NATIONAL OPEN UNIVERSITY OF NIGERIA

COURSE TITLE: APPLIED STATISTICS

COURSE CODE: ECO 726

FACULTY OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS

Course Content Developer
Anthony Ilegbinosa IMOISI (PhD)
Department of Economics
Faculty of Arts, Management and Social Sciences
Edo University Iyamho, Edo State

Course Content Editor
Professor Anthony A. AKAMOBI
Department of Economics, Faculty of Social Sciences
Chukwuemeka Odumegwu Ojukwu University
Igbariam, Anambra State.

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

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

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

 Printed: 2018 ISBN: 978-058-023-X.

CONTENT

Introduction

Course Content

Course Aims

Course Objectives

Working through This Course

Course Materials

Study Units

Textbooks and References

Assignment File

Presentation Schedule

Assessment

Tutor-Marked Assignment (TMAs)

Final Examination and Grading

Course Marking Scheme

Course Overview

How to Get the Most from This Course

Tutors and Tutorials

Summary

INTRODUCTION

Statistical economics is a branch of economics that takes care of analysis of economic phenomenon. In this course, you will be introduced to a number of analytical tools in statistics; At this juncture, you will be exposed to the fundamental assumptions, formulae as well as calculations of the topics under consideration. In addition, you will be taught the decision criteria under each topic.

COURSE CONTENT

This course will expose you to different statistical tools that economist can utilize in economic analysis. This course is built on the foundation of elementary statistics and elementary economics in the understanding of real life situation.

COURSE AIMS

There are twelve study units in the course and each unit has its objectives. You are advised to go through the objectives of each of them and bear them in mind as you go through each of the unit. The general objectives include the following;

- Exposing you to fundamental statistical tools that can be applied in economics,
- Apply these tools to real life situation,
- Expose the students to economic interpretation of all calculated coefficients

COURSE OBJECTIVES

There are general and specific-units objectives of the course. These are set in order to achieve the purpose of this course. The units' objectives are itemized at the beginning of each unit; and students ought to go through them before working through each unit. Students can as well refer to them in the course of their study to make sure that they are keeping up with the pace of the teaching. This will aid the students in achieving the task involved in the course. The objectives act as study guides, such that every student could know if he or she is coming in terms with the knowledge of each unit set objectives.

On successful completion of the course, the student ought to be able to:

• expand the learning horizons of the subject

• apply statistical tools in economic problems

WORKING THROUGH THIS COURSE

This course needs spending quality time to study. The content of this course is comprehensive, wide-ranging and presented in a clear and understandable language. The presentation style is sufficient and the contents are easy to comprehend. To successfully complete this course, it is essential to read the study units, reference materials as well as other materials on the course. Each unit has a self-assessment exercise. Students will be required to submit assignments for assessment purposes and there will be final examination at the end of the

course. Students should take sufficient advantage of the tutorial sessions because it is a good opportunity to share ideas with their fellow course mates. The course will take approximately fourteen weeks and the components of the course are outlined in details under the course material sub-section.

COURSE MATERIALS

The key components of the course are:

1. Course Guide
2. Study Units
3. Textbooks
4. Assignment
5. Presentation Schedule

STUDY UNITS

There are three Modules in this course divided into twelve study units as follows:

MODULE 1: BASIC SAMPLING AND SURVEY

Unit 1: Sampling Survey and Sampling Distribution

Unit 2: Sampling Distribution of the Mean $(\bar{X})$

Unit 3: Sampling Distribution of Difference of Two Means and Sum $(\bar{x} - \bar{x})$

Unit 4: Sampling Distribution of Proportion ($\hat{P}$)

MODULE 2: ESTIMATION AND STATISTICAL TEST OF SIGNIFICANCE

Unit 5: Estimation: Point and Interval Estimation

Unit 6: Z-test and T-test

Unit 7: ANOVA/F-test

Unit 8: Chi-Square

MODULE 3: TEST OF HYPOTHESIS AND REGRESSION ANALYSIS

Unit 9: Hypothesis Testing

Unit 10: Simple Regression

Unit 11: Multiple Regression

Unit 12: Time Series

TEXTBOOKS AND REFERENCES

Every unit has a list of references and further reading. Endeavour to get sufficient amount of those textbooks and materials listed. The textbooks and materials are meant to broaden your knowledge of the course.

