NATIONAL OPEN UNIVERSITY OF NIGERIA

STATISTICS FOR ECONOMIST II

ECO 254

FACULTY OF SOCIAL SCIENCES

DEPARTMENT OF ECONOMICS

COURSE GUIDE

Course Developer
Samuel Olumuyiwa Olusanya
Economics Department, National Open University of Nigeria

Course Editor
Dr. Ganiyat A. Adesina-Uthman
Economics Department, National Open University of Nigeria

Course Reviewer
Samuel Olumuyiwa Olusanya
Economics Department, National Open University of Nigeria

CONTENT

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 1Probability 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 2Hypotheses 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 3Sampling 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 dairy or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working breach unit.
3. Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they get behind with their course work. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.
4. Turn to Unit 1 and read the introduction and the objectives for the unit.
5. Assemble the study materials. Information about what you need for a unit is given in the `Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your set books on your desk at the same time.
6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.
7. Up-to-date course information will be continuously delivered to you at the study centre.
8. Work before the relevant due date (about 4 weeks before due dates), get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments no later than the due date.
9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.
10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.
11. When you have submitted an assignment to your tutor for marking do not wait for it 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

This module introduces you to Probability distribution. The module consists of 3 units which include: Analysis of probability distribution, continuous random variables and other probability distribution.

Unit One: Analysis of Probability Distribution
Unit Two: Continuous Random Variables
Unit Three: Other Probability Distributions

UNIT One: ANALYSIS OF PROBABILITY DISTRIBUTION

Unit Structure

1.1 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.

1.2. Learning Outcome

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

i. understand the meaning of probability distribution
ii. Understand the application of probability
iii. Understand different types of probability distribution
iv. Analyze probability problem.

1.3. 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 value 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?

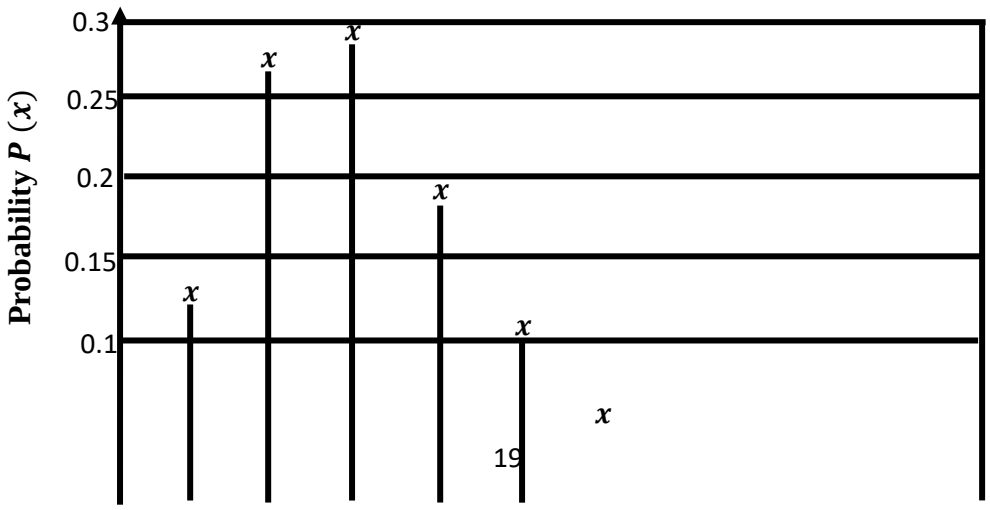
Problem III: The (arithmetic) mean life (continuous play on a specific cassette player) of a make and type of battery is 22 hours with a standard deviation of 0.11 hour. What is the probability a battery will last no more than 21hours?

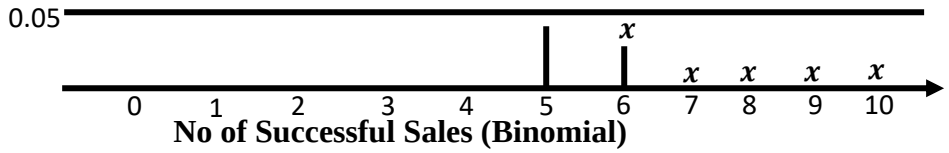
In problem I, the random variable of interest is the number of Sales per day. We refer to it as a random variable since the outcome in any one day is uncertain and may be considered to be dependent on chance. A random variable is a quantity resulting from a random experiment that by chance can assume 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≤ × <1) hour; 1≤ ×<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 Exercises 1

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

1.4. Discrete and Continuous Probability Distribution

Discrete (individually countable) outcomes, such as 1, 2, 3, yes, no, and true or false, are represented by a discrete distribution, which is a probability distribution. The binomial distribution, for instance, is a discrete distribution that assesses the likelihood that an event will occur "yes" or "no" over a specified number of trials, given the likelihood of the event in each trial, such as flipping a coin 100 times and getting "heads."

There are two types of statistical distributions: discrete and continuous. All outcomes greater than 0 (including those whose decimal points go on forever, like pi = 3.14159265), for example, are used to create a continuous

distribution. Probability theory and statistical analysis are fundamentally based on the ideas of discrete and continuous probability distributions, as well as the random variables they describe.

However, 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:

$$\sum_{u} \Pr(\times = u) = 1$$

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) = $\frac{1}{2n}$ for n = 1, 2, ..., we have the sum of probabilities ½ + ¼ + ⅛ + ... =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.

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 ⋋ (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 opposite 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.

By one convention, a probability distribution μ is called continuous if its cumulative distribution function F(x) = $\mu(-\infty, x)$ is continuous and therefore, the probability measure of single tons $\mu\{x\}$= 0 for all x.

Self-Assessment Exercises 2

Critically differentiate between Discrete and Continuous probability distribution with mathematical example.

1.5. Properties of Probability Distributions

Probabilities associated with individual values of x cannot be negative or greater than 1 since this has no meaning. We saw in the previous analysis under

the probability distribution that $0 \leq p(x) \leq 1$. Second, the sum of the probabilities of all possible valves of x must equal 1, e.g. by summing (the addition rule) we mean either x=0 or x=1 or x= 2. x=20 i.e. there is probability of 1 (certainty) that one of the mutually exclusive outcomes will occur. Thus for the discrete case:

$$\sum_{-\infty}^{+\infty} P\,(x) = 1 \_\_\_\_\_ eq\,(i)$$

However, in practice, the range of possible values will not generally go from minus infinity (- ∞) to plus infinity (+ ∞), the range only being stipulated here for completeness. For continuous distribution readers with knowledge of the calculus will appreciate that f(x) describes the curve of the probability density function:

$$\int_{-\infty}^{+\infty} f\,(x)\,d\,x\, = 1 \_\_\_\_\_ eq\,(ii)$$

The integral sign (∫) embraces (in theory) all possible values from minus infinity to plus infinity (- ∞ to + ∞) and sums the probability by evaluating the total area under the curve. The area occurring. You may ask the reason why we don't just sum the individual probabilities as in the discrete case. However, you must remember in principle that the random variable x may take an infinite number of possible values with an infinite number of associated probabilities, and summing these is slightly problematic. This is why integral calculus helps out. For our purpose there is no need to know how to integrate, but only why the integral sign has appeared. More so, for probability density functions it is the area under the curve corresponding to a range of values for x that yield the probability of that range of values occurring. Thus:

$$\int_{a}^{b} f(x)dx = p(a < x < b) \_\_\_\_\_ eg\,(iii).$$

The probability distribution such as the one in e.g. (i) have an average (mean) value for x as well as a standard deviation and variance. However, the mean or expected value of x is defined by:

$$E(x)\begin{cases} \sum_{-\infty}^{+\infty} xp(x) discreterandomvariable \\ \int_{-\infty}^{+\infty} xf(x)dx\, continuousrandomvariable \end{cases}$$

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:

$$\sigma^2 \begin{cases} \sum_{-\infty}^{+\infty} (x-\mu)^2 \ P(x) discrete\ random\ variables \\ \int_{-\infty}^{+\infty} (x-\mu)^2 \ f(x) continous\ random\ varibles \end{cases}$$

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.

Self-Assessment Exercises 3

What are the basic properties of distribution?

1.6. SUMMARY

Each probability distribution is applied to a specific situation and analysis; for example, the probability that an event will occur plus the probability that it won't occur is equal to 1, mathematically Pr(an event will occur) + Pr(an event will not occur) = 1. Probability distributions are used for a variety of purposes, including the measurement of various possible outcomes, a random experiment, survey, and procedure of statistical inferences. As a result, there are numerous applications for probability distribution in various contexts.

1.7. REFERENCES/Further Reading

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.

1.8. Possible Answers to SAEs

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

Answers to SAEs 1

A probability distribution function indicates the likelihood of an event or outcome. Statisticians use the following notation to describe probabilities: p(x) = the likelihood that random variable takes a specific value of x. The sum of all probabilities for all possible values must equal 1.

Answers to SAEs 2

For a discrete distribution, probabilities can be ascribed to the values in the distribution; for instance, "the probability that the web page will have 12 clicks in an hour is 0.15." A continuous distribution, on the other hand, has an infinite number of possible values, and the probability associated with any specific value is infinite.

Answers to SAEs 3

There are three features of distributions. Shape, central tendency, and variability are the three factors that best sum up a distribution. In later chapters, we'll discuss central tendency (roughly speaking, the distribution's center) and variability (the size of the distribution).

