with obtaining the quantifiable characteristics of population known as population parameters rather than the sample statistic. But these population parameters are not easy to compute since we may not be able to obtain values for every unit of the population. A way out is to estimate these from samples. This process is known as parameter estimation. Sampling theory deals with the study of the relationships that exist between a given population and the samples drawn from the population.

2.0 OBJECTIVES

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

- know the meaning of sampling
- understand random samples
- understand random population parameters

3.0 MAIN CONTENT

3.1 Meaning of Sampling

If we draw an object from a pack of population, we have the choice of replacing or not replacing the object into the pack before we draw again. In this first case, a particular object can come up again and again, whereas in the second it can come up only once. Sampling where each member of the population may be chosen more than once is called sampling with replacement, while sampling where each member cannot be chosen more than once is called sampling without replacement.

But it should be noted that finite population that is sampled with replacement can theoretically be considered infinite since samples of any size can be drawn without exhausting the population. For most practical purposes, sampling from a finite population that is very large can be considered as sampling from an infinite population.

3.1.1 Populations

In statistics the term population has a slightly different meaning from the one given to it in ordinary speech. It need not refer only to people or to animate creatures - the population of Nigeria, for instance or the female population of Lagos. Statisticians also speak of a population of objects, or events, or procedures, or observations, including such things as the quantity of lead in urine, visits to the doctor, or surgical operations. A population is thus an aggregate of creatures, things, cases and so on.

Although a statistician should clearly define the population he or she is dealing with, they may not be able to enumerate it exactly. For instance, in ordinary usage the population of Nigeria denotes the number of people within Nigerian's boundaries, perhaps as enumerated at a census. But a physician might embark on a study to try to answer the question "What is the average systolic blood pressure of Nigerian men aged 30-69?" But who are the Nigerian men referred to here? Not all Nigerian men live in Nigeria, and the social and genetic background of those that do may vary. A surgeon may study the effects of two alternative operations for malaria. But one may ask a question that how old are the patients? What sex are they? How severe is their disease? Where do they live? And so on. The reader needs precise information on such matters to draw

valid inferences from the sample that was studied to the population being considered. Statistics such as averages and standard deviations, when taken from populations are referred to as population parameters.

3.1.2 Samples

A population commonly contains too many individuals to study conveniently, so an investigation is often restricted to one or more samples drawn from it. A well chosen sample will contain most of the information about a particular population parameter but the relation between the sample and the population must be such as to allow true inferences to be made about a population from that sample.

Consequently, the first important attribute of a sample is that every individual in the population from which it is drawn must have a known non-zero chance of being included in it; a natural suggestion is that these chances should be equal. We would like the choices to be made independently; in other words, the choice of one subject will not affect the chance of other subjects being chosen

SELF ASSESSMENT EXERCISES

Discuss the use of sampling with replacement and without replacement

3.2 Random Samples and Random Numbers

Clearly, the reliability of conclusions drawn concerning a population depends on whether the sample is properly chosen so as to represent the population sufficiently well, and one of the important problems of statistical inference is just how to choose a sample.

However, one way to do this for finite populations is to make sure that each member of the population has the same chance of being in the sample, which invariably called a random sample. Random sampling can be accomplished for relatively small populations by drawing lots, or it equivalently, by using a standard table of random numbers. But it should be noted that it is normally constructed for such purposes.

Finally, because inferences from sample to population cannot be certain, we must use the language of probability in any statement of conclusion.

SELF ASSESSMENT EXERCISE

Differentiate between a Random samples and Random numbers.

4.0 CONCLUSION

In this unit, we have discussed about population and sample. However, we can conclude that population is the entire pool from which a statistical sample is drawn. The information obtained from the sample allows statisticians to develop hypotheses about the larger population. Researchers gather information from a sample because of the difficulty of studying the entire population. In statistical equations, population is usually denoted with a capital 'N', while the sample is usually denoted with a lowercase 'n'.

5.0 SUMMARY

This unit has really enabled us to do justice to population and sample. However, Population sampling is the process of taking a subset of subjects that is representative of the entire population. The sample must have sufficient size to warrant statistical analysis.

6.0 TUTOR MARKED ASSIGNMENT

Differentiate between the following;

1. Population and a sample
2. Random Samples and Random Numbers

7.0 REFERENCES/FURTHER READINGS

Ademisan, T.Y. (2011) *Introduction to Statistics, a contemporary issue*, 1st edition, Mill world Publication limited.

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

UNIT 2 POPULATION PARAMETERS

CONTENTS

- 1.0 Introduction
- 2.0 Objective
- 3.0 Main Content
 - 3.1 Analysis of Population Parameters
 - 3.2 Sample Statistics
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

A population is considered to be known when we know the probability distribution $f(x)$ (probability function or density function) of the associated random variable x . For Example, if x is a random variable whose values are the number of defective machines found during a given 6-day in a week, then x has probability distribution $f(x)$.

2.0 OBJECTIVES

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

- analyse the population parameter
- understand sample statistics.

3.1 Analysis of Population Parameters

If x is normally distributed, we say that the population is normally distributed or that we have a normal population. More so, if x is binomially distributed, we say that the population is binomially distributed or that we have a binomial population.

However, there will be certain quantities that appear in $f(x)$, such as μ and σ in the case of the normal distribution or p in the case of the binomial distribution. However, other quantities such as median, mode and skewness can be determined in terms of these and all such quantities are called population parameters. But note that when we are given the population so that we know $f(x)$, then the population parameters are also known.

More so, when a problem arises when the probability distribution $f(x)$ of the population is not known precisely, although we may have some idea of it, or at

least be able to make some hypothesis concerning the behaviour of $f(x)$ e.g. if we may have some reason to suppose that a particular population is normally distributed, in this case the values of μ and σ and so we might wish to draw statistical inferences about their applications.

3.2 Sample Statistics

In this analysis, we may take a random samples from the population and then use these samples to obtain values that serve to estimate and test hypothesis about the population parameters.

Let us consider an example where we wish to draw conclusion about the heights of 24,000 adult students by examining only 200 students selected from the population. In this case, x can be random variable whose values are the various heights. To obtain a sample of size 200, we must first choose one individual at random from the population. However, this individual can have any one value, say x_1 of the various possible heights and we can call x_1 the value of a random variable X_1 where the subscript 1 is used since it corresponds to the first individual for the sample, who can have any one of the values x_1 of the possible heights and x_2 can be taken as the value of a random variable X_2 . We continue this process up to X_{200} since the sample is size 200. Let us assume that the sampling is with replacement so that the same individual could conceivably be chosen more than once. In this vein, the sample size is much smaller than the population size, sampling without replacement would give practically the same results as sampling with replacement.

More so, a general case where a sample size n will be described by the values x_1, x_2, \dots, x_n of the random variables X_1, X_2, \dots, X_n . This case of sampling with replacement, X_1, X_2, \dots, X_n would be independent, identically distributed random variables having probability function $f(x)$ and their joint distribution is as follows:

$$P(x_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = f(x_1)f(x_2) \dots f(x_n)$$

Any quantity obtained from a sample for the purpose of estimating a population parameter is called a sample statistic. If we subject it to calculation, a sample statistics for a sample of size n can be defined as a function of the random variables X_1, X_2, \dots, X_n that is $g(X_1, \dots, X_n)$. The function $g(X_1, \dots, X_n)$ is another random variable, whose values can be denoted by $g(X_1, \dots, X_n)$.

SELF ASSESSMENT EXERCISES

With clear explanation, differentiate between population parameter and sample statistics.

4.0 CONCLUSION

Population parameter is known when the probability distribution of say $f(x)$ is known of associated random variable X while sample statistic is any quantity obtained from a pooled of samples for the purpose of estimating a population parameter.

5.0 SUMMARY

This unit has done justice to the analysis of population parameter, its expansion and how sample statistics is used in estimating population parameter for a random variable.

6.0 TUTOR MARKED ASSIGNMENT

With the aid of explanation and examples of parameters estimation, write short note on the following:

- (a) Population parameter
- (b) Sample statistic
- (c) Randomness of a variable say $f(x)$
- (d) Sample in a pooled population.

7.0 REFERENCES/FURTHER READINGS

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

Murray, S., & John, S., & Alu, S. (2001) *Probability and Statics*, Schaum's easy outlines, Macgrw Hill Publication Company, New York.

UNIT 3 SAMPLING PARAMETERS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Sampling Error
 - 3.1.1 Sampling Distribution
 - 3.2 Sampling Distribution of the Mean
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Reading

1.0 INTRODUCTION

Sample parameters are those parameters that are used in estimating variables of selected population parameters. In this unit we shall take a look at some of those sampling parameters and their calculation.

2.0 OBJECTIVES

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

- Understand the meaning of sampling Error
- Understand sampling distribution
- Understands sampling distribution of mean.

3.0 MAIN CONTENT

3.1 Sampling Error

The values of a population parameter and that of the corresponding statistic are not always the same. If a difference occurs this difference is known as a sampling error. Consequently, sampling error (E) is defined as the difference between the sample statistic (s) and the population parameter being estimated (P). That is: $E = S - P$. However, if the parameters are under estimated, the sampling errors are negative errors while if they are over estimated, the sampling errors are positive errors.

3.1.1 Sampling Distribution

The value of a statistic obtain vary from one sample to another even when equal samples are selected from the same population using the same procedure. However, the statistics obtained from repeated selections, when estimated and organised into relative frequency distribution form a sampling distribution.

A sampling distribution is the set of all possible values of a particular statistic and you should note that there is sampling distribution of means, sampling distribution of variance, etc. For each type of this sampling distribution, one can compute the mean, variance, standard deviation, etc. Therefore, we can have mean and standard deviation of sampling distribution of means, variances, etc. Note that the standard deviation of the sampling distribution is known as the standard error.

SELF ASSESSMENT EXERCISES

Explain in detail, the meaning of sampling error and is it differs from sampling Distribution?

3.2 Sampling Distribution of the Mean

Assuming we draw n repeated independent samples of data from a population of size N and then calculates, the mean of each of these samples. Assuming the means of the samples are represented by $(X_1, X_2, \dots \dots X_n)$. The frequency distribution of the sample mean is called sampling distribution of the mean. Moreover, the mean of this distribution $\mu_{\bar{x}}$, when computed will be equal to the mean of the N population (μ) i.e.

$$\mu_{\bar{x}} = \frac{\bar{X}_1 + \bar{X}_2 + \dots \dots + \bar{X}_n}{n} = \mu$$

Therefore this means that the expectation of the sample mean is the same as the population mean. The formular above hold whether or not sampling is with replacement. Also the standard deviation of the sampling distribution of the mean is known as standard error of the mean and this is given as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

With the variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ if there is replacement and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

when the sampling is without replacement. If the population variance σ^2 is not known, we make use of the estimate below:

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

Example

Given a population consisting of numbers 1, 2, 3, 4, 5.

1. Write out all the possible samples of size two that can be obtained without replacement from the population.
2. Calculate the mean of the population.
3. Calculate the mean of the sampling distribution of mean. What do you observe?

Solution

If the samples of size two are selected without replacement, then we expect

$${}^5C_2 = \frac{5!}{3!2!}$$

Samples

(a) The samples are:

(1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5), (3,4), (3,5), (4,5),

(b) The mean of the population is:

$$\frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

(c) The mean of each samples

For (1,2) the mean \bar{x}_1 is $\frac{1 + 2}{2} = 1.5$

For (1,3) the mean \bar{x}_2 is $\frac{1 + 3}{2} = 2.0$

For (1,4) the mean \bar{x}_3 is $\frac{1 + 4}{2} = 2.5$

For (1,5) the mean \bar{x}_4 is $\frac{1 + 5}{2} = 3.0$

For (2,3) the mean \bar{x}_5 is $\frac{2+3}{2} = 2.5$

For (2,4) the mean \bar{x}_6 is $\frac{2+4}{2} = 3.0$

For (2,5) the mean \bar{x}_7 is $\frac{2+5}{2} = 3.5$

For (3,4) the mean \bar{x}_8 is $\frac{3+4}{2} = 3.5$

For (3,5) mean \bar{x}_9 is $\frac{3+5}{2} = 4.0$

For (4,5) mean \bar{x}_{10} is $\frac{4+5}{2} = 4.5$

Therefore, mean of sampling distribution of mean

$$\frac{1.5 + 2.0 + 2.5 + 3.0 + 2.5 + 3.0 + 3.5 + 3.5 + 4.0 + 4.5}{10}$$

$$\mu_{\bar{x}} = \frac{30}{10} = 3.$$

Thus shows that $\mu_{\bar{x}} = \mu$ as we mentioned earlier.

SELF ASSESSMENT EXERCISES

Suppose the population consists of bus fares given to three children in a day. If the bus fares are ~~₦~~20, ~~₦~~40 and ~~₦~~60 respectively,

- Write out all possible samples of size two that can be drawn from this population with replacement.
- Calculate the mean of the sampling distribution of means obtained.
- Compare this value with the population mean.
- Find the variance and standard deviation of the population.
- Compute the standard error of means and also the variance of means. What do you observe?