ASSIGNMENT FILE

There are assignments in this course and you are expected to do all of them by following the schedule set down for them in terms of when to attempt the homework and submit same for grading by your tutor.

PRESENTATION SCHEDULE

The presentation schedule included in your course materials offers you the significant dates for the completion of tutor-marked assignments and attending tutorials. Remember, you are obliged to submit all your assignments by the due date. You ought to safeguard against falling behind in your work.

ASSESSMENT

Your assessment will be based on Tutor-Marked Assignments (TMAs) and a final examination which you will write at the end of the course.

TUTOR-MARKED ASSIGNMENT

Assignment questions for the twelve units in this course are contained in the assignment file. You will be able to finish your assignments from the information and materials contained in your set books, reading and study units. Though, it is advantageous that you show that you have read and researched more extensively than the required minimum. You ought to use other references to have a wide-ranging perspective of the subject and also to give you an in-depth understanding of the subject.

When you have finished each assignment, send it, collectively with a TMA form, to your tutor. Ensure that each assignment gets to your tutor on or before the deadline set in the presentation schedule. If for any reason, you cannot finish your work on time, get in touch with your tutor before the assignment is due to talk about the likelihood of an extension. Extensions will not be granted after the due date except there are extraordinary conditions. The TMAs more often than not make up 30% of the entire score for the course.

FINAL EXAMINATION AND GRADING

The final examination for ECO726 will be two hours duration and have a value of 70% of the total course grade. The examination will be made of questions which reveal the kinds of self-assessment practice exercises and tutor-marked problems you have beforehand encountered. All areas of the course will be assessed

You ought to use the time between finishing the last unit and sitting for the examination to revise the whole course material. You may find it helpful to appraise 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.

Assessment	**Marks**
Assignment	
(Best three assignments out of the four marked)	30%
Final Examination	70%
Total	**100%**

COURSE OVERVIEW

The Table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, Applied Statistics (ECO726).

Units	**Title of Work**	**Week's Activities**	**Assessment (end of unit)**
	Course Guide		
MODULE 1: BASIC SAMPLING AND SURVEY			
1	Unit 1: Sampling Survey and Sampling Distribution	Week 1	Assignment 1
2	Unit 2: Sampling Distribution of the Mean $(\bar{X})$	Week 2	Assignment 1
3	Sampling Distribution of Difference of Two Means and Sum $(\bar{x} - \bar{x})$	Week 3	Assignment 1
4	Sampling Distribution of Proportion ($\hat{P}$)	Week 4	Assignment 1
MODULE 2: ESTIMATION AND STATISTICAL TEST OF SIGNIFICANCE			
1	Estimation: Point and Interval Estimation	Week 5	Assignment 2
2	Z-test and T-test	Week 6	Assignment 2
3	ANOVA/F-test	Week 7	Assignment 2
4	Chi-Square	Week 8	Assignment 2
MODULE 3: TEST OF HYPOTHESIS AND REGRESSION ANALYSIS			

1	Hypothesis Testing	Week 9	Assignment 3
2	Simple Regression	Week 10	Assignment 3
3	Multiple Regression	Week 11 & 12	Assignment 3
4	Time Series	Week 13	Assignment 3
	Examination	Week 14 & 15	
	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. Self-assessments are interspersed throughout 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 exercise 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 the help.

FACILITATORS/TUTORS AND TUTORIALS

There are some hours of tutorials (2-hour sessions) provided in support of this course. You will be notified of the dates, times and location of these tutorials, together with the name and phone number of your tutor, as soon as you are allocated a tutorial group.