UNIT 2 CONTINUOUS RANDOM VARIABLES

Unit Structure

2.1. INTRODUCTION

A continuous random variable is a random variable where the data can take infinitely many values. For example a random on 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:

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

Where the function f(x) has the properties:

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

2.2. Learning Outcome

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

i. Understand the Continuous probability distribution
ii. Know how to calculate the distribution functions for continuous random variables

2.3 Continuous Probability Distribution

A continuous probability distribution is one in which the random variable X is capable of taking on any value. The likelihood of X taking on any one particular value is zero because there are an unlimited number of possible values for it. As a result, we frequently use ranges of values (p(X>0) =.50). One illustration of a continuous distribution is the normal distribution. Continuous or discrete probability distributions are the two types of probability distributions. A continuous distribution has an infinitely wide range of values, making it uncountable. Time is one such example; you could count from 0 seconds to 1 billion, 1 trillion, and so on, indefinitely. A discrete distribution has a set of countable values.

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

$$P(q < M < p) = \int_{P}^{q} f(m)dm \_\_\_\_\_(2)$$

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.

 **Self-Assessment Exercises 1**

Find the constant b such that the function

$$f(n) = \begin{cases} bn^2 0 < x < 3 \\ 0 \; otherwise \; is \; the \; density \end{cases}$$

Function and then find P(1 <×< 2)

2.4 Distribution Functions for Continuous Random Variables

The definition of a continuous random variable's distribution function is the probability that the value of the continuous random variable will be less than or equal to a real number.

The function f(x) is used to represent the probability distribution curve for a continuous random variable. A probability density function, or f(x), is a function that yields the distribution's curve. The region between the x-axis and the function f(x) is defined as being equal to a probability. We DO NOT receive probabilities associated with the continuous random variable from the probability density function f(x). The probability is represented by the area under the distribution graph produced by the function f(x).

A continuous probability distribution has the following characteristics:

1. The distribution's total area under the curve is 1.
2. The area under the curve of the distribution between x=c and x=d represents the likelihood that the continuous random variable will have a value between c and d.

3. There is no chance (P(x=c)) that the continuous random variable would exactly equal a specific value.

However, the cumulative distribution function or distribution function for a random variable is defined by

$$f(x) = P(\times \leq x) \_\_\_\_\_ (3)$$

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

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

Self-Assessment Exercises 2

Find the distribution function for:

$$f(x) = \begin{cases} bx^2 \;\; 0 < x < 3 \\ 0 \; otherwise \; is \; the \; density \; function \; and \; then \; find \; P(1 < \times < 2). \end{cases}$$

2.5. SUMMARY

Continuous random variables may be included in continuous probability distributions and distributional functions. As a result, I think that at the end of this unit, you should be able to analyze and calculate continuous random variables.

2.6. REFERENCES/Further Reading

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.

2.7. Possible Answers to SAEs

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

Answers to SAEs 1

Note that if b≥0, then property1 (equation 1) is satisfied. So f(x) must satisfy property 2 (equation 2) in order for it to be a density function.

Therefore:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{3} bx^2\,dx = \frac{bx^3}{3}\int_{0}^{3} = 9b$$

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

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

Next,

$$P(1 <\times< 2) = \int_{1}^{2} \frac{1}{9}bx^2dx = \frac{x^3}{27}\int_{1}^{2} = \frac{8}{27} - \frac{1}{27} = \frac{7}{27.}$$

Answers to SAEs 2

$$f(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{0}^{x} \frac{1}{9}x^2\,dx = \frac{x^3}{27} where\ x \leq 3.$$

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

$$P(x \leq \times \leq x + \Delta x) = \int_{x}^{x+\Delta x} f(x)dx \_\_\_\_\_ (5)$$

so that if Δx is small, we have approximately

$$P(x \leq \times \leq x + \Delta x) + f(x)\Delta x \_\_\_\_\_ (6)$$

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

$$\frac{df(x)}{dx} = f(x) \_\_\_\_\_ (7)$$

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

UNIT 3 OTHER PROBABILITY DISTRIBUTIONS

Unit Structure

3.1. 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.

3.2. Learning Outcome

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

i. Understand the Multinomial distributions
ii. Understand the Hyper geometric distribution
iii. understand the uniform distributions
iv. know the differences between Cauchy, Gamma and Bata Distribution

3.3. The Multinomial Distribution

In statistics, the term "multinomial distribution" refers to a generalization of the "binomial distribution," which only accepts two values (such as success and failure). An extension of the binary distribution, which only accepts two values (such as success and failure) in statistics, to more than two values. The multinomial distribution is a distribution function for discrete processes that has set probability for each independently generated value, similar to the binomial distribution. Multinomial distributions are more helpful when all of the results are of interest, but processes involving multinomial distributions can also be studied using the binomial distribution by concentrating on one result of interest and grouping all of the other results into one category (simplifying the distribution to two values).

Applications in geology and biology frequently use multinomial distributions. For instance, the 19th-century Austrian botanist Gregor Mendel crossed two pea strains, one with green and wrinkled seeds and one with yellow and smooth seeds, resulting in strains with four distinct seeds: green and wrinkled, yellow and round, green and round, and yellow and wrinkled. He discovered the fundamental ideas of genetics by studying the ensuing multinomial distribution.

So, let us make some hypothetical examples such that: an event $A_{1,}A_{2,}$ ..... A_k are mutually exclusive, and can occur with respective probabilities $P_{1,}P_{2,}$ ..... P_k where $P_1 + P_2$ + ... + P_K + 1. If $X_{1,}X_{2,}$ ....., X_k are the random variables respectively, given the number of times that $A_{1,}A_{2,}$ ..... A_k occur in a total of n trials, so that $X_{1,}X_{2,}$+....., X_k = n, then P($X_1 = n_{1,}K_2 = n_{2,}$ .....,$X_k = n_k$) = $\frac{n}{n_1!\, n_2!\, \ldots\ldots\, n_k!} P_1^{n_1} P_2^{n_k} \ldots\ldots P_k^{n_k}$ _____(1)

where $n_1 + n_2 + \ldots n_k = n$, is the joint probability function for the random variables $x_{1,}x_{2,}$ ....., 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 $P(P_1 + P_2 + \cdots P_k)^n$.

Self-Assessment Exercises 1

What is meant by multinomial distribution?

3.4. The Hyper geometric Distribution

In statistics, a hypergeometric distribution is a distribution function that allows selections from two groups without replacing any of the group members. The absence of replacements in the hypergeometric distribution sets it apart from the binomial distribution. In statistics, a distribution function is used to choose between two groups without changing any of the group members. The absence of replacements in the hypergeometric distribution sets it apart from the binomial distribution. As a result, it is frequently used in random sampling to ensure statistical quality. A straightforward, everyday example would be choosing team members at random from a population of males and girls. Like the Poisson and binomial distributions, this distribution uses discrete variables and is a type of probability distribution. Combinatorial analysis has been used in it.

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.

Self-Assessment Exercises 2

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.

3.5. 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

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

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

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

The mean and variance are, respectively

$$\mu = \frac{1}{2}(a+b)_1 = \sigma^2 = \frac{1}{12}(b-a)^2 \_\_\_\_\_(3)$$

Self-Assessment Exercises 3

Calculate the mean and variance of a uniform distribution given that $(a+b) = {}^{4}/_{128}$ *and* $b = 6.40$ *while* $a = 2.10$.

3.6 Cauchy, Gamma and Beta Distribution

3.6.1 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*;

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

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.

3.6.2 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:

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

where $\boldsymbol{\Gamma}(\propto)$is the gamma function and the mean and the variance are given as follows;

$$\mu = \alpha\beta \; \sigma^2 = \alpha\beta^2 \_\_\_\_\_ (6)$$

3.6.3 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:

$$f(x) = \begin{cases} \dfrac{x^{\propto-1}(1-x)^{\beta-1}}{B(\propto,\beta)} & 0 < x < 1 \\ 0\ otherwise\ (\propto,\beta > 0) \end{cases} \_\_\_\_\_(7)$$

where $B(\propto,\beta)$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;

$$f(x) = \begin{cases} \dfrac{\Gamma\ (\propto +\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\ 0 < x < 1 \end{cases} \_\_\_\_\_(8)$$

where $\propto,\beta$are positive. Therefore we can then specify the mean and variance as follows;

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

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

$$x = \frac{\propto -1}{\_\_\_\_\_} \quad (10)$$

What is the difference between beta distribution and gamma distribution?

Self-Assessment Exercises 2

3.7. SUMMARY

A probability distribution is a mathematical function used in probability theory and statistics that estimates the likelihood that various possible outcomes of an experiment will occur. In terms of its sample space and the probability of events (subsets of the sample space), it is a mathematical description of a random phenomena. The probability distribution of X, for example, would be 0.5 (1 in 2 or 1/2) for X = heads and 0.5 for X = tails (assuming that the coin is fair) if it were used to represent the result of a coin flip (referred to as "the experiment"). Probability distributions are used more frequently to assess the relative frequency of a wide range of random variables. For discrete or continuous

variables, probability distributions can be defined in a variety of ways. Specific designations are given to distributions with unique characteristics or for purposes that are particularly crucial.

3.8. 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.

3.9. Possible Answers to SAEs

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

Answers to SAEs 1

When each trial contains a discrete number of potential outcomes, the multinomial distribution shows the likelihood of getting a particular set of counts. The multinomial distribution of response patterns is the foundation for the goodness-of-fit test that is the most straightforward.

Answers to SAEs 2

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(at least one of them would be a Nigerian) = 1 – P (None of them is a Nigerian)

For P(None is a Nigerian)h = 0.

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

$$P(atleastoneofthemwouldbeaNigerian) = 1 - 0.42 = 0.58$$

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

$$P\left(h = 1 = \frac{3_{C_1} \times 9 - 3_{C_{2-1}}}{9_{C_2}} = \frac{3_{C_1} \times 6_{C_1}}{9_{C_2}} = \frac{18}{36} = 0.5\right.$$

Answers to SAEs 3

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

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

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

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

Answers to SAEs 3

Beta distribution reduces to a uniform distribution in exceptional instances, while gamma distribution reduces to an exponential distribution. The negative binomial distribution is a generalization of the geometric distribution, just as the gamma distribution is a generalization of the exponential distribution.