4.0 CONCLUSION

The sampling parameter has been examined dealing with the sampling error, sampling distribution and its mean. However, examples of how to calculate the mean and the variance of the sampling distribution were also examined in this unit.

5.0 SUMMARY

In this unit, we have been able to analyse the issue in sampling parameters and its calculation varying from the mean, variance and the standard deviation.

6.0 TUTOR MARKED ASSIGNMENT

1. Write short note on the following:
 - (a) Sampling error
 - (b) Sampling Distribution
 - (c) Sampling Distribution of the mean
2. Given a population of numbers 2, 4, 6, 8, 10,
 - (a) Write out all the possible samples of size two that can be obtained without replacement, from this population.
 - (b) Calculate the mean of the population and the mean of sampling distribution of mean.

7.0 REFERENCES/FURTHER READINGS

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

Murray, S., & John, S., & Alu, S. (2001) *Probability and Statics*, Schaum's easy outlines, Macgrw Hill Publication Company, New York.

UNIT 4 CALCULATION OF SAMPLING DISTRIBUTION AND ESTIMATORS FOR MEAN VARIANCE

CONTENTS

- 1.0 Introduction
- 2.0 Objective
- 3.0 Main Content
 - 3.1 Estimators for Mean and Variance
 - 3.2 Sampling Distribution of Proportion
 - 3.3 Sampling Distribution of Differences and Sums
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

The calculation of sampling distribution is a step forward to look at different ways of obtaining distribution of proportion process of pooled data, sampling distribution of Differences and Sum. Moreover, analysing the estimators of mean and variance also gives us an insight to analyse the different mean score of say students in a class and its standard deviation from a population of large class of students.

2.0 OBJECTIVES

At the end of this unit you will be able to:

- estimate the mean and variance from a population
- calculate the sampling distribution of proportion
- calculate the sampling distribution of Differences and sums

3.0 MAIN CONTENT

3.1 Estimators for Mean and Variance

Assuming we select a sample of size n from a normally distributed population variance, how do you think we estimate μ and σ^2 ? The answer to this question is: the sample mean \bar{X} is an unbiased estimator of population mean μ while the sample variance σ^2 will be the unbiased estimator of σ^2 . Therefore, the unbiased estimator of the variance given as:

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

However, if we select a sample of size n from a normal population with mean μ and its variance σ^2 , then the sample mean follows a normal distribution analysis.

$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ since, as we have discussed in some of the unit in this course, the variance then become $\frac{\sigma^2}{n}$. The formula that can be taken as the standard form of sampling distribution of mean is therefore given as:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example

The mean mark of students in statistics test is 68 with standard deviation of 20. If samples consisting of 64 students each are obtained from the students population of 6,000, estimate the mean and standard deviation of the sampling distribution of mean if sampling is done with replacement.

Solution

Population mean $\mu = 68$

Since the sample mean is an unbiased estimator of the population mean, the mean of the sampling distribution is:

$$E(\bar{x}) = \mu$$

Therefore, $E(\bar{x}) = 68$

From the question, the selection of sample was with replacement the standard deviation of the sampling distribution becomes:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{64}} = \frac{20}{8} = 2.5$$

SELF ASSESSMENT EXERCISE

1. Show that the sampling distribution of mean:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

2. Given that the mean of a sample of 18 observation from a normal distribution of mean 100 and variance 64 is \bar{x} . Compute (i) $P(\bar{x} < 86)$ (ii) $P(76 < \bar{x} < 86)$.

3.2 Sampling Distribution of Proportion

The analytical formulae for the sampling distribution of proportion is given as

$$\delta_p^2 = \frac{P(q)}{n} = \frac{P(1-P)}{n}$$

P is the population proportion and $q = 1 - p$, and if the sample estimator is \hat{P} , the estimation of the variance will be

$$\text{Var}(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n}$$

The sample proportion is also an unbiased estimator of the population proportion.

Example

If 150 tosses are made of a fair coin, find the probability that between 38% and 58% will be heads.

Solution

The 150 tosses is a sample from infinite population of all possible tosses of the coin. However, since the coin is fair,

$$P = \frac{1}{2} \text{ and } q = 1 - P = \frac{1}{2}$$

$$38\% \text{ of tosses} = \frac{38}{100} \times 150 = 57 \text{ while } 58\% = 87$$

$$\mu = \text{expected number of heads} = np = 150 \times \frac{1}{2} = 75.$$

$$\delta = \sqrt{npq} = \sqrt{75} \times \frac{1}{2} = 6.12$$

$$\therefore P(-57 < \hat{P} < 87) = P\left(\frac{57 - 75}{6.12} < Z < \frac{87 - 75}{6.12}\right)$$

$$= P(-2.94 < Z < 1.96) = 0.5596 - 0.0053 \\ = 0.5543.$$

SELF ASSESSMENT EXERCISE

Differentiate between the formulae of sampling distribution of proportion and estimators for mean and variance.

3.3 Sampling Distribution of Differences and Sums

If we are given two normally independent random variables x_1 and x_2 whose means are μ_{P_1} and μ_{P_2} respectively and variances σ_{P_1} and σ_{P_2} , then the sampling distribution of difference:

- (i) Is a normal distribution
- (ii) Has mean equal to the difference of mean of the two variables i.e.

$$\mu_{P_1-P_2} = \mu_{P_1} - \mu_{P_2}$$

- (iii) Has a variance equal to the sum of the two variances i.e.

$$\sigma^2_{P_1-P_2} = \sigma^2_{P_1} + \sigma^2_{P_2}$$

For example, the sampling distribution of the difference of mean is

$$\text{and } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

when n_1 and n_2 are the samples sizes drawn from the two population and δ_1^2 and δ_2^2 the variances of the two populations.

More so, the sampling distributions of differences in proportion from two distributions and binomially distributed population is

$$\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = \hat{p}_1 - \hat{p}_2.$$

where \hat{p}_1 and \hat{p}_2 are proportion of successes.

Furthermore, the sampling distribution of sum are defined as

$$\mu_{\hat{p}_1 + \hat{p}_2} = \mu_{\hat{p}_1} + \mu_{\hat{p}_2} = \hat{p}_1 + \hat{p}_2$$

and

$$S_{\hat{p}_1 + \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{p}_2}{n_1} + \frac{\hat{p}_1 \hat{p}_2}{n_2}}$$

where \hat{p}_1 and \hat{p}_2 are proportion of successes

Furthermore the sampling distribution of sum are defined as:

$$\mu_{P_1+P_2} = \mu_{P_1} + \mu_{P_2} \text{ and } \sigma^2_{P_1+P_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example

Suppose you are given two populations with $P_1 = (40,60)$ and $P_2 = (50,90)$. Show that

(a) $\mu_{P_1+P_2} = \mu_{P_1} + \mu_{P_2}$ (b) $\mu_{P_1-P_2} = \mu_{P_1} - \mu_{P_2}$

(c) $\sigma^2_{P_1+P_2} = \sigma^2_{P_1} + \sigma^2_{P_2}$ for a sample drawn from each of the population.

Solution

Sum of sampling

The possible combinations of samples are $(40,50)$, $(40,90)$, $(60,50)$, and $(60,90)$.

\therefore sum of examples is $40 + 50$; $40 + 90$; $60 + 50$; $60 + 90$.
Sum of samples $90, 130, 110, \text{ and } 140$.

Therefore, mean of sum of samples is given by

$$\mu_{P_1+P_2} = \frac{90 + 130 + 110 + 150}{4} = \frac{480}{4} = 120.$$

For $(40,60)$

$$\mu_{P_1} = \frac{40 + 60}{2} = 50$$

For $(50,90)$

$$\mu_{P_2} = \frac{50 + 90}{2} = 70$$

$$\therefore \mu_{P_1} + \mu_{P_2} = 50 + 70 = 120 = \mu_{P_1+P_2}$$

The difference of samples

$$\therefore \mu_{P_1} - \mu_{P_2} = 50 - 70 = -20$$

Samples for differences of samples give us $40 - 50$; $40 - 90$; $60 - 50$; $60 - 90$

The sum becomes: $-10 + (-50) + 10 + (-30)$
 $= -10 - 50 + 10 - 30 = -80$

$$\mu_{P_1-P_2} = \text{mean} = \frac{-80}{4} = -20$$

Therefore

$$\mu_{P_1-P_1} = \mu_{P_1} - \mu_{P_2}$$

$$\sigma_{P_1+P_2}^2 = \text{variance of } 90,130,110, \text{ and } 150.$$

$$\sigma_{P_1+P_2}^2 = \frac{\Sigma(x-\bar{x})^2}{N}$$

$$\sigma_{P_1+P_2} = \frac{1}{4} [(90 - 120)^2 + (130 - 120)^2 + (110 - 120)^2 + (150 - 120)^2] = 500$$

$$\therefore \sigma_{P_1}^2 = \text{variance of } (40,60) = \frac{1}{2} [(40 - 50)^2 + (60 - 50)^2] = \frac{1}{2} (100 + 100) = 100$$

$$\sigma_{P_2}^2 = \frac{[(50 - 70)^2 + (90 - 70)^2]}{2} = 400$$

$$\sigma_{P_1}^2 + \sigma_{P_2}^2 = 100 + 400 = 500$$

$$\sigma_{P_1+P_2}^2 = 500 = \sigma_{P_1}^2 + \sigma_{P_2}^2$$

SELF ASSESSMENT EXERCISES

Suppose you are given two populations with $P_1 = (20,40)$ and $P_2 = (70,90)$. Show that

$$(a) \quad \mu_{P_1+P_2} = \mu_{P_1} + \mu_{P_2}$$

$$(b) \quad \mu_{P_1-P_2} = \mu_{P_1} - \mu_{P_2}$$

$$(c) \quad \sigma_{P_1+P_2}^2 = \sigma_{P_1}^2 + \sigma_{P_2}^2$$

4.0 CONCLUSION

We have vividly done justice to how to calculate; the estimators for mean and variance, sampling distribution of proportion plus sampling distribution of differences and sums. So at this point you should be able to work through this analysis mentioned above.

5.0 SUMMARY

In this unit, we have learnt on how to estimate and make a comparison between sampling distribution of differences and sums in one hand and how to estimate mean and variance.

6.0 TUTOR MARKED ASSIGNMENT

1. Writes short notes on the following:
 - (a) Estimators for mean and variance

- (b) Sampling Distribution of proportion
 - (c) Sampling Distribution of differences and sums
 - (d) Simple Random Sampling
 - (e) Sampling Error
2. The proportion of Boarders in two Federal Government Colleges for two consecutive years are: $\left(\frac{1}{3}, \frac{3}{4}\right)$ and $\left(\frac{1}{4}, \frac{3}{5}\right)$

If a sample is selected from each of the two populations, prove that:

$$\mu_{P_1 - P_2} = \mu_{P_1} - \mu_{P_2}$$

$$(a) \sigma_{P_1 - P_2} = \sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2}$$

The heights of two men were measured as 1.6 metres and 1.62 metres with standard errors of 0.04 metres and 0.015 metres respectively. Determine the mean and standard deviation of

- (a) The sum
- (b) The difference of the heights

7.0 REFERENCES/FURTHER READINGS

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

UNIT 5 FREQUENCY DISTRIBUTION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Analysis of Frequency Distribution
 - 3.2 Frequency Polygon
 - 3.3 Relative Frequency Distributions
 - 3.4 Calculation of Ungrouped Data for Frequency Distributions
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

If a sample (or even a population) is large, it is difficult to observe the various characteristics or to compute statistics such as mean and standard deviation. However, for this reason, it is useful to organize or group the raw data.

2.0 OBJECTIVES

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

- analyse the frequency distribution and frequency polygon
- know the relative frequency distributions and understand the calculation of ungrouped data for frequency distributions.

3.0 MAIN CONTENT

3.1 Analysis of Frequency Distribution

For example, suppose that a sample consists of the heights of 100 male students at Propero University. We can then arrange the data into classes or categories and determine the number of individuals belong to each class called a frequency distribution or frequency table.

Table showing the Heights of 100 male students at Propero University.

Table 1

Height (inches)	Number of Students
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8
Total	100

Moreover, the first class or category for example consists of height from 60 to 62 inches, indicated by 60 – 62, is called class interval. However, since 5 students have heights belonging to this class, therefore the class frequency is 5. Since the height that is recorded as 60 inches is actually between 59.5 and 60.5 inches, and we can go on to record the class interval as 59.5 – 60.5. The next class interval would then be 60.5 – 61.5, etc. However, the class interval 59.5 – 60.5, the number 59.5 and 60.5 are often called class boundaries. The width of the j th class interval, denoted by M_j , which is usually the same for all classes (in which cases it is denoted by m) and is the difference between the upper and lower class boundaries, that is $m=60.5 - 59.5=1$.