Module 2: Hypothesis Testing

This module introduces you to Hypothesis testing. The module consists of 5 units which include: meaning of hypothesis, the criterion of significance, statistical test for hypothesis, testing differences between two means and testing differences between matched samples.

Unit One: Meaning of Hypothesis

Unit Two: The Criterion of Significance

Unit Three: Statistical Test for Hypothesis

Unit Four: Testing Differences between Two Means

Unit Five: Testing Difference between Matched Samples

Unit one MEANING OF HYPOTHESIS

Unit Structure

1.1. Introduction
1.2. Learning Outcome
1.3. Statistical Hypothesis
1.4. Type 1 and Type 2 Errors
1.5. Summary
1.6. References/Further Readings/Web Resources
1.7. Possible Answers to Self-Assessment Exercises (SAEs)

1.1. 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.

3.2. Learning Outcome

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

i. know the meaning of hypotheses testing
ii. understand type 1 and type 2 errors
iii. calculate one-tailed and two-tailed tests
iv. calculate various hypothesis testing
v. Apply hypotheses testing to economics problems.

1.2 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**$(\boldsymbol{H_0})$ ∴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**$(\boldsymbol{H_0})$**:** 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,$(\mu_1 \neq \mu_2)$.Sothis 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 Exercises 1**

What are null and alternative hypotheses?

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 $(\beta)$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 Exercises 1**

Differentiate between the following terms between type 1 and type 2 errors

 **3.5. SUMMARY**

To determine if the available data sufficiently support a certain hypothesis, a statistical hypothesis test is a technique of statistical inference. We can use hypothesis testing to make probabilistic claims about the characteristics of the population.

 **3.6. REFERENCES/Further Reading**

Adedayo, O. A (2000). *Understanding Statistics*, 1st Edition, JAS publisher

Akoka, Lagos.

Koka, D. H. (2019). *Introduction to Statistical Analysis*, 1st Edition, Mac

Publisher

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

Statisticians test hypotheses using null and alternate hypotheses. The alternative hypothesis of a test expresses your research's prediction of an effect or relationship, whereas the null hypothesis of a test consistently predicts no effect or no association between variables.

Answers to SAEs 2

Rejecting a null hypothesis that is actually true in the population results in a type I error (false-positive); failing to reject a null hypothesis that is actually untrue in the population results in a type II error (false-negative).

Unit Two: THE CRITERION OF SIGNIFICANCE

Unit Structure

2.1. 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 date 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.

2.2. Learning Outcome

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

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

2.3. One-tailed and Two-tailed tests

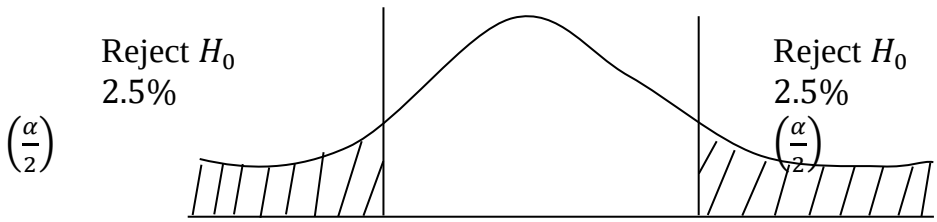
2.3.1. One-tailed test

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.

2.3.2. 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:

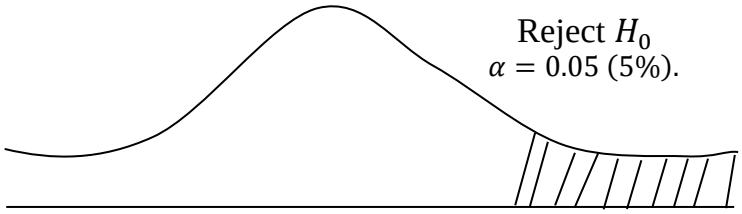


Two – tailed.

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 Exercises 1

Differentiate between one-tailed and two tailed tests.

2.4. SUMMARY

In this unit, there has been a lot of discussion on one and two tailed tests, and their graphs have also been looked at. Therefore, we can infer that one-tailed tests are used for one side of an asymmetric distribution with two tails, such as the normal distribution, which is common in estimating location, or for an asymmetric distribution with a single tail, such as the chi-squared distribution, which are common in measuring goodness-of-fit. This corresponds to specifying a direction. Only when there are two tails, such as in the normal distribution, are two-tailed tests appropriate, which equates to considering either direction significant.

2.5. REFERENCES/Further Reading

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

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

2.6. Possible Answers to SAEs

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

Answers to SAEs 1

The main difference between one-tailed and two-tailed tests is that one-tailed tests will only have one critical region whereas two-tailed tests will have two critical regions. If we require a $100(1-\alpha)$ $100\ (\ 1-\alpha\ )$ % confidence interval we have to make some adjustments when using a two-tailed test.

Unit Three: STATISTICAL TEST FOR HYPOTHESIS

Unit Structure

3.1. 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.

3.2. Learning Outcome

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

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

3.3. 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.

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

When the variances are known and the sample size is large, a z-test is a statistical test to assess whether two population means are different. A hypothesis test known as a z-test is one in which the z-statistic exhibits a normal distribution. A z-statistic, also known as a z-score, is a numerical representation of the outcome of a z-test.

In this situation, the population for which inferences is to be made is assumed to be normally distributed with mean$(\mu)$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.

Self-Assessment Exercises 2

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.

3.4. Statistical Test of a Single Population when the Population Variance is known

When we are either aware of the population variance (2) or when we have another option, we can test a hypothesis statistically using z-tests. Although our sample size, n30 n 30 '", is big, we are unaware of the population variance. When comparing a single population to a norm, such as to see if a town's average lifespan deviates from the national average, one-sample t-tests are performed.

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

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

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

Self-Assessment Exercises 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.

3.5. Interval Estimation for Mean of a Single Population

In statistics, interval estimate is the process of determining the interval, or range of values, within which a parameter, such as the mean (average) of a population, is most likely to fall. So, 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.

Self-Assessment Exercises 3

Suppose N = 25,$\bar{x}$ =20.4, $s = 8.0$, then $S_x = \frac{S}{\sqrt{n}} = \frac{8.0}{\sqrt{25}} = \frac{8.0}{5} = 1.6$
At $x = 0.05$, 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 $\bar{x} - t_{s_{\bar{x}}} \leq \mu \leq \bar{x} +$

3.6. SUMMARY

The statistical test for hypothesis has been the focus of this lesson, and we may infer that the process of hypothesis testing involves an analyst assessing a statistical hypothesis. The type of data used and the analysis's objectives influence the methodology the analyst uses. To accept or reject the null hypothesis is the objective.

3.7. REFERENCES/Further Reading

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

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

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

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

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.

Answers to SAEs 2

$H_0: \mu = 3.5$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 σ = 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.

Answers to SAEs 3

Confidence internal 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 $16.96\ and\ 23.84$.

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

$$\bar{x} - t_s\bar{x} \leq \mu \leq \bar{x}\ t\ t_{s_{\bar{x}}}$$

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).

Unit Four: TESTING DIFFERENCES BETWEEN TWO MEANS

CONTENTS

Unit Structure

4.1. 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.

4.2. Learning Outcome

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

i. understand the test of difference between two means of independent Samples
ii. Understand the meaning of confidence intervals for Difference between Two means.

4.3. 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:

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

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

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 formular is used when the population variance is known. But however, if the population variances are known, the standard score is used:

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

Where$\sigma_1^2$and $\sigma_2^2$are the two percent or population variances.
However, in some test books, they make use of

$$\frac{\bar{x}_{1-}\overline{x}_2}{\sqrt{\left(\frac{N_1 S_{1+}^2 N_2 S_2^2}{N_1+N_{2-2}}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$

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:

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

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 $\propto = 0.05$and $\propto = 0.01$?

Solution

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

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

Using the formular:

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

$$t = \frac{81-78}{\sqrt{\left(\frac{21\times12\times19\times10}{22+20-2}\right)\left(\frac{1}{22}+\frac{1}{20}\right)}} = \frac{3}{\left(\sqrt{\frac{252+190}{40}}\right)\left(\frac{1}{22}+\frac{1}{20}\right)}$$

$= \frac{3}{\frac{\sqrt{442}}{40} \times \frac{42}{440}} = 2.92$

Degree of freedom $= (22 + 10) - 2 = 40$.

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.

 **Self-Assessment Exercises 1**

Using brandP petrol for the mean number of kilometres covered by 22 similar kekemarwa 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?

4.4. Confidence intervals for Differences between Two means

The estimate of the absolute difference in means of the desired outcome variable between the comparison groups is given by the confidence interval for the difference in means. Making a determination as to whether there is a statistically significant difference between comparison groups is frequently interesting.

The estimate of the absolute difference in means of the desired outcome variable between the comparison groups is given by the confidence interval for the difference in means. Determining whether there is a statistically significant difference between comparison groups is frequently interesting. Whether the observed difference is more than what may be predicted by chance will determine this assessment.

The confidence intervals for the difference in means provide a range of likely values for (μ_1-μ_2). It is important to note that all values in the confidence interval are equally likely estimates of the true value of (μ_1-μ_2). If there is no difference between the population means, then the difference will be zero (i.e., (μ_1-μ_2).= 0). Zero is the null value of the parameter (in this case the difference in means). If a 95% confidence interval includes the null value, then there is no statistically meaningful or statistically significant difference between the

groups. If the confidence interval does not include the null value, then we conclude that there is a statistically significant difference between the groups. Since none of the confidence intervals contain the null value, zero, there is a statistically significant difference in means between men and women for each of the attributes in the aforementioned table. Be aware, however, that although the 95% confidence intervals do not include zero, certain of the means (such systolic and diastolic blood pressure) are not significantly different between men and women. Accordingly, the means differ slightly, but statistically in a way that is significant. If the sample size is sufficiently big, as it is in this example, it may be able to show that the differences between groups are statistically significant even when they are slight.