Moreso, the midpoint of the class interval, which can be taken as representative of the class is called mark corresponding to the class interval 60 – 62 is 61.

SELF ASSESSMENT EXERCISES

Discuss the meaning of frequency distribution in detail.

3.2 Frequency Polygon

A graph for frequency distribution can be supplied by a histogram or by a polygon graph often called a frequency polygon that connect the midpoints of the tops in the histogram. The graph below shows the graph seems to indicate that the sample is drawn from a population of heights that is normally distributed.

Fig 1

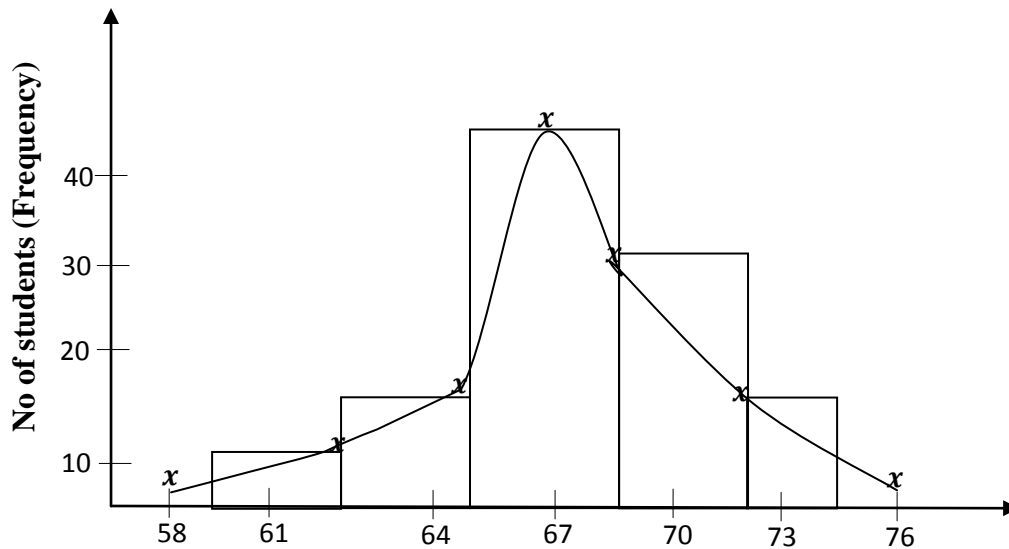


Figure 1 Showing Frequency Polygon

SELF ASSESSMENT EXERCISES

Discuss in detail the term “a Frequency Polygon”

3.3 Relative Frequency Distributions

If we recorded the relative frequency or percentage rather than the number of students in each class, the result would be what we can call relative or percentage frequency distribution. For example the relative or percentage frequency corresponding to the class 63-65 is $\frac{18}{100}$ or 18%.

Furthermore, the corresponding histogram is similar to that of the table 1 in unit 3.1 except that the vertical axis is relative frequency instead of frequency and the sum of the rectangular area can then be 1 or 100%. We can consider a relative frequency as a probability distribution in which probabilities are replaced by relative frequencies. But in statistical analysis, relative frequencies can be thought of as empirical probabilities, and relative frequency distributions is also known as empirical probability distributions.

SELF ASSESSMENT EXERCISES

Explain with detailed examples relative frequency distributions

3.4 Calculation of Ungrouped Data for Frequency Distributions

The unorganized data collected during investigation is known as raw data. Suppose a lecturer records raw scores of 50 students in a statistics test as follows:

58	15	81	79	92	58	69	32	45	56
41	85	43	52	61	75	85	69	56	49
57	87	89	49	85	45	69	75	65	61
25	72	67	58	84	60	32	57	69	68
73	42	65	55	74	58	36	78	68	79

When this data is arranged in ascending or descending order, we have an array of data as follows.

15	25	32	32	36	41	42	43	45	45
49	49	52	55	56	56	57	57	58	58
58	58	60	61	61	65	65	67	68	68
69	69	69	69	72	73	74	75	75	78
79	79	81	84	85	85	85	87	89	92

However, the array we can easily identify the highest as 92 and the lowest as 15. The difference between these two numbers is known as the range that is the range = Highest value – lowest value = $92 - 15 = 77$.

We can also arrange the table into categories or groups or classes. Normally, we expect that the classes should be between 5 and 20. The scores of the students can be tabulated using classes 11-20, 21-30, etc. We then use tallies to form the table. Tallies are strokes used for counting and a value of 5 tallies is denoted by 4 vertical strokes and one diagonal stroke as **HHH**. This is to facilitate ease for counting. Therefore, the students' score we obtain is as follows:

Classes	Tally	Frequency
11 – 20	I	1
21 – 30	II	2
31 – 40	III	3
41 – 50	HHH II	7
51 – 60	HHHHHH I	11
61 – 70	HHHHHH I	11
71 – 80	HHH III	8
81 – 90	HHH II	7
91 – 100	I	1

SELF ASSESSMENT EXERCISE

The following are the scores of 20 students in a test

72	59	58	60	59
53	82	68	74	63
59	52	50	88	78
57	80	85	90	60

Using intervals 50 – 54, 55 – 59, etc, draw a frequency table.

4.0 CONCLUSION

In this unit, we have discuss about frequency distribution and we deduced from the analysis above that a frequency distribution is a representation, either in a graphical or tabular format, which displays the number of observations within a given interval. The intervals must be mutually exclusive and exhaustive. Frequency distributions are usually used within a statistical context.

5.0 SUMMARY

This unit has vividly takes a look at frequency distribution and various calculations on the said topic was examined and frequency polygon was also discussed. This unit would have put you the thorough teaching of frequency distribution.

6.0 TUTOR MARKED ASSIGNMENT

1. The following are the scores of 20 students in an examination

66	52	70	61	79
64	82	72	63	81
43	44	84	56	40
42	41	88	59	45

2. Choose a class interval and draw a frequency table.
3. The following are the marks of 20 students in a statistics examination in Babatunde Olusegun Olusanya college;

36	22	27	71	49
34	52	62	73	41
55	66	67	36	20
42	41	65	59	45

7.0 REFERENCES/FURTHER READING

Adedayo, O. A (2000). *Understanding Statistics*, JAS publisher Akoka, Lagos

Faraday, D.F. (2009). *Statistics for Business Management*, 1st edition, Junit Press limited, Lagos.

MODULE 4 T-TEST, F-TEST AND CHI SQUARE ANALYSIS

CONTENTS

UNIT 1 T-TEST

UNIT 2 F-TEST

UNIT 3 CHI SQUARE TEST

UNIT ONE T TEST

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 History of T Statistics
 - 3.2 Unpaired and paired two Sample t tests.
 - 3.3 The t test formula
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

A t-test is any statistical test in which the test statistic follows a student's t distribution if the null hypothesis is supported. It is also called students t test in the name of its founder "students". T test is used to compare two different set of values. It is generally performed on a small set of data. T test is generally applied to normal distribution which has a small set of values. The test compares the mean of two samples. T test uses mean and standard deviations of two samples to make comparison.

2.0 OBJECTIVES

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

- Understand the history of t statistics
- know the unpaired and paired two sample t tests.
- Understand the t test formula

3.0 MAIN CONTENT

3.1 History of T Statistic

The T statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness Brewery in Dublin, Ireland (“Student” was his pen name). However, Gosset had been hired due to Claude Guinness’s policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness’s industrial processes.

Gosset devised the t-test as a cheap way to monitor the quality of stout. The student t-test work was submitted to and accepted in the Journal *Biometrika*, the Journal that Karl Pearson had co-founded in 1908.

Company policy at Guinness forbids its chemists from publishing their findings, so Gosset published his mathematical work under the name “Student”. Guinness had a policy of allowing technical staff leave for study (popularly called study leave), which Gosset used during the first two terms of the 1906-1907 academic year in Professor Karl Pearson’s Biometric Laboratory at University College London. Furthermore, Gosset’s identity was then known to fellow statisticians and Editor-in-Chief Karl Pearson. It is not clear how much of the work Gosset performed while he was at Guinness and how much was done when he was on study leave at University College, London.

3.1.1 Uses of T-Test

T-test is used for the following:

- (a) A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.
- (b) A two-sample location test of the null hypothesis that the means of two populations are equal. All such tests are usually called Student’s tests. It should be noted that that name should only be used if the variances of the two populations are also assumed to be equal: the form of the test used when this assumption is dropped is sometimes called WELCH’S TEST. These tests are often referred to as “unpaired” or “independent samples” t-tests, as they are typically when the statistical units underlying the two samples being compared are non-overlapping.
- (c) A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, assuming we measure the size of a breast cancer patient’s tumour before and after a treatment. If the treatment is effective, we can only expect the tumour size for many of the patients to be smaller

following the treatment and it is called a paired or repeated measures t-test.

- (d) A test of whether the slope of a regression line differs significantly from D.

3.1.2 Assumptions

The assumptions of a t-test are as follow:

- (a) z follows a standard normal distribution under the null hypothesis.
- (b) S^2 follows a λ^2 distribution (chi-square) with p degrees of freedom under the null hypothesis, where p is positive constant.
- (c) z and s are independent
However, in a specific type of t-test, these conditions are consequences of the population being studied, and of the way in which the data are sampled. For instance, the t-test comparing the means of two independent samples, the following assumptions should be met.
- (d) Each of the two populations being compared should follow a normal distribution, and this can be tested using a normality test, or it can be assessed graphically using a normal quartile plot.
- (e) If using Student's original definition of the t-test, the two populations being compared should have the same variance (using F-test or assessable graphically using a $Q - Q$ plot), but if the sample size in the two groups being compared are equal, student's original t-test is highly the best to the presence of unequal variances.
- (f) The data used to carry out the test should be sampled independently from the two populations being compared.

SELF ASSESSMENT EXERCISE

List the uses and assumptions of t-test statistics.

3.2 Unpaired and Paired Two-Sample T-Tests

Two-sample t-tests for a difference in mean involve independent samples and overlapping samples. The paired t-tests are of form of blocking and have greater power than unpaired tests when the paired units are similar with respect to noise factors that are independent of membership in the two groups being compared. But you should note that the paired t-test can be used to reduce the effects of confounding factors in an observational study.

3.2.1 Independent (Unpaired) Samples

The independent samples t-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. For example, suppose we are evaluating the effect of malaria outbreak and we enrol 200 subjects to the treatment group and 100 subjects to the control group. In this situation, we have two independent samples and would use the unpaired form of the t-test. The randomization is not essential here that is if we contacted 200 people by phone and obtained each person's age and gender and then used a two-sample t-test to see whether the mean ages differ by gender, this would also be an independent samples t-test, even though we can see that the data are observational.

3.2.2 Paired Samples

Paired samples t-tests consist of a simple of matched pairs of similar units, or one group of units that has been tested twice which sometimes we call repeated measures t-test. An example of the repeated measures t-test would be where subjects are tested again after treatment with a headache lowering medication. By comparing the same patient's numbers before and after treatment, we are effectively using each patient as their own control. That way, the correct rejection of the null hypothesis that is of no difference made by the treatment and can become much more likely with statistical power increasing because the random between patient variations has now been eliminated. In this analysis, each analysis is half way that is each paired half depends on the other paired half, therefore the version of student's t-test has only $n/2 - 1$ degrees of freedom where n is the numbers of observations and the pairs then becomes the individual test units and the sample has to be doubled to achieve the same number of degrees of freedom.

Furthermore, a paired samples t-test is based on a "matched pairs sample" results from an unpaired sample that is used to make up a paired sample by using more variables that are measured with the variable of interest.

Matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of the other measured variables and it is used in observational studies to reduce or

eliminate the effects of – s confounding factors. Finally, it should be noted that paired sample tests is also called “dependent sample t-tests”.

3.3 The T-Test Formula

The T test formula is given as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where \bar{x}_1 = mean of first set values

\bar{x}_2 = mean of second set values

S_1 = standard deviation of first set of values

S_2 = standard deviation of second set of values

n_1 = total number of values in first set

n_2 = total number of values in second set

The formulae for standard deviation is given by:

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

Where x = values given

\bar{x} = mean

n = total number values

Example

***** the t-test value for the following two-sets of values

Solution

Formulae for mean

$$\bar{x} = \frac{\Sigma x}{n}$$

Formulae for standard deviation

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

Calculate for first set: Number of items in first set, $n_1 = 4$

Mean for the first set or data, $x_1 = 6.5$

SELF ASSESSMENT EXERCISES

Differentiate between paired and unpaired samples

4.0 CONCLUSION

We can conclude in this unit that the test statistic in the t-test is known as the t-statistic and the t-test looks at the t statistic, t-distribution and degree of freedom to determine the p-value (probability) that can be used to determine whether the population differ.

5.0 SUMMARY

The unit has seriously dealt with the analysis of t-test statistic and how it is used in paired and unpaired samples. However, the uses and the assumptions of t-test was also examined and at this junction I belief you must have understand the meaning and how t-test is used in a different situation.

6.0 TUTOR ASSESSMENT ASSIGNMENT

1. In what percentage increase is paired samples differ from unpaired samples of a t-test statistics
2. Discuss in detail the assumption of t-test statistics
3. Find the t-test value for the following two set of data

x_1	9	10	11	12
x_2	2	4	6	8

4. Given the following hypothesis bellow;

Null hypothesis: Good teaching has no significant effect on Students performance.

Alternative hypothesis: Good teaching has a significant effect on Students performance.

Using 5% level of significance, and the t calculated is 1.9442. calculate the t statistics and accept/reject the null or alternative hypothesis.

7.0 REFERENCES/FURTHER READINGS

Richard, M, O. (2001). *The Analysis of Understanding Statistics*, 2nd Edition, Millworld Publication Limited.

Sawilowsky, S. (2005). Misconceptions Leading to Choosing the t-Test over the Wilcoxon Mann – Whitney U Test for Shift in Location Parameter, *Journal of Modern Applied Statistical Method* 4(2) 598-600.

UNIT 2 F TEST

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Examples of F tests
 - 3.2 Formulae and analysis of f test Statistics
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

An F test can be defined as any statistical test in which the test statistics has an F distribution under a null hypothesis situation and it is usually used when comparing statistical models in a data set so that we can identify the mode that best fits the population where the date were sampled. F test arises when we have a model that has been fitted to a data of least square method. F Test was coined by George, W. Snedeor but he used it to honour Sir Ronald A. Fisher and Fisher initially developed the statistics as the variance ratio in 1920.

2.0 OBJECTIVES

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

- Know the various examples of f tests Statistics
- Understand formulae and analysis of f tests Statistics

3.0 MAIN CONTENT

3.1 Examples OF F-Tests

The common examples of F tests in a statistical analysis are as follows:

- (a) Looking at the situation where the hypothesis that the means of given set of normally distributed populations that all the parameters having the same standard deviation are equal. We can call this the best known F test commonly used in statistical test and it plays a key role in the analysis of variance (ANOVA) that was dealt with in statistics for Economist 1.
- (b) In the case where the hypothesis that a proposed regression model actually fit the data to be analysed.

- (c) Vividly looking at the hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested with each other.
- (d) F test is also used in some statistical procedures such as Scheffe's method for multiple comparisons adjustment in linear models.

SELF ASSESSMENT EXERCISES

Discuss the use of F test in a statistical test.

3.2 Formulae and Analysis of F Test Statistic

Majority of F tests in one way analysis of variance is used to examine whether the expected values of a quantitative variable within different/ several pre-defined groups differs from one to another. For instance, assuming that success of students at WAEC level compare four ways of achieving success, the ANOVA F test can be used to investigate whether any of the achievement of success is on average of hard work or crooked way of passing, to the others versus the null hypothesis that all four ways of achieving success yield the same mean response. This can be said to be an example of what is called 'OMNIBUS' test, meaning that a single test is performed to detect any several possible differences, or we could carry out pair wise tests among the success achievement (for instance in the success example, we could carry out six tests among pairs of success).

However, the advantages of the ANOVA F-test is just that we do not need to pre-specify which success achievement are to be compared, and we do not need to adjust for making multiple comparisons but the disadvantage of the ANOVA F-test is that if we reject the null hypothesis, we do not know which success achievement can be said to be significantly different from the others that is if we perform the F test at level α we cannot state that the success achievement pair with the greatest mean difference is significantly different at level α .

Let us now specify the formulae for one-way ANOVA F tests Statistics is:

$$F = \frac{\textit{explained variance}}{\textit{unexplained variance}}$$

or

$$F = \frac{\textit{between - group variability}}{\textit{within - group variability}}$$

The "explained variance" or "between-group variability is

$$\sum_i n_i(\bar{y}_i - \bar{y})^2 / (k - 1)$$

Where \bar{y}_i is the sample mean in the i^{th} group, n_i is the number of observations in the group. However, \bar{y} denotes the overall mean of the data and k is the number of groups.

The “unexplained variance” or “within-group variability” is given as:

$$\sum_{ij} (y_{ij} - \bar{y}_i)^2 / (N - K)$$

where y_{ij} is the j^{th} observation in the i^{th} out of K groups and N is the overall sample size. More so, the F test statistic follows the F -distribution with $k - 1, N - k$ degrees of freedom under the null hypothesis. But you should note that the statistic will be large if the between group variability is large relative to the within group variability, which is unlikely to occur if the population means of the groups all have the same value and when there are only two groups for one-way ANOVA F test, $F = t^2$ where t is the student's statistic.

More so, for regression analysis problem, the F - test statistic is given:

$$F = \frac{\left(\frac{Rss_1 - Rss_2}{P_1 - P_2} \right)}{\left(\frac{Rss_2}{n - P_2} \right)}$$

Where Rss_i is the residual sum of squares of model i . If your regression model has been calculated with weights, then replace Rss_i with X^2 , the weighted sum of squared residuals.

Using the hypothesis testing, under the null hypothesis is rejected if the F -calculated from the data is greater than the critical value of the F -distribution for some desired false rejection probability (say for example 0.05) and you should also note that the F -test is a wild test.

SELF ASSESSMENT EXERCISE

Discuss the use of F -test in statistical analysis.

4.0 CONCLUSION

F test has a lot of uses in statistical analysis and its test the overall significant of the parameters. Therefore the formulae of F test makes use of ANOVA and it is also specify that it is used in regression analysis.

5.0 SUMMARY

The unit has critically discussed the F test in statistical analysis as well as the uses of the test were also discussed as well as the formulae both in ANOVA analysis and regression analysis.

6.0 TUTOR MARKED ASSESSMENT EXERCISE

1. List the common examples of F-tests in statistical analysis.
2. Discuss the advantages and disadvantages of F-test ANOVA
3. Specify the formulae and define the parameter of F-test statistical with explained variance and unexplained variance.

7.0 REFERENCES FOR FURTHER READING

Wemimo, A. (2007). *Undersatnding the Concept of Statistics, a contemporary approach*, 1st edition, Merrinlyn Press limited.

UNIT 3 CHI-SQUARE ANALYSIS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Examples of Chi-square distribution
 - 3.2 Application of Chi-square analysis
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

A Chi-square can said to be a measurement of how expectations are compared to results. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables and be drawn from a large enough sample. For example, the result of tossing a coin 100 times would meet these criteria. Furthermore, chi-square test (X^2 test) is a statistical hypothesis test where the sampling distribution of the test statistic, is a chi-squared distribution when the null hypothesis H_0 is true.

2.0 OBJECTIVES

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

- Know the examples of chi-square distribution
- Understand the application of chi-square analysis

3.0 MAIN CONTENT

3.1 Examples of Chi-Square Distribution

1. Pearson's Chi-Square Test

This is a statistical test that is applied to categorical data to investigate how likely it is that any observed difference between the sets arose by chance and it is good for unpaired data that can be seen from large samples.

Moreover, it is used to assess the two types “test of goodness of fit” and tests of independence. The test of goodness of fit analyse whether or not the observed frequency distribution is different from the theoretical distribution while the test

of independence analyse whether the paired observations on two variables, expressed in a contingency table are independent of one another.

However the test can be calculated by calculating the chi-squared test statistic, X^2 which shows the normalised sum of squared deviations between observed and the theoretical frequencies and you then determine the degree of freedom (df) of the statistic which means the numbers of frequencies reduced by the number of parameter of the distribution. After you must have done this, you then compare the X^2 calculated value to the tabulated value using the degree of freedom.

(2) Discrete Uniform Distribution

The formula for the discrete Uniform Distribution is given as:

$$E_i = \frac{N}{n}$$

when N observations are divided by n cells, but the degree of freedom reduction gives $p = 1$ because the observed frequencies O_i are constrained to sum N .

It should be noted that the degree of freedom are not based on the number of observation as with a student's t-test or F-test distribution. For example, when you are testing for a fair six sided die, you can only have five degrees of freedom because there are six parameters to be observed from 1 to 6. The numbers of times you rolled the die does not determine the number of degree of freedom.

Moreover, the chi-square test statistic is as follows:

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

where X^2 = Pearson's cumulative test statistic which approaches a X^2 distribution.

O_i = observed frequency
 E_i = Expected (theoretical) frequency supported by null hypothesis.
 n = Number of cells in table

And the degree of freedom is $(r - 1)(n - 1)$ where r is the numbers of rows and n is the number of cells in the table (column).

(3) Yate's correction for continuity

This is also called Yate's chi-squared test and it is used when testing for independence in a contingency table. However, the formulae below shows the Yate's corrected version of Pearson's chi-square statistic.

$$X_{Yates}^2 = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i}$$

where:

O_i = observed frequency
 E_i = Expected (theoretical) frequency
 N = Number of events.

(4) Cochran – Mantel Statistics

These are collections of test statistics that is used for the analysis of stratified categorical data. It shows the comparison of two groups on a different categorical response and it is used when the effect of the explanatory variable on the response variable is influenced by covariates that can be controlled. It is also used in observational studies where the random assignment of subjects to different treatment cannot be controlled but the influencing covariate can.

(5) Mc Nemar's Test

This is a statistical test that is used on paired nominal data. It makes use of 2x2 contingency tables to determine whether the row and column marginal frequencies are equal and its application is in the area of test in genetics where the transmission disequilibrium test for detecting linkage disequilibrium.

(6) Turkey's Test of Additivity

This is an approach use in ANOVA (that is a region analysis involving two qualitative factors) to detect whether the factor variables are additively related to the expected value of the response variables. It should be noted that turkey called turkey's one-degree of freedom test.

SELF ASSESSMENT EXERCISES

List and explain the examples of chi-square test in statistical analysis.

3.2 Application of Chi-Square Analysis

3.2.1 Application on Type of Data

There are two types of data of a random variable which if you can recall from the previous discussion; we said we have numerical and categorical data. Chi-square is used to examine whether the distributions of categorical variable in question differ from another and its yield data in categories, whereas, the numerical variables yield data in numerical form.

For example, responses from respondents that “what is your major”, “Do you own a house?” are called categorical data analysis, but when a question like “what is your age?” or “how tall are you?” are called Numerical data which can be discrete or continuous.

Let us use a table to briefly explain this:

Table 1

Data type	Question type	Possible Responses
Categorical	What is your marital status?	Single or married?
Numerical	Discrete- How many houses do you own?	Two or three
Numerical	Continuous- How tall are you?	66 Inches

Note that a discrete data arise from counting process that counting the in variables involves in say 1, 2, 3, while continuous data arise from a measuring process of trying to get the size of clothes, height and weight. Also you should bear in mind that chi-square tests can only be used on actual numbers and not on percentages, proportion, means, etc.

3.2.2 2×2 Application of Contingency Table

In this type of analysis, we can say that there are several types of chi-square test but it depends on how the data are collected and the null hypothesis that is being tested.

Let’s then look at the simplest case of a 2×2 contingency table. Later,. A, b, c and d are used to represent the contents of the cells, then we will have the following table:

Table 2: General notation for a 2x2 contingency table:

Variable 1

Variable 2	Data type 1	Data type 2	Total
Category 1	A	B	$a + b$
Category 2	C	D	$c + d$
Total	a+c	b+d	$a + b + c + d = N$

From the above table, the chi-square statistic is calculated by the formulae:

$$X^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + d)(c + d)(b + d)(a + c)}$$

Note that the four components of the denominator are the four totals from the table column and rows.

For example, let us assume we conduct a drug a drug test on a group of female women and you hypothesized that the female women receiving the drug would show increase in heart rates compared to those that did not receive the drug. Assuming you conduct the study and collected the following data:

Null Hypothesis(H_0): The female women whose heart rate increased are independent of drug treatment.

Alternative Hypothesis (H_1): The Female women whose heart rate increased are associated with drug treatment.

Table 3 The Drug Test Results

	Heart Rate increased	No Heart Rate Increased	Total
Treated	46	18	64
No treated	40	31	71
Total	86	49	135

Using the formulae specified above:

$$\begin{aligned}
&= \frac{135 [(146)(31) - (18)(40)]^2}{(64)(71)(49)(80)} \\
&= 135 \frac{[(1426 - 720)]^2}{(45.44)(3920)} \\
&= \frac{135(706)^2}{17812480} \\
&= \frac{6728860}{17812480} \\
&= 3.77
\end{aligned}$$

The next step is to know the degree of freedom, therefore the degree of freedom equal (number of columns minus one) \times (number of rows minus one). Moreover, the $Df = (c - 1)(R - 1)$.

$$\therefore Df = (2 - 1)(2 - 1) = 1.$$

But the chi-square statistic ($X^2 = 3.77$) and using alpha level significance and $df=1$. Looking through the chi-square table with 1 degree of freedom, the chi-square calculated ($\alpha = 0.05$) at 0.05 level of significance is 3.841.