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

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

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

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

Self-Assessment Exercises 2

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.

4.5. SUMMARY

The point estimate has the form x1x2 when analyzing the variance between two means, and Equation 5.3 is once more used to represent the standard error. The difference in sample means under the null hypothesis constitutes the null value, to sum up. The test statistic Z is employed, just like in Chapter 4, to figure out the p-value.

4.6. REFERENCES/Further Reading

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

Lagos
Babatunde, K. A. (2021). *Statistics for Economics*. 1st Edition, MAL Publication.

Datel, R.O. (2013). *Statistics for Business and Management Studies*, 2nd

Edition, Merrigon Press Company limited.

4.7. Possible Answers to SAEs

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

Answers to SAEs 1

The word "better" implies that the hypothesis is directional. So

$$H_0: \mu_1 = \mu_2$$

and the difference is due to chance versus $H_0: \mu_1 > \mu_2$

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

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{\sqrt{s_1^2 + s_2^2}}{N}}$$

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 $\propto$ level of 0.05 is 2.021 and for 60 is 2.00.

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

Value for $40 + \frac{value\ for\ 60 - value\ for\ 40}{60-40} \times (42.40)$

$$= 2.021 + \left(\frac{2.000 - 2.021}{20}\right) x^2$$

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 $\overline{x}_1 - \overline{x}_2$ 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. $(\sigma_1^2 = \sigma_2^2)$.

Answers to SAEs 2

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

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

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

$$= [(81 - 78) - 2.02 \times 1.03)] \leq \mu_1 - \mu_2 \leq [(81 - 78) + 2.02 \times 1.03)]$$

$$(3 - 2.08 \leq \mu_1 - \mu_2 \leq 3 + 2.08$$

$$0.92 \leq \mu_1 - \mu_2 \leq 3 + 2.08$$

$$\therefore 0.92 \leq \mu_1 - \mu_2 \leq 5.08$$

which is the 95% confidence interval at $\propto = 0.01$ 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

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

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

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

Unit Five: TESTING DIFFERENCE BETWEEN MATCHED SAMPLES

Unit Structure

5.1. 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.

5.2. Learning Outcome

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

i. know how to calculate the mean of the population of different score of t statistic
ii. understand variance of the difference of scores

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

Pair	1st born IQ-Score ($\overline{x}_1$)	2nd born IQScore ($\overline{x}_2$)
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:

$$H_0: \mu_0 = 0 \; vs \; H_1: \mu_0 \neq 0$$

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:

$$\frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}}$$

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

So the statistics become:

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

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

$$S_\Delta^2 = \Sigma \frac{\left(\Delta - \overline{\Delta}\right)^2}{N-1}$$

and N is the number of samples.

Self-Assessment Exercises 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.

5.4. SUMMARY

The variance of the difference scores and the mean of the distinct means have demonstrated that the hypothesis can be tested at two different levels of confidence, 95% and 99%. However, utilizing the t statistic and variance of mean formular, there is a discrepancy between their computation and conclusion.

5.5. REFERENCES/Further Reading

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

Wesley, H.F. (2010). *Statistics and Economics, a broader approach*, 1st Edition, Queror publication limited.

5.6. Possible Answers to SAEs

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

Answers to SAEs 1

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

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

The difference between means is as follows:

Weight After	Weight Before	$\Delta = x_1 - x_2$	$(\Delta - \overline{\Delta})$	$(\Delta - \overline{\Delta})^2$
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

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

The variance of the difference of scores

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

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

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

(a) At 95% confidence level, $\propto = 0.05$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 $\propto = 0.01$ 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.

Module 3: Sampling Theory

This module introduces you to Sampling theory. The module consists of 5 units which include: population and sample, population parameters, sampling parameters, Calculation of Sampling Distribution and Estimators for Mean Variance and frequency distribution

Unit One: Population and Sample
Unit Two: Population Parameters
Unit Three: Sampling Parameters

Unit Four: Calculation of Sampling Distribution and Estimators for Mean Variance
Unit 5Five: Frequency Distribution

Unit One: POPULATION AND SAMPLE

Unit Structure

1.1. Introduction
1.2. Learning Outcome
1.3. Meaning of Sampling
1.4. Random Samples and Random Numbers
1.5. Summary
1.6. References/Further Readings/Web Resources
1.7. Possible Answers to Self-Assessment Exercises (SAEs)

1.1. 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. These 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.

1.2. Learning Outcome

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

i. know the meaning of sampling
ii. Understand random samples
iii. Understand random population parameters

1.3 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.

1.3.1 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.

1.3.2 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 1**

Discuss the use of sampling with replacement and without replacement

1.4 Random Samples and Random Numbers

Every person is given a number according to the random number procedure. The next step is to choose a subset of the population at random using a random number generator or random number tables. To create random numbers, you can also utilize Microsoft Excel's random number function (RAND).

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 Exercises 2**

Differentiate between a Random samples and Random numbers

1.5. SUMMARY

On the other hand, we can infer that the population is the full set from which a statistical sample is taken. Statistics experts can create hypotheses about the wider population using the information they learn from the sample. Due to the challenge of examining the complete population, researchers collect data from samples. While the sample is typically represented by a lowercase 'n,' the population is typically represented by a capital 'N' in statistical formulae.

1.6. REFERENCES/Further Reading

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

1.7. Possible Answers to SAEs

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

Answers to SAEs 1

Sampling without replacement is the process of selecting a subset of observations at random; once an observation is chosen, it cannot be chosen again. sampling with replacement, in which an observation may be chosen more than once and a subset of observations is chosen at random.

Answers to SAEs 2

Random and non-random sampling are the two primary types of sampling. The sampling strategy known as random sampling is one in which there is an equal chance of selecting each sample.

The sample that is randomly selected is a fair representative of the entire population. If the sample selected at all does not accurately reflect the population, sampling error results. Non-random sampling refers to a sampling method where the sample is chosen based on criteria other than pure chance. In other words, bias exists in non-random sampling.

Unit Two: POPULATION PARAMETERS

Unit Structure

2.1. 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).

2.2. Learning Outcome

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

i. Analyse the population parameter
ii. Understand sample statistics.

2.3 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.

Self-Assessment Exercises 1

What are population statistics parameters?

2.4. 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 $X200$ 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 $x_{1,}x_2 \dots \dots x_n$ of the random variables $X_1, X_2, \dots \dots X_n$ This case of sampling with replacement, $X_1, X_2, \dots \dots X_n$ would be independent, identically distributed random variables having probability function $f(x)$ and their joint distribution is as follows:

$$P(x_1 = x_{11}\, X_2 = x_2 \dots \dots X_n = x_n) = f(x_1)f(x_2) \dots \dots f(x_n)$$

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 $X_1, X_2, \dots \dots X_n$that is $g(X_1, \dots \dots X_n)$ The function $g(X_1, \dots \dots X_n)$ is another random variable, whose values can be denoted by $g(X_1, \dots \dots X_n)$.

Self-Assessment Exercises 2

What is Sample Statistics

2.5. SUMMARY

A population parameter is known when the related random variable X's probability distribution, let's say f(x), is known, whereas a sample statistic is any number collected from a pooling of samples in order to estimate a population parameter.

2.6. REFERENCES/Further Reading

Adedayo, O. A (2000).*Understanding Statistics*, 1st JAS publisher Akoka, Lagos

Murray, S., & John, S., & Alu, S. (2001) *Probability and Statics*, 1st Edition Schaum's easy Outlines, Macgrw Hill Publication Company,New York.

2.7. Possible Answers to SAEs

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

Answers to SAEs 1

In statistics, a population refers to all the members of a group of people or things. A population can be large or small depending on what you are interested in studying. A parameter is data that describes the entire population, while a statistic is data that describes a sample of that population

Answers to SAEs 2

Any value calculated from your sample data is referred to as a sample statistic (or simply a statistic). The sample average, median, sample standard deviation, and percentiles are a few examples. Because a statistic is based on data gathered

through random sampling, which is a random experiment, it is a random variable.

Unit Three: SAMPLING PARAMETERS

Unit Structure

3.1. 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.

3.2. Learning Outcome

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

i. Understand the meaning of sampling Error
ii. Understand sampling distribution
iii. Understand sampling distribution of mean.

3.3. 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.

3.3.1. 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 1

What is the difference between sampling error and sampling Distribution

3.4. 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 $(X_1, X_2, \ldots\ldots X_n)$. The frequency distribution of the sample mean is called sampling distribution of the mean. Moreover, the mean of this distribution$\mu_{\bar{x}_1}$, when computed will be equal to the mean of the N population(μ)i.e.

$$\mu_{\bar{x}} = \frac{\bar{X}_{1+}\bar{X}_2 t \ \ldots\ldots\ t\ \bar{X}_n}{n} = \mu$$

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:

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

 **Self-Assessment Exercises 2**

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?

 **3.5. SUMMARY**

The sampling parameter has been explored in relation to the sample distribution, sampling mean, and sampling error. However, this unit also looked at instances of how to compute the sampling distribution's mean and variance.

 **3.6. REFERENCES/Further Reading**

Adedayo, O. A (2000).*Understanding Statistics*, 1st Edition, JAS publisher Akoka, Lagos

Murray, S., & John, S., & Alu, S. (2001) *Probability and Statics*, 1st Edition Schaum's easy Outlines, Macgrw Hill Publication Company,New York.

 **3.7. Possible Answers to SAEs**

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

Answers to SAEs 1

Sampling distributions represent sampling error and precision using the width of the curves. Tighter distributions represent lower error and more precise estimates because they cluster more tightly around the population value.