Decision Rule

Therefore, the decision rule is that if the chi-square calculated is greater than the chi-square tabulated, we accept the alternative hypothesis (H_1) and reject the null hypothesis (H_0) OR if the chi-square tabulated is greater than the chi-square calculated, we accept the null hypothesis (H_0) and reject the alternative hypothesis H_1 .

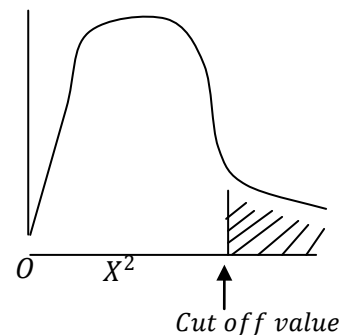
But in the case of our example, the chi-square calculated value is 3.77 and chi-square tabulated is 3.841, therefore, the chi-square tabulated (3.841) is greater than the chi-square calculated (3.77) then conclude that the female women whose heart increased is independent of drug treatment.

Table 4

The table below is an example of chi-square table.

Probability Level (Alpha)

Degrees of Freedom	Level of Significance (5%)	Level of Significance (1%)
$\nu = 1$	3.841	$\sqrt{\nu}$
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.812
8	15.507	20.090



3.2.3 Chi – Square Goodness of Fit (One Sample Test)

This is a test that allows us to compare a collection of categorical data with the theoretical distribution. It is also used in genetic theory analysis.

Now let us take an example of a collection of data on genetic of AA genotype and AS genotype of Mr Olusanya Samuel’s Family.

H_0 : AA genotype is better to AS in Olusaya’s family.

H_i : AA genotype is not better to AS in olusanya’s family.

	A	S	Totals
A	20	52	72
S	43	55	68
Totals	63	77	140

The genotype ration is 85 of AA and 55 of AS, while the cross bridging of AA to AS is 3: 1 and we expect that 75 of AA to 65 of AS. Do you think the result is different?

We can then go on to calculate the chi-square statistic by completing the following steps:

- (1) For each observed number in the table, subtract the corresponding expected number($O - E$)
- (2) Take the square of the difference $[(O - E)^2]$
- (3) You then divide the squares obtained for each cell in the table by the expected number for that cell $[(O - E)^2/E]$
- (4) Take the sum of all the values for $(O - E)^2/E$ and this is the chi-square statistics.

	Observed (O)	Expected (E)	(O - E)	(O - E)²	(O - E)²/E
AA	85	75	10	100	1.33
AS	55	65	-10	100	0.15
Total	140	140			1.48

Therefore $X^2 = 1.48$.

Using 5% level of significance (0.05) of the probability table below:

Table 5

PROBABILITY LEVEL (ALPHA)

Degrees of Freedom	Level of Significance (5%) (0.05)	Level of Significance (1%) (0.01)
$\nu = 1$	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.812
8	15.507	20.090

From the table, the chi-square tabulate at 5% level of significance (0.05%) is 3.84 while the chi-square calculated is 1.48. therefore, since the tabulated value is greater than the calculated value, we accept H_0 and reject H_1 then conclude

that AA genotype is not better to AS genotype in Babatunde Olusegun Olusanya's family.

More so, it should be noted that the chi-square test of independence make use of a contingency table that has r rows and c columns and the chi-square test here is the test of independence. In the test of independence both H_0 and H_1 are as follows:

H_0 : The two categorical variables are independent

H_1 : The two categorical variables are related

And the chi-square equation is given as the sum of all the $(F_o - F_e)^2 / F_e$ where F_o is the frequency of the observed data and F_e is the frequency of the expected values. The table 6 below becomes:

Table 6

	Category I	Category II	Category III	Row Totals
Sample A	<i>a</i>	<i>b</i>	<i>c</i>	$a + b + c$
Sample B	<i>d</i>	<i>e</i>	<i>f</i>	$d + e + f$
Sample c	<i>g</i>	<i>h</i>	<i>i</i>	$g + h + i$
Column Totals	$a + d + g$	$b + e + h$	$c + f + i$	$a + b + c + d + e + f + g + h + i = N$

Therefore, we need to calculate the expected values for each cell in the table and we perform that by using the row and the column total divided by the grand total (N), for instance, the expected value of cell 'a' will be $(a + b + c)(a + d + g) / N$.

SELF ASSESSMENT EXERCISES

Given the table below for open end question for strongly agree (SA), strongly disagree (SD), disagree (D) and agree (A) in a result for studying the impact of population on growth in Nigeria:

Variable	Responses
Strong Agree	62.40
Agree	23.20
Strongly Disagree	10.20

Disagree	4.20
Total	100

Hypothesis

H_0 :- Population has no impact on growth in Nigeria

H_1 :- Population has an impact on growth in Nigeria.

Required to calculate the chi-square statistics

3.4. Chi-Square Goodness of Fit Test

Chi-square goodness of fit test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company. The sample problem at the end of the lesson considers this example.

3.4.1. When to Use the Chi-Square Goodness of Fit Test

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

3.4.2. State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis (H_0) and an alternative hypothesis (H_a). The hypotheses are stated in such a way that they

are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

For a chi-square goodness of fit test, the hypotheses take the following form.

H_0 : The data are consistent with a specified distribution.

H_a : The data are *not* consistent with a specified distribution.

Typically, the null hypothesis (H_0) specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis (H_a) is that *at least* one of the specified proportions is not true.

3.4.3. Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the chi-square goodness of fit test to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square goodness of fit test is described in the next section, and demonstrated in the sample problem at the end of this lesson.

3.4.4. Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic.

- Degrees of freedom. The degrees of freedom (DF) is equal to the number of levels (k) of the categorical variable minus 1: $DF = k - 1$.
- Expected frequency counts. The expected frequency counts at each level of the categorical variable are equal to the sample size times the hypothesized proportion from the null hypothesis

$$E_i = np_i$$

where E_i is the expected frequency count for the i th level of the categorical variable, n is the total sample size, and p_i is the hypothesized proportion of observations in level i .

- Test statistic. The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \sum [(O_i - E_i)^2 / E_i]$$

where O_i is the observed frequency count for the i th level of the categorical variable, and E_i is the expected frequency count for the i th level of the categorical variable.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

3.4.5 Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Example:

Solomon Ogooluwa Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans, and 10% are All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
 - Null hypothesis: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
 - Alternative hypothesis: At least one of the proportions in the null hypothesis is false.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square goodness of fit test of the null hypothesis.
- **Analyze sample data.** Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

$$DF = k - 1 = 3 - 1 = 2$$

(E_i)	=	n	*	p_i
(E_1)	=	100	*	0.30 = 30
(E_2)	=	100	*	0.60 = 60
(E_3)	=	100	*	0.10 = 10

$$X^2 = \sum [(O_i - E_i)^2 / E_i]$$

$$X^2 = [(50 - 30)^2 / 30] + [(45 - 60)^2 / 60] + [(5 - 10)^2 / 10]$$

$$X^2 = (400 / 30) + (225 / 60) + (25 / 10) = 13.33 + 3.75 + 2.50 = 19.58$$

where DF is the degrees of freedom, k is the number of levels of the categorical variable, n is the number of observations in the sample, E_i is the expected frequency count for level i, O_i is the observed frequency count for level i, and X^2 is the chi-square test statistic.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.

We use the Chi-Square Distribution Calculator to find $P(X^2 > 19.58) = 0.0001$.

- **Interpret results.** Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5.

3.4.6. Chi-Square Test of Homogeneity

This lesson explains how to conduct a **chi-square test of homogeneity**. The test is applied to a single categorical variable from two different populations. It is used to determine whether frequency counts are distributed identically across different populations.

For example, in a survey of TV viewing preferences, we might ask respondents to identify their favorite program. We might ask the same question of two different populations, such as males and females. We could use a chi-square test for homogeneity to determine whether male viewing preferences differed significantly from female viewing preferences. The sample problem at the end of the lesson considers this example.

3.4.7 When to Use Chi-Square Test for Homogeneity

The test procedure described in this lesson is appropriate when the following conditions are met:

- For each population, the sampling method is simple random sampling.
- The variable under study is categorical.
- If sample data are displayed in a contingency table (Populations x Category levels), the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

3.4.8 State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

Suppose that data were sampled from r populations, and assume that the categorical variable had c levels. At any specified level of the categorical variable, the null hypothesis states that each population has the same proportion of observations. Thus,

$$\begin{aligned} H_0: P_{\text{level 1 of population 1}} &= P_{\text{level 1 of population 2}} = \dots = P_{\text{level 1 of population } r} \\ H_0: P_{\text{level 2 of population 1}} &= P_{\text{level 2 of population 2}} = \dots = P_{\text{level 2 of population } r} \end{aligned}$$

$$\dots$$

$$H_0: P_{\text{level } c \text{ of population } 1} = P_{\text{level } c \text{ of population } 2} = \dots = P_{\text{level } c \text{ of population } r}$$

The alternative hypothesis (H_a) is that *at least* one of the null hypothesis statements is false.

3.4.9 Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- **Test method.** Use the chi-square test for homogeneity to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square test for homogeneity is described in the next section.

3.4.10 Analyze Sample Data

Using sample data from the contingency tables, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic. The analysis described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:

$$DF = (r - 1) * (c - 1)$$

where r is the number of populations, and c is the number of levels for the categorical variable.

- **Expected frequency counts.** The expected frequency counts are computed separately for each population at each level of the categorical variable, according to the following formula.

$$E_{r,c} = (n_r * n_c) / n$$

where $E_{r,c}$ is the expected frequency count for population r at level c of the categorical variable, n_r is the total number of observations from population r , n_c is the total number of observations at treatment level c , and n is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \sum [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

where $O_{r,c}$ is the observed frequency count in population r for level c of the categorical variable, and $E_{r,c}$ is the expected frequency count in population r for level c of the categorical variable.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

3.4.11 Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Example:

In a study of the television viewing habits of children, a developmental psychologist selects a random sample of 300 first graders - 100 boys and 200 girls. Each child is asked which of the following TV programs they like best: The Lone Ranger, Sesame Street, or The Simpsons. Results are shown in the contingency table below.

	Viewing Preferences			Row total
	Lone Ranger	Sesame Street	The Simpsons	
Boys	50	30	20	100
Girls	50	80	70	200
Column total	100	110	90	300

Do the boys' preferences for these TV programs differ significantly from the girls' preferences? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
 - Null hypothesis: The null hypothesis states that the proportion of boys who prefer the Lone Ranger is identical to the proportion of girls. Similarly, for the other programs. Thus,

$$H_0: P_{\text{boys who prefer Lone Ranger}} = P_{\text{girls who prefer Lone Ranger}}$$

$$H_0: P_{\text{boys who prefer Sesame Street}} = P_{\text{girls who prefer Sesame Street}}$$

$$H_0: P_{\text{boys who prefer The Simpsons}} = P_{\text{girls who prefer The Simpsons}}$$
 - Alternative hypothesis: At least one of the null hypothesis statements is false.
- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square test for homogeneity.
- **Analyze sample data.** Applying the chi-square test for homogeneity to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = \frac{(n_r * n_c)}{n}$$

$$E_{1,1} = \frac{(100 * 100)}{300} = 33.3$$

$$E_{1,2} = \frac{(100 * 110)}{300} = 36.7$$

$$E_{1,3} = \frac{(100 * 90)}{300} = 30.0$$

$$E_{2,1} = \frac{(200 * 100)}{300} = 66.7$$

$$E_{2,2} = \frac{(200 * 110)}{300} = 73.3$$

$$E_{2,3} = \frac{(200 * 90)}{300} = 60.0$$

$$X^2 = \sum \left[\frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \right]$$

$$X^2 = (50 - 33.3)^2/33.3 + (30 - 36.7)^2/36.7 + (20 - 30)^2/30$$

$$\begin{aligned}
& + (50 - 66.7)^2/66.7 + (80 - 73.3)^2/73.3 + (70 - 60)^2/60 \\
X^2 &= (16.7)^2/33.3 + (-6.7)^2/36.7 + (-10.0)^2/30 + (-16.7)^2/66.7 + \\
& (3.3)^2/73.3 + (10)^2/60 \\
X^2 &= 8.38 + 1.22 + 3.33 + 4.18 + 0.61 + 1.67 = 19.39
\end{aligned}$$

where DF is the degrees of freedom, r is the number of populations, c is the number of levels of the categorical variable, n_r is the number of observations from population r , n_c is the number of observations from level c of the categorical variable, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count in population r for level c , and $O_{r,c}$ is the observed frequency count in population r for level c .

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.39.

We use the Chi-Square Distribution Calculator to find $P(X^2 > 19.39) = 0.0000$. (The actual P-value, of course, is not exactly zero. If the Chi-Square Distribution Calculator reported more than four decimal places, we would find that the actual P-value is a very small number that is less than 0.00005 and greater than zero.)

- **Interpret results.** Since the P-value (0.0000) is less than the significance level (0.05), we reject the null hypothesis.

Note: Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and the expected frequency count was at least 5 in each population at each level of the categorical variable.