Answers to SAEs 2

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

$$5_{C_2} = \frac{5!}{3!\,2!}$$

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.

Unit Four: CALCULATION OF SAMPLING DISTRIBUTION AND ESTIMATORS FOR MEAN VARIANCE

Unit Structure

4.1. 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.

4.2. Learning Outcome

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

i. Estimate the mean and variance from a population
ii. Calculate the sampling distribution of proportion
iii. Calculate the sampling distribution of Differences and sums

4.3. 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:

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

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.

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

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

Self-Assessment Exercises 1

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.

4.4. Sampling Distribution of Proportion

The population proportion (p) is the mean of all sample proportions (p-hat) when repeated random samples of a given size n are collected from a population of values for a categorical variable, where the proportion in the category of interest is p. By dividing the total number of successes, or chances, by the sample size "n," the Sampling Distribution of Proportion calculates the proportion of success, or a probability that specific events will occur. As a result, p = x/n is used to define the sample proportion. If the random sample of 'n' is obtained using replacement, the proportion sampling distribution obeys the binomial probability law. For instance, if the population is infinite and the chance of an event occurring is "," then the probability of the event not occurring is "1-." Now calculate the proportion of success 'p' for each of the conceivable sample sizes 'n' selected from the population. The sampling distribution of proportion obeys the binomial probability law if the random sample of 'n' is obtained with replacement. Such as, if the population is infinite and the probability of occurrence of an event is **'π',** then the probability of non-occurrence of the event is **(1-π).** Now consider all the possible sample size 'n' drawn from the population and estimate the proportion 'p' of success for each

Therefore, 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

$\text{Var}\left(\hat{P}\right) = \frac{\hat{P}(1-\hat{P})}{n}$ Theample proportion is also an unbiased estimator of the population proportion.

Self-Assessment Exercises 2

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

4.5. Sampling Distribution of Differences and Sums

If we are given two normally independent random variables $x_1 and\ x_2$whose means are $\mu_{P_1} and\ \mu_{P_2}$ respectively and variances $\sigma_{P_1} and\ \sigma_{P_2}$, 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

and $\sigma_{\bar{x}_1-x_2} = \sqrt{\sigma_{\bar{x}_1} + \sigma_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

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:

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

 **Self-Assessment Exercises 3**

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

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

 **3.5. SUMMARY**

It is the probability distribution of the given statistic, as obtained in statistics using a random sample. It offers a broad method for drawing conclusions from statistics. The parameter used to determine sample statistics is called an estimator. This means that the sample size, N, or the number of scores used to compute the mean, is equal to the population variance divided by N, the sample size. The variance of the sampling distribution of the mean hence decreases as sample size increases.

 **3.6. REFERENCES/Further Reading**

Adedayo, O. A (2000).*Understanding Statistics*, 1st Edition, JAS publisher

Akoka, Lagos

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

Population mean μ = 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:

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

Answers to SAEs 2

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% of tosses $= \frac{38}{100} \times 150 = 57$ while $58\% = 87$
μ = expected number of heads $= np = 150 \times \frac{1}{2} = 75.$

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

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

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

Answers to SAEs 3

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 $90, 130, 110, and\ 140.$

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: $-10 + (-50) + 10 + (-30)$
$=-10 - 50 + 10 - 30 = -80$

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

Therefore

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

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

$$\sigma^2{}_{P_1+P_2} = \frac{\Sigma(x-\overline{x})^2}{N}$$

$$\sigma_{P_1+P_2} = \frac{1}{4}[(90 - 120)^2 + (130 - 120)^2 + (110 - 1\,20)^2 + (150 - 120)^2\,]$$
$$= 500$$

$$\therefore \sigma_{P_1}^2 = variance\ of\ (40{,}60)$$

$$= \frac{1}{2}\left[\left(40-50\right)^2 + \left(60-50\right)^2\right] = \frac{1}{2}(100+100) = 100$$

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

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

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

Unit Five: FREQUENCY DISTRIBUTION

Unit Structure

5.5. Summary
5.6. References/Further Readings/Web Resources
5.7. Possible Answers to Self-Assessment Exercises (SAEs)

5.1. 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.

5.2. Learning Outcome

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

i. Analyse the frequency distribution and frequency polygon
ii. Know the relative frequency distributions
iii. Understand the calculation of ungrouped data for frequency Distributions.

5.3. Analysis of Frequency Distribution

Frequency Distribution: What Is It? A representation of the number of observations inside a specific interval, either graphically or tabulatedly, is called a frequency distribution. The difference between frequency and distribution is how frequently a value appears within an interval. A representation of the number of observations inside a specific interval, either graphically or tabulatedly, is called a frequency distribution. The difference between frequency and distribution is how frequently a value appears within an interval. The interval size is determined by the data being examined and the analyst's objectives. The intervals must be thorough and exclusive of one another. In a statistical context, frequency distributions are frequently utilized. In general, a normal distribution chart can be used to represent frequency distributions.

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 61.5 and 62.5 inches, and we can go on to record the class interval as 59.5 – 62.5. The next class interval would then be 62.5 – 65.5, etc. However, the class interval 59.5 – 62.5, the number 59.5 and 62.5 are often called class boundaries. The width of the jth class interval, denoted by Mj, 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=62.5 – 59.5=3.

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

 **Self-Assessment Exercises 1**

What is a frequency distribution?

5.4. Frequency Polygon

The shape of the data and the patterns that a specific data collection exhibits are depicted by frequency polygons. Let's have a look at how to construct a frequency polygon step by step, either with or without a histogram. Data can be represented graphically using frequency polygons. It is used to show trends and the shape of the data. Although it can be drawn without one, it is typically drawn with the aid of a histogram. A histogram, which is used to illustrate frequency distributions, is a collection of rectangular bars with no spaces between them.

Steps to Draw a Frequency Polygon

1. Label the horizontal axis with the class intervals for each class. The frequency will be plotted on the vertical axis.

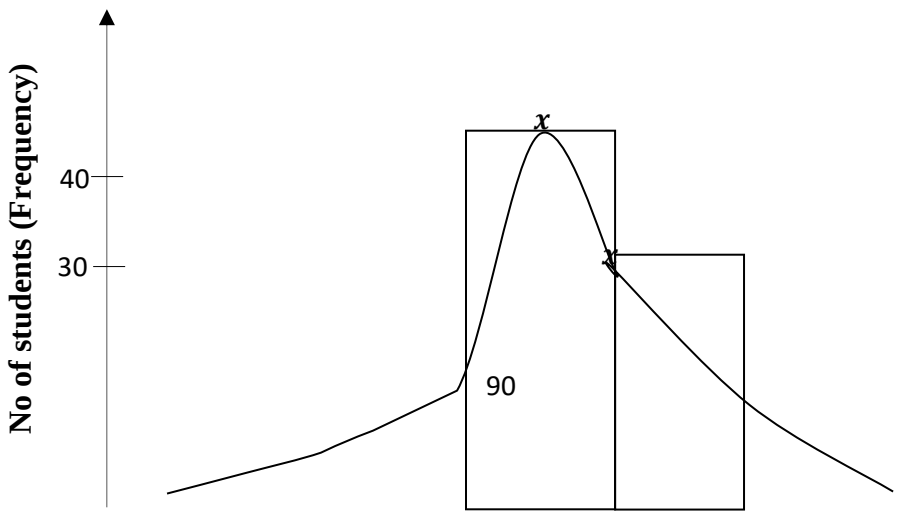
2. Determine the grade for each class period. Calculating a class grade is as follows:
Classmark is equal to (Upper limit minus Lower limit) / 2.
3. Label the horizontal axis with all of the class grades. It is sometimes referred to as the midpoint of each class.
4. Plot the frequency that was given to you in accordance with each class mark. The frequency is always shown by the height. Be sure to plot the frequency against the class mark rather than any class's upper or lower limit.
5. Use a line segment to connect every point that was plotted. It will result in a kinked curve.

6. This curve is known as the frequency polygon.

Keep in mind that the method described above is used to create a frequency polygon without creating a histogram. A histogram can alternatively be created by first placing rectangular bars against the specified class intervals. In order to acquire the frequency polygon, you must then combine the midpoints of the bars. In a histogram, the bars won't be separated by any spaces.

Further, 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



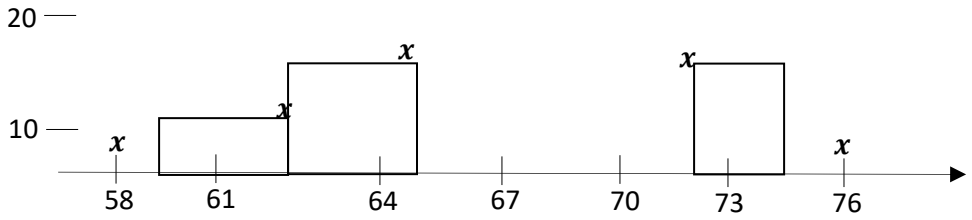


Figure 1 Showing Frequency Polygon

5.4.1. Relative Frequency Distributions

Relative frequency distributions are tabular summaries of data sets that display the relative frequency of items in various non-overlapping classes. The percentage or portion of all items in a class that make up the relative frequency is called a fraction or proportion.

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.

5.4.2. 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 ~~IIII~~. 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	~~IIII~~ II	7
51 – 60	~~IIIIIIII~~ I	11
61 – 70	~~IIIIIIII~~ I	11
71 – 80	~~IIII~~ III	8
81 – 90	~~IIII~~ II	7
91 – 100	I	1

 **Self-Assessment Exercises 2**

1. What is the concept of frequency polygon?
2. What is frequency distribution?

 **5.5. SUMMARY**

A frequency distribution is a representation that shows the number of observations inside a specific interval, either in a graphical or tabular manner. We have discussed frequency distribution in this unit, and we inferred from the

analysis above that this is what it is. The intervals must be thorough and exclusive of one another. In a statistical context, frequency distributions are frequently utilized.