Self Assessment Exercise

Adejoke Uthman Company prints baseball cards. The company claims that 40% of the cards are rookies, 70% veterans, and 20% are All-Stars. Suppose a random sample of 200 cards has 70 rookies, 75 veterans, and 10 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

3.5 Chi-Square Test for Independence

This lesson explains how to conduct a **chi-square test for independence**. The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference. The sample problem at the end of the lesson considers this example.

3.5.1 When to Use Chi-Square Test for Independence

The test procedure described in this lesson is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

3.5.2 State the Hypotheses

Suppose that Variable A has r levels, and Variable B has c levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

H_0 : Variable A and Variable B are independent.

H_a : Variable A and Variable B are not independent.

The alternative hypothesis is that knowing the level of Variable A *can* help you predict the level of Variable B.

Note: Support for the alternative hypothesis suggests that the variables are related; but the relationship is not necessarily causal, in the sense that one variable "causes" the other.

3.5.3 Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- **Test method.** Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.

3.5.4 Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:

$$DF = (r - 1) * (c - 1)$$

where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute $r * c$ expected frequencies, according to the following formula.

$$E_{r,c} = (n_r * n_c) / n$$

where $E_{r,c}$ is the expected frequency count for level r of Variable A and level c of Variable B, n_r is the total number of sample observations at level r of Variable A, n_c is the total number of sample observations at level c of Variable B, and n is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \sum [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

where $O_{r,c}$ is the observed frequency count at level r of Variable A and level c of Variable B, and $E_{r,c}$ is the expected frequency count at level r of Variable A and level c of Variable B.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Example:

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

	Voting Preferences			Row total
	Republican	Democrat	Independent	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

H_0 : Gender and voting preferences are independent.

H_a : Gender and voting preferences are not independent.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square test for independence.
- **Analyze sample data.** Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = (n_r * n_c) / n$$

$$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$$

$$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$$

$$X^2 = \sum [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

$$X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40$$

$$+ (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$$

$$X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$$

$$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, n_r is the number of observations from level r of gender, n_c is the number of observations from level c of voting preference, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count when gender is level r and voting preference is level c , and $O_{r,c}$ is the observed frequency count when gender is level r voting preference is level c .

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

We use the Chi-Square Distribution Calculator to find $P(X^2 > 16.2) = 0.0003$.

- **Interpret results.** Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

Note: Specifically, the approach is appropriate because the sampling method was simple random sampling, the variables under study were categorical, and the expected frequency count was at least 5 in each cell of the contingency table.

Self Assessment Exercise

Discuss in details the chi-square test for independence.

4.0 CONCLUSION

Chi-square analysis is one of the statistical test that measure how the expectations is being compared to the results and different examples of chi-square were examined such as Pearson's chi-square test, Yate's correction for continuity, Cochran – Mantel-Haenzel Statistics, Mac Nemar's test and turkey's test of additivity.

5.0 SUMMARY

The unit has vividly examined and discussed the analysis of chi-square statistic and the applications to several calculations such as 2×2 contingency table, categorical and numerical analysis and the chi-square of Goodness of fit for one sample.

Therefore I belief at this point you must have really understood the use and calculation of chi-square statistic.

6.0 TUTOR MARKED ASSIGMENT

1. Given the table below for open end question for strongly agree (SA), strongly disagree (SD), disagree (D), agree (A) and Neutral (N) in a result for studying the impact of population on growth in Nigeria:

Variable	Responses
Strong Agree	43.22
Agree	22.10

Strongly Disagree	48.74
Disagree	12.32
Neutral	23.62
Total	150

Hypothesis

H_0 :- Inflation has no impact on growth in Nigeria

H_1 :- Inflation has an impact on growth in Nigeria.

Required to calculate the chi-square statistics

2. Given the table below for open end question for strongly agree (SA), strongly disagree (SD), disagree (D), agree (A) and Neutral (N) in a result for studying the impact of population on growth in Nigeria:

Variable	Responses
Strong Agree	12.33
Agree	11.34
Strongly Disagree	43.44
Disagree	10.23
Neutral	22.66
Total	100.00

Hypothesis

H_0 :- Human Capital development has no impact on growth in Nigeria

H_1 :- Human Capital development has an impact on growth in Nigeria.

Required to calculate the chi-square statistics

3. Write short note on the following:

(a) Pearson's Chi-Square

- (b) Discrete Uniform distribution
- (c) Yate's correction for continuity
- (d) Cochran-Mantel Statistics
- (e) Mc Nemar's test
- (f) Turkey's test of additivity

4. Differentiate between 2 x 2 application of contingency table and Chi-square Goodness of fit (one sample test).

7.0 REFERENCES/FURTHER READING

Lee, I.G. (2008). *Working through the Statistics*, a broader approach, 1st edition, Leepy Publication limited.

Ramsey, J.J. (2011). *Statistics and Probability theory*, 2nd edition, Mill world Press limited.

MODULE 5 SIMPLE LINEAR REGRESSION ANALYSIS AND ITS APPLICATION

- Unit 1 Meaning of Regression Analysis
- Unit 2 Simple Linear Regression Analysis
- Unit 3 Application of Simple Linear Regression Analysis

UNIT 1 MEANING OF REGRESSION ANALYSIS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Uses of Regression Analysis
 - 3.2 Three conceptualization of Regression analysis
 - 3.3 Regression Models
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

The term regression was introduced by Francis Galton. In a famous paper, Galton found that, although there was a tendency for all parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to more or “regress” towards the average height in the population as a whole. Therefore, the height of the children of unusually tall and unusually short parents tends to move toward the average height of the population. However, regression analysis is a statistical process for estimating the relationship among variables. It also includes many techniques for modelling and analyzing several variables, which the focus is on the relationship between a dependent variable and one or more independent variables. Specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one independent variable is varied while the other independent variables are fixed.

2.0 OBJECTIVES

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

- Know the uses of regression analysis
- Understand the three conceptualizations of regression analysis
- Understand regression models

3.0 MAIN CONTENT

3.1 Uses of Regression Analysis

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to know, which of the independent variables are closely related to the dependent variable and to establish the form of these relationship whether positive relationship or negative relationship. Regression analysis is also used in casual relationship between a linear model that is between the dependent variable to an independent variables, but it should be noted that correlation does not imply causation like linear regression analysis.

More so, different techniques are used in carrying out the regression analysis and such method such as linear regression and ordinary least squares regression are parametric, and in that we can define regression function in terms of its finite number of unknown parameters that are estimated from the data. However, the non-parametric regression means the techniques that allow the regression function to lie in a specified set of functions, which may be finite – dimensional.

When we are performing a regression analysis methodology in practice, it depends on the type of the data generating process and it can be related or possibly related to the regression approach that is being used. Moreover, since the true type of data generating process is generally not known, regression analysis often then depends to some extent on making assumptions about this process and these assumptions are testable if a sufficient quantity of data is available. Regression analysis is also used for prediction for future planning and however with little effects or questions of causality based on observational data, regression methods can give misleading result.

SELF ASSESSMENT EXERCISES

Discuss the used of regression analysis

3.2 Three Conceptualizations of Regression Analysis

A researcher faced with a large amount of raw data will want to summarize it in a way that presents essential information without too much distraction. Examples of data reduction include frequency table or group specific means and variances. Like most methods in statistics, the objective is to predict, as closely as possible, an array of observed values of the dependent variable based on a simple function of independent variables. Obviously, predicted values from regression models are not exactly the same as observed ones. Characteristically, regression partitions on observation into two parts.

$$\boxed{\text{Observed}} = \boxed{\text{Structural}} + \boxed{\text{Stochastic}}$$

The observed part represents the actual values of the dependent variable at hand. The structural part denotes the relationship between dependent and independent variables.

The stochastic part is the random component unexplained by the structural part. Moreover, the last term may be regarded as the sum of three components; omitted structural factors, measurement error, and a wise. Omitting structural factors is inevitable in social science research because we can never claim to understand and measure all causal structures affecting a dependent variable. However, measurement error refers to inaccuracies in the way in which the data are recorded, reported or measured. Random noise reflects the extent to which human behaviour or occurrence of events is subject to uncertainty (i.e. stochastic influences).

Furthermore, how to interpret regression models is contingent on one's conceptualization about what regression does to data. We can then propose three different conceptualizations.

Causation: $\boxed{\text{Observed}} = \boxed{\text{True Mechanism}} + \boxed{\text{Disturbance}}$

Prediction: $\boxed{\text{Observed}} = \boxed{\text{Predicted}} + \boxed{\text{Error}}$

Description: $\boxed{\text{Observed}} = \boxed{\text{Summary}} + \boxed{\text{Residual}}$

SELF ASSESSMENT EXERCISES

Explain the transmission of observed value in regression analysis that yields structural and stochastic variables.

3.3 Regression Model

The conceptualizations discussed in unit 3.2 provide three different views of qualitative analysis. The first approach corresponds most closely to what might be perceived as a view in classical econometrics in which the model accurately represents the “true” causal mechanism that generates the data. However, the researcher's goal is to specify a model to uncover the data generating mechanism, or “true” causal model. The first approach can be viewed as an attempt to get as close as possible to a deterministic model. More modern

approaches would argue that there is no “true” model but rather some models are more useful, more interesting, or closer to the truth than others.

More so, the second approach is more directly applicable to fields like engineering where, given a relationship between explanatory variables and a response variable, the goal is to make useful response predictions for new data. For instance, suppose that the strength of a material is related to temperature and pressure during the manufacturing process. However, assuming that we produce sample of materials by varying temperature and pressure in a systematic way. One objective of modelling might be to find values of temperature and pressure that give the material maximum strength. Social scientists also employ this modelling approach in forecasting and may use this approach to identify people at risk of a particular outcome based on certain characteristics.

The Hurd approach reflects the current view in modern econometrics and statistics in which a model serves to summarize the basic features of data without distorting them. A principle called OCCAM’S RAZOR, or the LAW OF PARSIMONY, is often invoked when assessing competing explanations of the same phenomena. However, when applied to statistical models, this principle means that if two models equally explain the observed facts, the simpler model is preferred until new evidence proves otherwise. This approach differs from the first view in the sense that the question asked is not whether the model is “true” but whether it corresponds to the facts. The facts usually require formalization based on past research or theory. The model is then specified in accordance with theory or previous research.

Therefore, these conceptualizations are not mutually exclusive; the applicability of a particular interpretation hinges on concrete situations, particularly the nature of the research design and objectives. With most applications in social sciences utilizing observational data, our conclusion is to favour the last interpretation. That is, the primary goal of statistical modelling is to summarize massive amounts of data with simple structures and few parameters. With this conceptualization of regression models, it is important to keep in mind the trade off between accuracy and parsimony. Moreover, we desire accuracy in a model in the sense that we want to preserve maximum information and minimize errors associated with residuals.

Finally, we prefer parsimonious models. More often than not, the desire to preserve information can only be achieved by building complicated model, which comes at the expense of parsimony or simplicity. However, the tension between accuracy and parsimony is so fundamental to social science research.

SELF ASSESSMENT EXERCISE

Differentiate between the following:

- (a) Parsimony and accuracy of a model
- (b) Statistical modelling and statistical variables
- (c) Models and variables

4.0 CONCLUSION

Regression analysis has been defined in so many ways and has a lot of uses in statistical and economic modelling. Regression analysis also comes in form of observing the variables included in the model and making a prediction for future planning.

5.0 SUMMARY

The unit has seriously dealt with the meaning of regression analysis and its uses were also examined in details. However, the three conceptualization of regression model were also discussed and regression model was then analyzed to enable us understand the techniques involved in regression modelling.

6.0 TUTOR MARKED ASSIGNMENT

1. Discuss the phenomenon of relationship between dependent variables and independent variables in a regression analysis.
2. Highlight the uses of regression analysis in a statistical analysis
3. What are conceptualizations of Regression Analysis?

7.0 REFERENCES FOR FURTHER READING

Daniel, A.P & Yu, X. (1999).*Statistical Methods for categorical data Analysis*, Academic Press, inc. 1999.

Armstrong, J. S. (2012). Illusion in Regress Analysis, *International Journal of forecasting (forthcoming)* pg 28.

David, A. F. (2005).*Statistical models, theory and practice*, Cambridge University Press, 2005.

Chian, C. L. (2003).*Statistical Methods of Analysis*, World scientific, ISBN pg 981 – 938.

UNIT 2 SIMPLE/LINEAR ANALYSIS AND ITS APPLICATION CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Examples of Chi-square distribution
 - 3.2 Application of Chi-square analysis
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In statistics, linear regression is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables denoted x . In the case of one explanatory variable is called simple linear regression.

For more than one explanatory variable, it is called a multiple regression. However, in a linear regression, data are modelled using linear prediction functions and unknown model parameters are estimated from the data. Such models are called linear models. Linear regression could also be referred to as a model in which the conditional mean of y , given the value of x , is an affine function of x . More so, linear regression could also be a model in which the median, or some other quantile of the conditional distribution of y is given x is expressed as a linear function of x .

2.0 OBJECTIVES

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

- Know the uses of simple linear regression
- Understand linear regression analysis

3.0 MAIN CONTENT

3.1 Uses of Simple Linear Regression

Linear regression focuses on the conditional probability distribution of y given x rather than the joint probability distribution of y and x . Linear regression was the first type of regression analysis to be studied vigorously and to be used extensively in practical applications and this is because models which depend linearly on their unknown parameter are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of resulting estimates are easier to determine.

Therefore, linear regression has a lot of practice uses, such as the followings:

- (a) If the goal is prediction or forecasting or reduction, linear regression can be used to fit a predictive model to an observed data set of y and x values and after developing such a model, if an additional value of x is then given without its accompanying value of y , the fitted model can be used to make a prediction of the value of y .
- (b) Given a variable y and a number of variables X_1, \dots, X_p that may be related to y , linear regression analysis can be applied to quantify the strength of the relationship between y and the X_i , to examine which X_i may have no relationship with y at all and to identify which subsets of the contain redundant information about y .

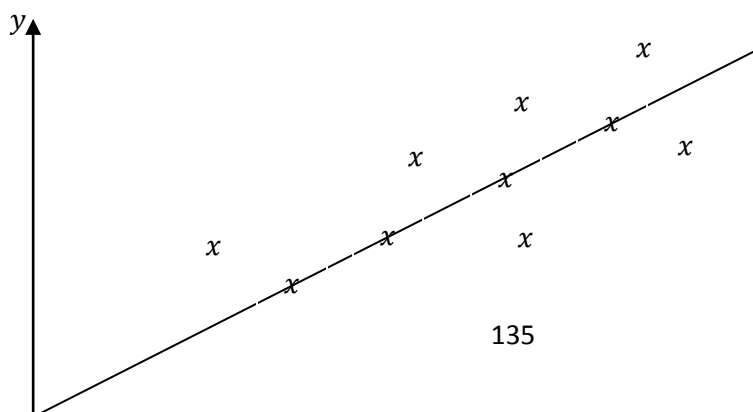
SELF ASSESSMENT EXERCISE

Discuss the uses of Simple Linear Regression in Statistical analysis.

3.2 Linear Regression Analysis

Sometimes the value of the graph of the data of a bi variant data may not all lie on a straight line. In that situation, we have a scatter diagram, an example is shown below.

Fig 1



x

—————→
A Scater Diagram

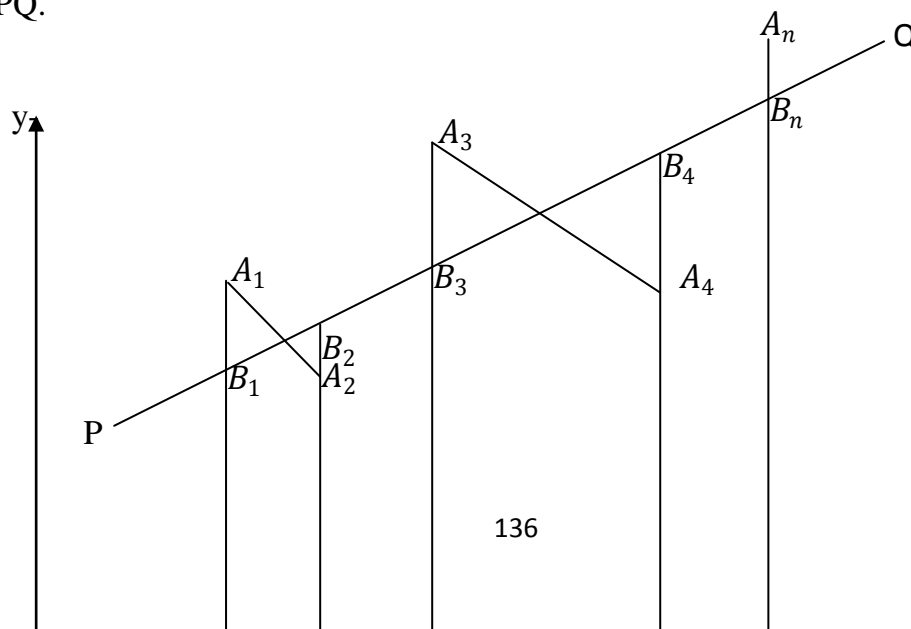
The line of best fit is then drawn to join those points that are in straight line on the diagram together and it is also called regression line given by the equation $y = n_1x + C$ where e is the error term in the straight line equation which if minimised will be zero and the result becomes $y = Mx + C$.

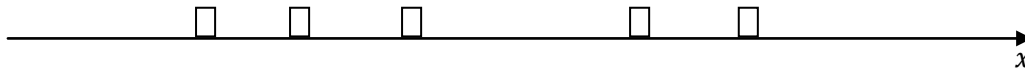
There are various methods of fitting the regression line. One of the methods is the free hand method where the line is fitted into the scatter diagram by eye, with the person drawing the line making efforts to be fair to all the points. This is a method adapted by students in physics practical where student plot point and end up with a scatter diagram and they are usually requested to fit in a line of best fit. The lines obtained by the student is not exactly the same and using hand to draw the line of best fit is a subjective and inadequate method. The other methods exist; such as the method of grand mean, the method of semi average and the least square method.

Further, we can then say that the least square methods is the best and most accurate and reliable method of fitting in a line of best fit.

Assuming that $A_1(n_1, y_1), A_2(n_2, y_2) \dots \dots A_n (X_n, y_n)$ are parts of a scattered diagram whose line of best fit is PQ with regression equation $y = Mx + C$.

Let $A_1, B_1, A_2, B_2 \dots \dots A_n, B_n$ be the distance between the points and the diagram in PQ.





Residual A_1B_1 has a distance of $[y_1 - (Mx_1 + C)]$, A_2B_2 has length $[y_2 - (Mx_2 + C)]$, A_3, B_3 has a distance of $[y_3 - (Mx_3 + C)]$... A_nB_n has a distance of $[A_n - (Mx_n - C)]$.

The sum of squares of the residuals:

$$S = [y_1 - Mx_1 + C]^2 + (y_2 - Mx_2 + C)^2 + \dots + [y_n - (Mx_n - C)]^2$$

$$= (y_1 - Mx_1 - C)^2 + (y_2 - Mx_2 - C)^2 + \dots + (y_n - Mx_n - C)^2$$

However, the method of least square for the regression line is such that s is minimum. Obtaining partial derivatives of s with respect to m and c , we get:

$$\frac{j_s}{j_m} = 2X_1(y_1 - Mx_1 - C) - 2x_2(y_2 - Mx_2 - C) - \dots - 2X_N(y_N - mx_n - C) = 0.$$

When S is a minimum

$$x_1y_1 + x_2y_2 + \dots + x_ny_n - m(n_1^2 + n_2^2 + \dots + n_N^2) - C(n_1 + n_2 + \dots + n_N) = 0$$

Using summation notation (Σ) we get $\Sigma xy - m\Sigma x^2 - C\Sigma x = 0$ that $\Sigma xy = m\Sigma x^2 + C\Sigma x$.

More so,

$$\frac{ds}{jc} = -2(y_1 - Mx_1 - C) - 2(y_2 - Mx_2 - C) - \dots - 2(y_N - MX_N - C) = 0$$

where S is a minimum. Dividing through by -2 we obtain.

$$(y_1 - Mx_1 - C) + (y_2 - Mx_2 - C) + \dots + (y_N - Mx_N - C) = 0$$

and then let us use set notation we have:

$$\Sigma y - m \Sigma x - N_c = 0$$

$$\Sigma y - m \Sigma x - N_c.$$

You should note that equation (1) and (2) are known as the normal equations: where m is the regression coefficients. Therefore, we then have:

$$\Sigma x_1, \Sigma x_1, \Sigma x^2, \Sigma x.$$

Example 1

The table below shows the units of fertilizer used and the units of yield in a science laboratory experience.

Units of fertilizer	1	2	3	4	5
Units of yield	20	15	18	12	10

Find the regression equation of x and y . Find the estimated value of y and $x = 2.5$

Solution

X	y	yx	X^2	y^2
1	20	20	1	400
2	15	30	4	225
3	18	54	9	324
4	12	48	16	144
5	10	50	25	100
Total 15	75	202	55	1193

Using the normal equations

$$\Sigma xy = m\Sigma x^2 + C\Sigma x \quad (1)$$

$$\Sigma y = m\Sigma x + N_c \quad (2)$$

we then obtain

$$202 = 55m + 15c \quad (1)$$

$$75 = 15m + 5c \quad (2)$$

(2) $\times 3$ gives

$$225 = 45m + 15c \quad (3)$$

(1) $- 3$ gives

$$-23 = 10m$$

$$m = -2.3$$

Substituting $m = -2.3$ into (1) we obtain
 $C = 21.9$

Therefore, the regression equation is

$$y = mx + c = -2.3x + 21.9.$$

$$x = 2.5$$

$$y = -2.3 \times 2.5 + 21.9 = 16.15$$

However, the regression coefficient m and the intercept can also be computed using the standard formulas as follows:

If $y = mx + c$ is the regression equation of y on x or therefore we have:

$$\text{Mor } \beta_1 = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$\text{M or } \beta_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\text{Covariance of } xy}{\text{VAR of } x}$$

and

$$C \text{ or } \beta_0 = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{N\Sigma x^2 - (\Sigma x)^2}$$

Example 2

Solve Example 1 above using formulae method.

Solution

$$\Sigma x = 15, \Sigma y = 75, \Sigma xy = 202$$

$$\Sigma x^2 = 55, \Sigma y^2 = 11.93$$

$$\text{M or } \beta_1 = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{5 \times 202 - 15 \times 75}{5 \times 55 - (15)^2}$$

$$= \frac{1010 - 1125}{275 - 225}$$

$$= \frac{-115}{50}$$

$$= -2.3.$$

$$C \text{ or } \beta_0 = \frac{\Sigma y \Sigma x^2}{N \Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{75 \times 55 - 15 \times 202}{5 \times 55 - (15)^2}$$

$$= \frac{4125 - 3030}{275 - 225}$$

$$= \frac{1095}{50}$$

$$= 21.9$$

Therefore, the regression equation $y = Mx + c$ or $y = \beta_0 + \beta_1 x_1$ becomes
 $y = 21.9 - 2.3x_1$

Also note that when β_1 is formed β_0 can be found with the formula

$$\beta_0 = \frac{1}{N} \Sigma y - \beta_1 \Sigma x$$

In the last example $\beta_0 = \frac{1}{5} (75 - 2.3 \times 15)$

$$\beta_0 = \frac{1}{5} (75 + 34.5)$$

$= 21.9$ as we before.

Regression of x on y

In this case, the axes are reversed and the equation is of the form $x = ay + b$ the coefficient becomes:

$$b = \frac{\Sigma x \Sigma y^2 - \Sigma y \Sigma xy}{N \Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{Cory}{VARy}$$

$$a = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$$

And b can also be found using $b = \frac{1}{N}(\Sigma x - a\Sigma y)$.

Example 3

Use the data in example 1 above to compute the regression of x on y .

Solution

$$N = 5, \Sigma x = 15, \Sigma y = 75, \Sigma xy = 202, \Sigma x^2 = 55, \Sigma y^2 = 1193$$

$$\begin{aligned} a &= \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2} \\ &= \frac{5 \times 202 - 15 \times 75}{5 \times 1193 - (75)^2} \\ &= \frac{1010 - 1125}{5965 - 5625} \\ &= \frac{-115}{340} = -0.34 \end{aligned}$$

$$\begin{aligned} b &= \frac{1}{N}(\Sigma x - a\Sigma y) = \frac{1}{5}[15 - (-0.34) \times 75] \\ &= 5.07 \end{aligned}$$

or

$$\begin{aligned} a &= \frac{\Sigma x \Sigma y^2 - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2} \\ &= \frac{15 \times 1193 - 75 \times 202}{5 \times 1193 - (75)^2} \\ &= \frac{17895 - 15150}{340} \end{aligned}$$

$$= \frac{2745}{340}$$

$$= 8.07.$$

Therefore the equation is $x = ay + 6$ so,

$$x = -0.34 t 8.07$$

SELF ASSESSMENT EXERCISE

- i. Discuss the importance of simple linear equation to statistical analysis or future prediction.
- ii. The table below shows the score of 6 male students in Olusanya Samuel College in mathematics examination.

No of students	1	2	3	4	5	6
Male Students score	22	14	12	20	16	10

You are required to calculate:

- (a) The regression equation y on x and x on y .
- (b) Find the estimated value of y when $x = 2.5$ and x when $y = 2.2$.

4.0 CONCLUSION

Simple linear regression was examined in this unit and we got to know that it is the relationship between the dependent and independent variable. However, various calculation of the straight line graph was also examined in this unit that calculation of regression equation $y = \beta_0 + \beta_1 X_1$ of y on x and x on y respectively.