5.6. REFERENCES/Further Reading

Adedayo, O. A (2000).*Understanding Statistics*, 1st Edition, JAS publisher Akoka, Lagos

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

5.7. Possible Answers to SAEs

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

Answers to SAEs 1

Definition. Frequency distributions are visual displays that organise and present frequency counts so that the information can be interpreted more easily. Frequency distributions can show absolute frequencies or relative frequencies, such as proportions or percentages.

Answers to SAEs 2

1. A line graph showing class frequency plotted against class midpoint is known as a frequency polygon. You can get it by connecting the midpoints of the rectangles' tops in the histogram.
2. A frequency distribution is a visualization of the number of observations within a certain interval that can be either graphical or tabular in nature. The distribution is the pattern of the variable's frequency, whereas the frequency is how frequently a value occurs within an interval.

Module 4: T-Test, F-Test and Chi-Square Analysis

This module introduces you to T-Test, F-Test and Chi-Square Analysis. The module consists of 3 units which include: meaning of econometrics, methodology, computer and econometrics, basic econometrics models: linear regression and importance

Unit One: T-Test
Unit Two: F-Test
Unit Three: Chi-Square

Unit One: T-TEST

Unit Structure

1.1. Introduction
1.2. Learning Outcome
1.3. History of T Statistics
1.4. Unpaired and paired two Sample t tests
1.5. The t test formula
1.6. Summary
1.7. References/Further Readings/Web Resources
1.8. Possible Answers to Self-Assessment Exercises (SAEs)

1.1. 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.

1.2. Learning Outcome

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

i. Understand the history of t statistics
ii. Know the unpaired and paired two sample t tests.
iii. Understand the t test formula

1.3. 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.

Gossest 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.

1.3.1 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.

1.3.2. 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 Exercises 1**

What are the assumptions of the t-test?

1.4. 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.

1.4.1. Independent (Unpaired) Samples

The independent samples t-test is used when two separates 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.

1.4.2 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".

1.4.3. The T-Test Formula

The T test formula is given as:

$$t = \frac{\bar{x}_{1-}\bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^2}{n_2}}}$$

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

Formulae for mean

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

Formulae for standard deviation

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

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 2

1. What is the unpaired two-sample t-test?
2. What does the t-test value tell you?

3.5. SUMMARY

The p-value (probability) that can be used to assess if the population differs is determined by the t-test, which considers the t statistic, t-distribution, and degree of freedom. This unit has led us to the conclusion that the test statistic in the t-test is known as the t-statistic.

3.6. REFERENCES/Further Reading

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.*

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

The distribution of populations is the fundamental presumption used in the majority of parametric tests. The following prerequisites must be met in order to perform a t-test: ratio scale or interval scale measurements; simple random extraction; normal data distribution; a suitable sample size; and homogeneity of variance.

Answers to SAEs 2

1. The mean of two independent groups is compared using the unpaired two-samples t-test. Consider the situation where 100 people had their weight measured: 50 women (group A) and 50 men (group B). If women's mean weights (mA) and men's mean weights (mB) differ greatly, that would be desirable to know.

2. The t-value measures the size of the difference relative to the variation in your sample data. Put another way, T is simply the calculated difference represented in units of standard error. The greater the magnitude of T, the greater the evidence against the null hypothesis

Unit Two: F TEST

Unit Structure

2.1. Introduction
2.2. Learning Outcome
2.3. F tests
2.3.1.. Formulae and analysis of f test Statistics

2.5. Summary
2.6. References/Further Readings/Web Resources
2.7. Possible Answers to Self-Assessment Exercises (SAEs)

2.1. 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. Snedeaor but he used it to honour Sir Ronald A. Fisher and Fisher initially developed the statistics as the variance ratio in 1920.

2.2. Learning Outcome

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

i. Know the various examples of f tests Statistics
ii. Understand formulae and analysis of f tests Statistics

2.3. F-Tests

Any statistical test with an F-distribution for the test statistic under the null hypothesis is known as an F-test. In order to determine which statistical model better represents the population from which the data were sampled, it is most frequently applied when contrasting models that have been fitted to data sets. Exact "F-tests" are typically required after least squares fitting of the models to the data. In honor of Ronald Fisher, George W. Snedecor came up with the name. In the 1920s, Fisher created the statistic as the variance ratio. The analysis of the following cases is one frequent example of the application of F-tests:
the idea that a group of normally distributed populations with the same standard deviation all have the same means. This F-test, which is arguably the most well-known, is crucial to the analysis of variance (ANOVA), which seeks to determine whether a suggested regression model adequately accounts for the data. The theory that a data set in a regression analysis follows the simpler of two proposed linear models that are nested within each other is called the Lack-of-Fit Sum of Squares.

Additionally, certain statistical techniques also employ F-tests, such as Scheffé's approach for multiple comparisons adjustment in linear models.

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.

2.3.1. 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 $\propto$ we cannot state that the success achievement pair with the greatest mean difference is significantly different at level $\propto$.

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

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

or

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

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

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

Where $Y_i$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 $yij$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, is the residual sum of squares of model i. If your regression model has been calculated with weights, then replace Rss_iwithX^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 Exercises 1

What is the difference between ANOVA and F-test?

3.5. SUMMARY

The F test is widely used in statistical analysis and is used to determine whether a set of parameters is generally significant. As a result, ANOVA is employed in regression analysis and is specifically mentioned in the F test formulas.

3.6. REFERENCES/Further Reading

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

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

The F-test is the ratio of the mean squared error between these two groups, and ANOVA separates the within-group variance from the between-group variance.

UNIT 3 CHI-SQUARE ANALYSIS

Unit Structure

3.1. Introduction

3.2. Learning Outcome
3.3. Chi-square distribution
3.4. Application of Chi-square analysis
3.5. Summary
3.6. References/Further Readings/Web Resources
3.7. Possible Answers to Self-Assessment Exercises (SAEs)

3.1. INTRODUCTION

A Chi-square can said to be a measurement of how expectations are compared to results. The date 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.

3.2. Learning Outcome

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

i. Know the examples of chi-square distribution
ii. Understand the application of chi-square analysis

3.3. 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^2which 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^2calculated value to thetabulated 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 constrasted 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 sic 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.

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

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

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

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$(r - i)(n - 1)$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.

$$X^2_{Yates} = \sum_{i-1}^{N} \frac{(1O_i - E_i 1 - 0.5)^2}{E_i}$$

where:

O_i= observed frequency

E_i = Expected (theoretical) frequency

N = Number of events.

(4) Cochram – 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 ransom 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 dis-equilibrium.

(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 1

What are examples of a chi-square test?

3.4. Application of Chi-Square Analysis

3.4.1 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.

$3.4.2.$ 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 $2x2$ 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$(\boldsymbol{Ho})$ **:** The female women whose heart rate increased are independent of drug treatment.

Alternative Hypothesis ($\boldsymbol{H_1}$): 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:

$$= \frac{135\,[(146)(31)-(18)(40)]^2}{(64)(71)(49)(80)}$$

$$= 135\,\frac{[(1426-720)]^2}{(45.44)(3920)}$$

$$= \frac{135(706)^2}{17812480}$$

$$= \frac{6728860}{17812480}$$

$= 3.77$

The next step is to know the degree of freedom, therefore the degree of freedom equal (number of columns minus one) × (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 $(\propto = 0.05)$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.

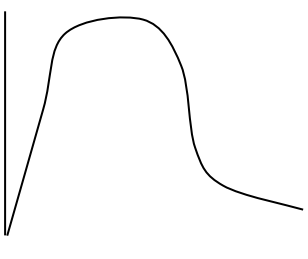
But in the care of our example, the chi-square calculated value is 3.77 and chi-square tabulated is 3.841, therefore, the chi0squared tabulated (3.841) is greater than the chi-square calculated (3.77) then conclude that the female women whose heart increased is independent of drug treatment.

Table 4

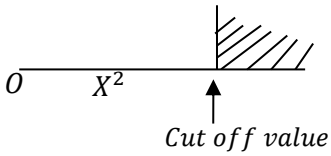
The table below is an example of chi-square table.

Probability Level (Alpha)

Degrees of Freedom	Level of Significance (5%)	Level of Significance (1%)



$v = 1$	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



.

3.4.3. 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	**Expected**	$(\boldsymbol{O - E})$	$(\boldsymbol{O - E})^2$	$(\boldsymbol{O - E})^2/\boldsymbol{E}$

	$(\boldsymbol{O})$	$(\boldsymbol{E})$			
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)
$v = 1$	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 Ho and reject Hi then conclude that AA genotype is not better to AS genotype in 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 Ho and Hi 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	a	b	c	$a+b+c$
Sample B	d	e	f	$d+e+f$
Sample c	g	h	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 2

1. What is the meaning of a chi square distribution?
2. What type of data is chi squared used for?

3.5. SUMMARY

A chi-square distribution is a continuous probability distribution. The shape of a chi-square distribution depends on its degrees of freedom, k. The mean of a chi-square distribution is equal to its degrees of freedom (k) and the variance is 2k.

3.6. 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.

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

Based on the data, which must be unprocessed, random, taken from independent variables, taken from a large sample, and mutually exclusive, a Chi-Square statistic test is computed. In plain English, two sets of statistical data are compared, such as the outcomes of flipping a fair coin.

Answers to SAEs 2

1. A continuous probability distribution is a chi-square distribution. A chi-square distribution's k degrees of freedom determine its shape. A chi-square distribution has a mean equal to the number of degrees of freedom (k) and a variance of 2k.

2. The Chi-square test analyzes categorical data. It means that the data has been counted and divided into categories. It will not work with parametric or continuous data. It tests how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

Module 1: Simple Linear Regression Analysis and Its Application

This module introduces you to Simple linear Regression analysis and its Application. The module consists of 3 units which include: meaning of regression analysis, simple linear regression analysis and application of simple regression analysis.

Unit One: Meaning of Regression Analysis
Unit Two: Simple Linear Regression Analysis
Unit Three: Application of Simple Linear Regression Analysis

Unit One: MEANING OF REGRESSION ANALYSIS

Unit Structure

1.1. Introduction
1.2. Learning Outcome
1.3. Uses of Regression Analysis
1.4. Three conceptualization of Regression analysis
1.5. Regression Models
1.6. Summary
1.7. References/Further Readings/Web Resources
1.8. Possible Answers to Self-Assessment Exercises (SAEs)

1.1 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.

1.2. Learning Outcome

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

i. Know the uses of regression analysis
ii. Understand the three conceptualizations of regression analysis
iii. Understand regression models

1.3. 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.

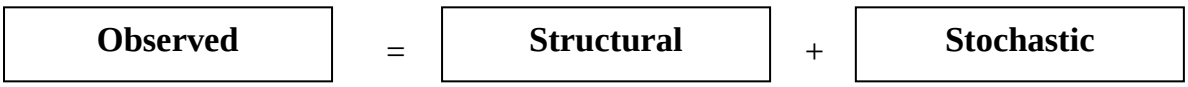
More so, different techniques are used in carrying out the regression analysis and such method such as linear regression and ordinary least squares regression are parametric, and in that we can define regression function in terms of its finite number of unknown parameters that are estimated from the 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.

1.3.1. 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

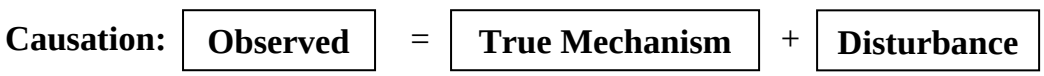
as possible, an array of observed values of the dependent variable based on a simple function of independent variables. Obviously, predicted values from regression models are not exactly the same as observed ones. Characteristically, regression partitions on observation into two parts.

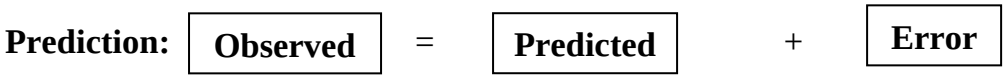


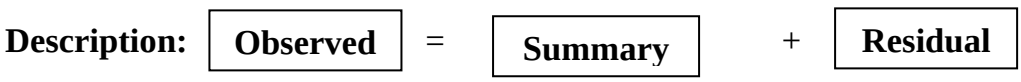
The observed part represents the actual values of the dependent variable at hand. The structural part denotes the relationship between dependent and independent variables.

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.







1.3.2. 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 Exercises 1**

1 Differentiate between Parsimony and accuracy of a model
2 Differentiate between Statistical modelling and statistical variables
3 What is regression model?

3.5. SUMMARY

Regression analysis has various applications in statistical and economic modeling and is described in a wide variety of ways. Regression analysis can also take the form of examining the model's variables and providing predictions for upcoming planning.

3.6. REFERENCES/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) 2(1).*

David, A. F. (2005). *Statistical models, theory and practice*, 1st Edition Cambridge University Press, 2005.

Chian, C. L. (2003).*Statistical Methods of Analysis*, World scientific, ISBN, 981 – 938.

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

1. Parsimonious models are simple models with great explanatory predictive power. They usually explain data with a minimum number of parameters or predictor variables while Model accuracy is defined as the number of classifications a model correctly predicts divided by the total number of predictions made. It's a way of assessing the performance of a model, but certainly not the only way

Answers to SAEs 2

2. Statistical modeling is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world. A statistical model is a collection of probability distributions on a set of all possible outcomes of an experiment while a variable is any characteristic, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables

Answers to SAEs 3

A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.

Unit Two: SIMPLE/LINEAR ANALYSIS AND ITS APPLICATION

Unit Structure

2.1. INTRODUCTION

In statistics, linear regression is an approach for modelling the relationship between s 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.

2.2. Learning Outcome

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

i. Know the uses of simple linear regression
ii. Understand linear regression analysis

2.3. 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:

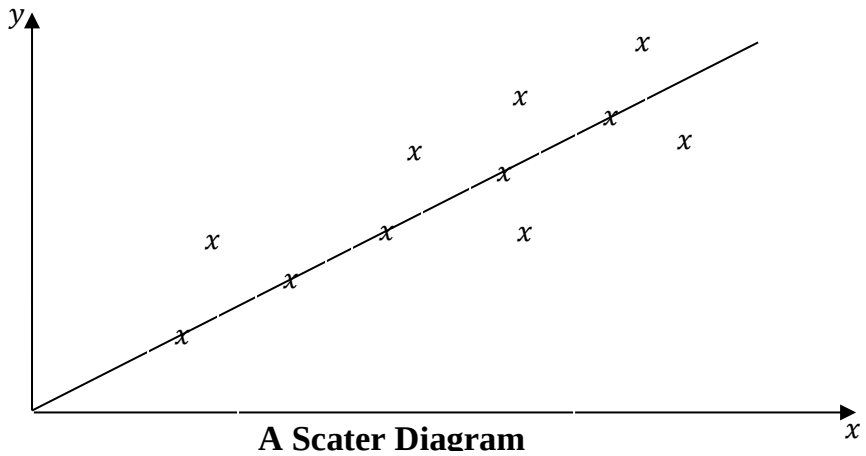
(a) If the goal is prediction or forecasting or reduction, linear regression can be used to fit a predictive model to an observed data set of y and x values and after developing such a model, if an additional value of x is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.

(b) Given a variable y and a number of variables $X_1, \ldots \ldots X_P$that may be related to y, linear regression analysis can be applied to quantity the strength of the relationship between y and theX_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.

2.3.1. 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



A Scater Diagram

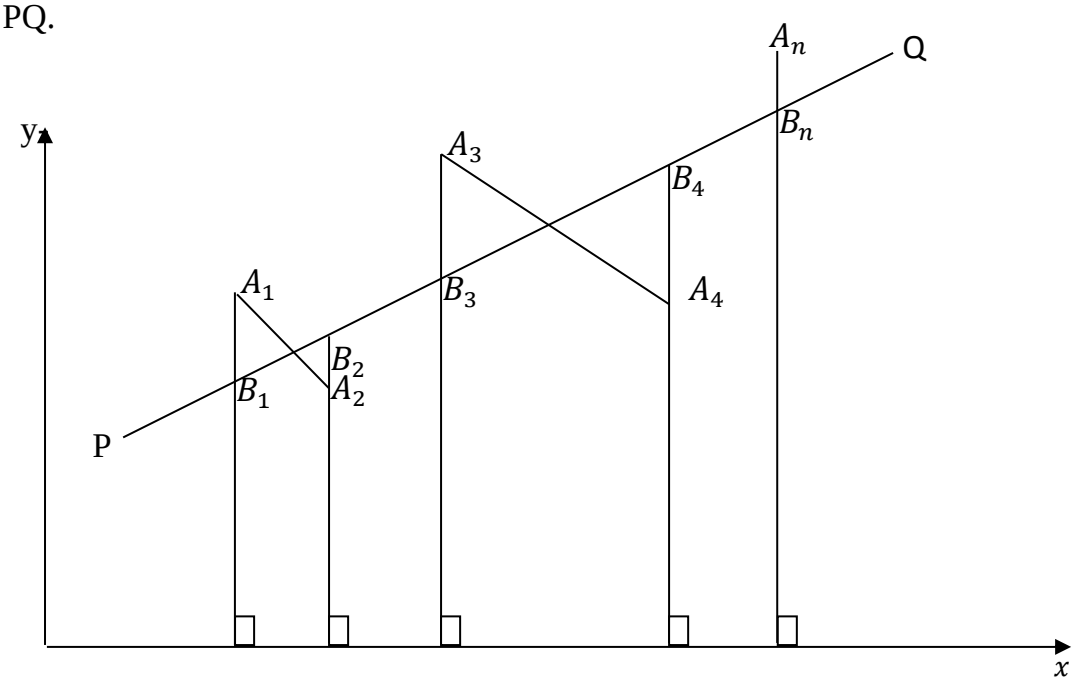
The line of best fit is then drawn to join those points that are in straight line on the diagram together and it is also called regression line given by the equation$y = n_1x + C\ te$where e is the error term in the straight line equation which if minimised will be zero and the result becomes $y = Mx + C$.

There are various methods of fitting the regression line. One of the methods is the free land 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 $A_1(n_1, y_1), A_2(n_2, y_2) \ldots\ldots An\ (X_n, y_n)\ are$ parts of a scattered diagram whose line of best fit is PQ with regression equation $y = Mx + C$.

Let $A_1,\ B_1, A_2B_2 \ldots\ldots\ A_nB_n$be the distance between the points and the diagram in PQ.