5.0 SUMMARY

This unit has done justice to the simple linear regression analysis and the meaning and the uses of simple linear regression analysis were discussed in detail and also the calculation of the straight line regression equation and the following were derived in the unit.

Method 1

$$(1)\Sigma xy = m\Sigma x^2 + C\Sigma x$$

or

$$\Sigma xy = \beta_1 \Sigma x_1^2 + \beta_0 \Sigma x_1$$

$$(2) \quad \Sigma y = m \Sigma x + N_C$$

or

$$\Sigma y = \beta_1 \Sigma x_1 + N\beta_0$$

Method 2

$$\beta_1 \text{ or } m = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$\beta_0 \text{ or } C = \frac{N\Sigma x^2 - \Sigma x \Sigma xy}{N\Sigma x^2 - (\Sigma x)^2}$$

Method 3

$$\beta_0 \text{ or } m = \frac{Corxy}{Var x}$$

$$= \Sigma \frac{(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$\beta_0 \text{ or } C = \frac{1}{N} = (\Sigma y - m \Sigma x)$$

and it should be noted that the choice of the formulae will depend on what data is easily available.

6.0 TUTOR MARKED ASSIGNMENT

1. List and explain at least two uses of Simple Regression Analysis.
2. The following table is the numbers of children of Olusanya Samuel family with their body weight;

No. of children	1	2	3	4	5	6	7
Body weight (Kg)	66	55	110	88	72	54	48

You are required to calculate:

- (i) The regression equation of parameter β_0 and β_1 of y and x .
- (ii) The regression equation of parameter β_0 and β_1 of x on y .

(iii) Specify the regression equation.

3. The following table is the numbers of children of Dr Adegoke's family with their body weight;

No. of children	1	2	3	4	5	6	7	8	9	10	11
Body weight (Kg)	33	29	78	109	74	58	42	90	87	49	113

You are required to calculate:

- (iv) The regression equation of parameter β_0 and β_1 of y and x .
- (v) The regression equation of parameter β_0 and β_1 of x on y .
- (vi) Specify the regression equation.

7.0 REFERENCES/FURTHER READINGS

Adedayo, A. O. (2006). *Understanding statistics*, pg 157 – 163, the Publisher, Akoka, Lagos.

Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, pg 267-288, *Jstor edition*.

Nievergelt, Y. (1994). *Total Least Square, State of the art Regression in Numerical Analysis*, 1st edition.

UNIT 3 APPLICATION OF SIMPLE LINEAR ANALYSIS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 How to interpret simple regression analysis results.
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

Application of Simple Linear regression analysis is the way by which we subject different data to statistical analysis by using computer software such as SPSS, to analyse and predict the relationship between the dependent variable and independent variable. For example, we can develop a model to investigate the impact of unemployment on Nigerian Economy. The model will have both dependent and independent variables to be able to predict whether there is any element of impact of unemployment in Nigerian Economy or not.

However, the relationship can either tell us whether it has a direct or positive relationship or indirect or negative relationship on each other and from the analysis, we can deduce or conclude whether there is an impact or not.

2.0 OBJECTIVES

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

- Understand what a simple regression analysis can deduce from its application.
- How to interpret the parameter using t and f-tests respectively.

3.0 MAIN CONTENT

3.1 How to Interpret Simple Regression Analysis Results

Regression analysis generates an equation to describe the statistical relationship between one or more predictor variables and response variables. After you use statistical software to fit a regression model, and verify the fit by checking the residual plots, you will want to interpret the results.

3.1.1 How to Interpret the P-Values in Linear Regression Analysis

The P-values for each term test the null hypothesis that the coefficient is equal to zero (no effect). However, a low p value (<0.05) indicates that you reject the

null. In other words, predictor that has how p-value is likely to be meaningful addition to you model because changes in the predictor’s value are related to changes in the response variable.

However, a large (insignificant) p-value suggests that changes in the predictor are not associated with changes in response.

Let us take a hypothetical example that is shown below:

Table 1

Coefficients:

Term	coefficient	S.E Coefficient	T statistics	P
Constant	0.4324	1.0241	0.3216	0.000
East	1.2331	1.3214	1.7321	0.092
South	2.3201	0.2240	5.6631	0.000
North	-1.4321	0.6431	-12.6431	0.000

From the table above, we can see that the p-value of South and North are significant because both their p-values are 0.000, but the predictor value for East is 0.092 greater than the common alpha (α) of 0.05 and this indicates that it is not statistically significant.

3.1.2 How to Interpret the Regression Coefficients for Linear Relationships

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model.

Moreover, the key to understanding the coefficients is to think about them as slopes and invariably they are often called slopes.

Let’s use an example below to explain.

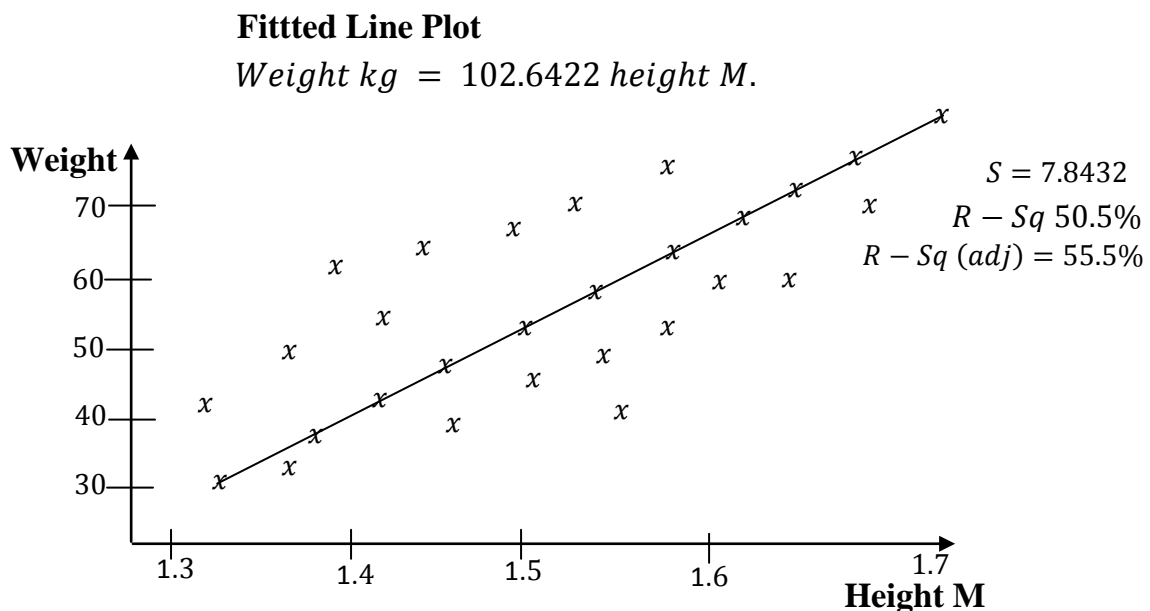
Table 2

Coefficients

Term	Coefficient	S-E Coefficient	T Statistics	P
Constant	-3.4601	18.4421	-1.6472	0.000
Height M	102.6422	16.0041	1.6632	0.000

The fitted line plot shows the same regression results graphically.

Figure 1



The equation shows that the coefficient for height is 102.6422 kilograms. The coefficient indicates that for every additional meter in height you can expect weight to increase by an average of 102.6422 kilograms.

From the graph above, a fitted line graphically shows the same information. For example if you move left right along the x axis by an amount that denotes a one meter change in height, the fitted line rises or fall by 102.6422 kilograms. However, these heights are from middle –school aged boys and ranges from 1.3m to 1.7m. The relationship is only valid within this data range, so we could not actually shift up or down the line by a full meter in this case.

In the fitted lines were flat (a slope coefficient of zero), the expected value for weight would not change no matter how far up and down the line you may go. Therefore, when we have low p-value, it shows that the slope is not zero, but

indicates that the changes in the predictor variable are associated with changes in the response value.

Let us then take a full Simple Linear Regression:

$$\text{A model is specified as: } GRT = f(GHE) \Rightarrow GRT = \beta_0 + \beta_1 GHE$$

Simple regression result

Dependent variable: GDP

Method:- Least Square

Year of analysis: 1980 – 2013

Included observations – 33

Variable	Coefficient	Std Error	T - Statistics	P
C	0.53101	0.39431	0.64311	0.000
GHE	1.23421	0.66401	1.36101	0.000
R Square	0.66400		F test	23.43202

Where the model specification is given as;

$$GDP = f(GHE) \dots\dots\dots (1)$$

$$GDP = a_0 + a_1GHE + e \dots\dots\dots(2)$$

Where GDP is the Gross domestic product (proxy for economy development)
 GHE is the Government health expenditure in Nigeria and ‘e’ is the error term

Hypothesis

Null Hypothesis (Ho): Government health expenditure has no significant effect on economy development in Nigeria.

Alternative Hypothesis (H1): Government health expenditure has a significant effect on economy development in Nigeria.

From the table above, we can say that we have a simple linear regression result. The c is the constant and it is also called the intercept and in regression result we do not normally interpret this, but the model shows the impact of government health expenditure on economy development in Nigeria. So let us start by interpreting the other parameters. We can say that there is a direct/positive relationship between Government health expenditure and economy development in Nigeria that is 1unit increase in economy

development will lead 1.23421 unit increases in government health expenditure. However, the t calculated from the result is given to be 1.3101 while the t tabulated using 5% (0.05) level of significance which is called the t tabulated is given to be 1.96 from the t statistical table. The decision rule for t test is that since the t tabulated (1.96) is greater than the t calculated, we accept the null hypothesis and reject the alternative hypothesis, then conclude that Government health expenditure has a significant effect on economy development in Nigeria and also conclude that the parameter government health expenditure is a good explanatory variable. More so the f test for the overall significance of the parameters, and the f calculated from the result is 23.4320 while the f tabulated using 5% (0.05) level of significance is 12.70 from the f statistical table. The decision rule is that since the f calculated is greater than the f tabulated, we accept the alternative hypothesis and reject the null hypothesis, and then we conclude that the overall parameter of the model is statistically significant. Finally, the coefficient of multiple determinations (R square) is 0.66400 from the result above which is 66.4%. This means that about 66.4% of the dependent variable (GDP) has been explained by the explanatory variable (GHE) while the remaining 33.6% are not present in the model or are outside the model.

SELF ASSESSMENT EXERCISE

Explain what steps you will take in interpreting the predictor for value in a regression analysis result.

4.0 CONCLUSION

How to interpret the p values and the coefficient of a regression analysis has shown an important tool in predicting to the future of analysis that may prove difficult when looking through its theoretical analysis. However, regression model take care of the situation of research problem and analyse it completely by running the analysis of dependent and independent variables on one another to get a suggested value for policy maker and planning analysis for the present and the future.

5.0 SUMMARY

This unit has briefly touch the application of simple linear regression analysis to statistical analysis of computer software such as stata, e-views and SPSS (Special Package for Social Science) to run an analysis that will leads to getting close to solution of a certain policy problems in the present and future justification.

6.0 TUTOR MARKED ASSIGNMENT

1. Define the following terms with detailed examples:

- (a) Predictor Value
 - (b) Coefficient of a regression analysis
 - (c) Model specification
2. Given the p values below, interpret the implications.

P values
0.000
0.042
0.001
0.032

3. Given the regression result below with the model is given as:

$$GDP = f(INF)$$

$$GDP = \beta_0 + \beta_1 INF$$

The parameter is looking at the impact of inflationary rate on economic growth.

Hypothesis

Null hypothesis (Ho): Inflationary rate has no significant impact on economic growth in Nigeria

Alternative hypothesis (H1): Inflationary rate has a significant impact on economic growth in Nigeria

Dependent variable: - GDP
Year of analysis: - 1980-2013
Included variable: - 33
Method: Least Square

Variable	Coefficient	T stat.	Prob.
C	0.43260	0.6431	0.000
INF	-1.36471	1.6612	0.000
R-Square	46.4201	F test	21.64320

You are required to interpret the regression result above and test for statistical significance of t and f test.

4. Given the regression result below with the model is given as:

$$GDP = f(MS)$$

$$GDP = \beta_0 + \beta_1 MS$$

The parameter is looking at the impact of money supply on Economic growth.

Hypothesis

Null hypothesis (Ho): Money Supply has no significant impact on economic growth in Nigeria

Alternative hypothesis (H1): Money Supply has a significant impact on economic growth in Nigeria

Dependent variable: - GDP

Year of analysis: - 1990-2013

Included variable: - 23

Method: Least Square

Variable	Coefficient	T stat.	Prob.
C	0.42311	0.78843	0.000
INF	2.87742	1.94222	0.000
R-Square	0.88211	F test	18.88972

You are required to interpret the regression result above and test for statistical significance of t and f test.

7.0 REFERENCES/FURTHER READING

Olomeko, O. A. (2012). *Regression Analysis*, 1st edition, pg 88 Millworld Publication Limited, Lagos.