Residual A_1B_1has a distance of$[y_1 - (Mx_1 + C)]$,A_2B_2has length $[y_2 - (Mx_2 + C)], A_3, B_3 has$ a distance of$[y_3 - (Mx_3 + C)] \ldots\ldots\ A_nB_n$ has a distance of$[A_N - (Mx_N - C)]$.
The sum of squares of the residuals:

$$S = [y_1 - Mx_1 + C)]^2 + (y_2 - Mx_2 + C)]^2\ t \ldots . t\ [y_n - (Mx_N - C)^2]$$
$$= (y_1 - Mx_1 - C)^2 + (y_2 - Mx_2 - C)^2 t, \ldots (y_n - Mx_n - C)^2$$

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:

$$\frac{j_s}{j_m} = 2X_1(y_1 - Mx_1 - C)$$
$$- 2x_2(y_2 - Mx_2 - C) -, \ldots - 2X_N(y_N - mx_n - C) = 0.$$

When $S\ is\ a\ minimum$

$x_1y_1, +x_2y_2t \ldots x_ny_n) - m(n_1^2 + n_2^2\ t \ldots t\ x_N^2) - C\ (n_1 + n_2\ t \ldots . t\ n_N) = 0$

Using summation notation $(\Sigma)\ we\ get\ \Sigma xy - m\Sigma x^2 - C\Sigma x = 0$ that $\Sigma xy = m\Sigma x^2 + \mathrm{C}\Sigma x.$

More so,

$\frac{ds}{jc} = -2(y_1 - Mx_1 - C) - 2(y_2 - Mx_2 - C) \ldots . -2(y_N - MX_N - C) = 0$

where S is a minimum. Dividing through by $-2\ we\ obtain.$

$$(y_1 - Mx_1 - C) + (y_2 - Mx_2 - C)t \ldots . t(y_N - Mx_N - C) = 0$$

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. Solve using formulae method.

Solution

$$\Sigma x = 15, \Sigma y = 75, \Sigma xy = 202$$
$$\Sigma x^2 = 55, \Sigma y^2 = 11.93$$

$$M \; or \; \beta_1 = \frac{N\Sigma xy - \Sigma x \Sigma y}{\frac{N\Sigma x^2 - (\Sigma x)^2}{=5\times202-15\times75}}$$
$$\frac{}{5\times55-(15)^2}$$

$$= \frac{1010 - 1125}{275 - 225}$$

$$= \frac{-115}{50}$$

$$= -2.3.$$

$$C \; or \; \beta_0 = \frac{\Sigma y \Sigma x^2}{N\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{75 \times 55 - 15 \times 202}{5 \times 55 - (15)^2}$$

$$= \frac{4125 - 3030}{275 - 225}$$

$$= \frac{1095}{50}$$

$$= 21.9$$

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 $\beta_0$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:

$$b = \frac{\Sigma x \Sigma y^2 - \Sigma y \Sigma xy}{N\Sigma y^2 - (\Sigma y^2)}$$

$$= \frac{Corxy}{VARy}$$

$$a = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$$

And b can also be found using$b = \frac{1}{N}(\Sigma x - a\Sigma y)$.

Example 3

Use the data in above to compute the regression of $x\ on\ y$.

Solution

$$N = S, \Sigma x = 15, \Sigma y = 75, \Sigma xy = 202, \Sigma x^2 = 55, \Sigma y^2 = 1193$$

$$a = \frac{N\Sigma xy - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{5 \times 202 - 15 \times 75}{5 \times 1193 - (75)^2}$$

$$= \frac{1010 - 1125}{5965 - 5625}$$

$$= \frac{-115}{340} = -0.34$$

$$b = \frac{1}{N}(\Sigma x - a\Sigma y) = \frac{1}{5}[15 - (-0.34) \times 75]$$

$$= 5.07$$

or

$$a = \frac{\Sigma x \Sigma y^2 - \Sigma x \Sigma y}{N\Sigma y^2 - (\Sigma y^2)}$$

$$= \frac{15 \times 1193 - 75 \times 202}{5 \times 1193 - (75)^2}$$

$$= \frac{17895 - 15150}{340}$$

$$= \frac{2745}{340}$$

$$= 8.07.$$

Therefore the equation is $x = ay + 6\ so,$

$$x = -0.34\ t\ 8.07$$

Self-Assessment Exercises 2

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 $x = 2.5$

3.5. SUMMARY

This unit looked at simple linear regression, and we learned that it is the relationship between the dependent and independent variables. However, several calculations of the straight line graph that calculated regression equation y =_0+_1 X_1ofy on x and x on y were also investigated in this unit.

3.6. REFERENCES/Further Reading

Adedayo, A. O. (2006).*Understanding statistics*, 1st Edition, the Publisher, Akoka, Lagos.

Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, 267-288, Jstor Edition.

Nievergelt, Y. (1994).*Total Least Square, State of the art Regression in Numerical Analysis*, 1st Edition.

3.7. Possible Answers to SAEs

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

Answers to SAEs 1

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

$$\Sigma x_y = m\Sigma x^2 + C\Sigma x \_\_\_\_\_ (1)$$

$$\Sigma_y = m\Sigma x + N_c \_\_\_\_\_ (2)$$

we then obtain

$$202 = 55m + 15c \_\_\_\_\_ (1)$$
$$75 = 15m\ t5c \_\_\_\_\_ (2)$$

$(2)x3\ gives$

$$225 = 45m\ t\ 15c \_\_\_\_\_ (3)$$

$(1) - 3\ gives$

$$-23 = 10m$$
$$m = -2.3$$

Substituting $m = -2.3\ into\ (1) we\ obtain$

$$C - 21.9$$

Therefore, the regression equation is

$$y = mxtc = -2.3x + 21.9.$$

$$x = 2.5$$
$$y = -2.3\ \times 2.5 + 21.9 = 16.15$$

However, the regression coefficient m and the intercept can also be computed using the standard formulas as follows:

$If\ y = mxtc\ is\ the\ regression\ equation\ of\ y\ on\ x\ or\ therefore\ we\ have$:

$$Mor \quad \beta_1 = \frac{N\Sigma xy - \Sigma x\Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$M\ or\ \beta_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \overline{x})^2} = \frac{Covariance\ of\ xy}{VAR\ of\ x}$$

and

$$C\ or\ \beta_o^c = \frac{\Sigma y\Sigma x^2 - \Sigma x\Sigma xy}{N\Sigma x^2 - (\Sigma x)^2}$$

Unit Three: APPLICATION OF SIMPLE LINEAR ANALYSIS

CONTENTS

Unit Structure

3.4. Summary
3.5. References/Further Readings/Web Resources
3.6. Possible Answers to Self-Assessment Exercises (SAEs)

3.1. INTRODUCTION

Application of Simple Linear regression analysis is the way by which we subject different data to statistical analysis by using computer software such strata, e-view 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 both have dependent and independent variables to be able to predict is deduced whether is there any element of impact of unemployment in Nigerian Economy or not.

However, the relationship can either tells 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.

3.2. Learning Outcome

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

i. Understand what a simple regression analysis can deduce from its application.
ii. How to interpret the parameter using to test for t and f-tests respectively.

3.3 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.

3.3.1 How to Interpret the P-Values in Linear Regression Analysis

The P-values for each term tests 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 thep-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 ($\propto$)of 0.05 and this indicates that it is not statistically significant.

3.3.2 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

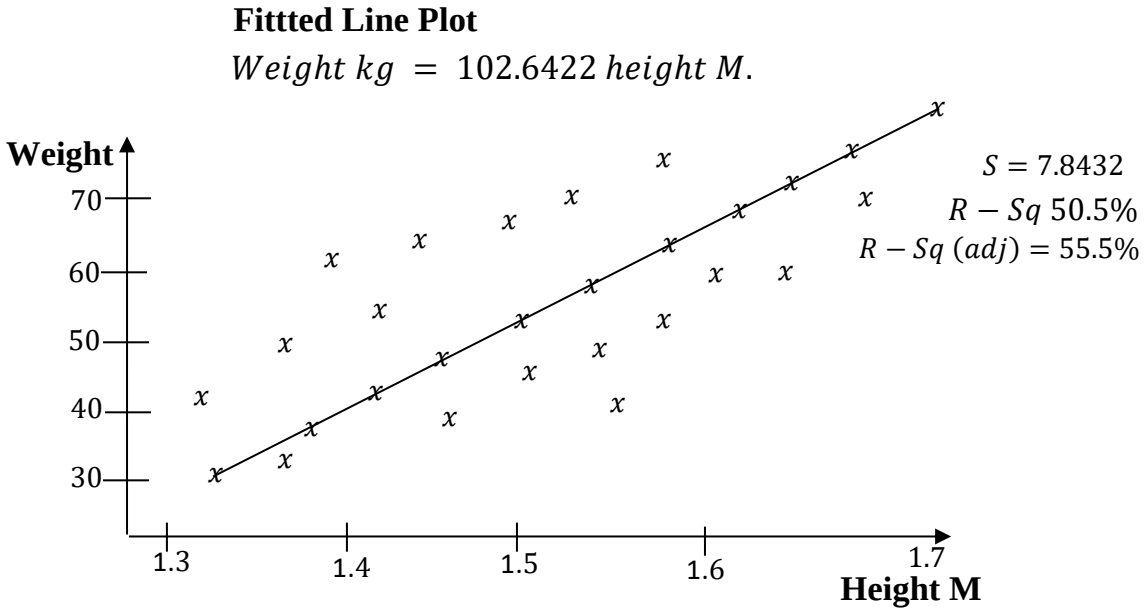
Coefficients

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 theslope 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:

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 –2022
Included observations – 42

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) ..... (1)
GDP = ao + a1GHE + e .....(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 calculated 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 Exercises 1**

Explain what steps you will take in interpreting the predictor for value in a regression analysis result.

 **3.5. SUMMARY**

An important tool in forecasting the future of analysis that may be challenging when looking at its theoretical analysis is understanding how to interpret the p values and coefficient of a regression analysis. Regression models, on the other hand, deal with the circumstances surrounding the research topic and analyze it thoroughly by running the analyses of dependent and independent variables against one another to produce a suggested value for policy makers and planning analysis for the present and the future.

 **3.6. REFERENCES/Further Reading**

Olomeko, O. A. (2012).*Regression Analysis*, 1st edition, pg 88 Millworld

Publication Limited, Lagos.

 **3.7. Possible Answers to SAEs**

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

Answers to SAEs 1

When the independent variable's value rises, the dependent variable's mean tends to rise as well, according to a positive coefficient. When the independent

variable rises, the dependent variable is thought to tend to fall, according to a negative coefficient.

.