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

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

Contact your tutor if:

• you do not understand any part of the study units or the assigned reading

• you have difficulty with the self-assessment exercises

• you have a question or problem with an assignment, with your tutor's comments on any 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

On successful completion of the course, you would have developed critical thinking and analytical skills (from the material) for efficient and effective application of statistics to solve economic problems. However, to gain a lot from the course please try to apply everything you learn in the course to term paper writing in other related courses. We wish you success with the course and hope that you will find it both interesting and useful.

TABLE OF CONTENT

MODULE 1: BASIC SAMPLING AND SURVEY

Unit 1: Sampling Survey and Sampling Distribution

Unit 2: Sampling Distribution of the Mean $(\bar{X})$

Unit 3: Sampling Distribution of Difference of Two Means and Sum $(\bar{x} - \bar{x})$

Unit 4: Sampling Distribution of Proportion ($\hat{P}$)

MODULE 2: ESTIMATION AND STATISTICAL TEST OF SIGNIFICANCE

Unit 5: Estimation: Point and Interval Estimation

Unit 6: Z-test and T-test

Unit 7: ANOVA/F-test

Unit 8: Chi-Square

MODULE 3: TEST OF HYPOTHESIS AND REGRESSION ANALYSIS

Unit 9: Hypothesis Testing

Unit 10: Simple Regression

Unit 11: Multiple Regression

Unit 12: Time Series

MODULE 1: BASIC SAMPLING AND SURVEY

Unit 1: Sampling Survey and Sampling Distribution

Unit 2: Sampling Distribution of the Mean $(\bar{X})$

Unit 3: Sampling Distribution of Difference of Two Means and Sum $(\bar{x} - \bar{x})$

Unit 4: Sampling Distribution of Proportion $(\hat{P})$

UNIT 1: SAMPLING, SURVEY AND SAMPLING DISTRIBUTION

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

- 3.1 Population
- 3.2 Sample
- 3.3 Sampling theory
  - 3.3.1 Sampling Techniques
  - 3.3.2 Non Probability Sample Design
  - 3.3.3 Probability Sample Design
- 3.4 Sampling Distribution of Parameter Estimate
- 3.5 Basic Descriptive Measures of Population and Sample

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/Further Reading

1.0 INTRODUCTION

In statistics, data are collected from a carefully selected sample from the population and are studied in order to learn something about the population which the data represents. In order to generate a meaningful statistics, Researchers generalize from what they find in the figure at hand (sample) to the wider phenomenon which they represent (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 as it

seems to compute since we may find it difficult to obtain values for every unit of the population. This can be made easy by a way of estimating samples (an abstract) of the population. This process is known as parameter estimation. Sampling theory deals with the study of the relationships that exist between a given population and the samples drawn from the population. From this theory, researchers can use a relatively small number of cases (a sample) as the bases for making inferences for all the cases (a population).

2.0 OBJECTIVES

At the end of this unit you should be able to understand and have knowledge of the following;

i. Population
ii. Sample
iii. Sampling theory
iv. Parameter estimation

3.0 MAIN CONTENT

3.1 Population

Statistical inference is the process by which conclusions are drawn on the basis of sample about the population from which sample is drawn. In order words, it is a process by which conclusion are drawn about some measure or attribute of a population based upon analysis of sample.

But before we talk on sampling we need to know about population. Methodologically, a population is the ―aggregate of all cases that conform to some designated set of specifications, it might be finite, when it consists of a given number of values or it may be infinite, when it includes an infinite number of values of the variable. In most cases values of population are hardly known, what we usually have is a certain number of values that any particular variable has assumed and which have been recorded in one way or the other. Such data form a sample from the population, for example, a population may be composed of all the residents in a specific country, neighbourhood, legislators, houses, records, and so on. The specific nature of the population depends on the research problem. If you are investigating the pattern of consumption in a particular city, you might define the population as all the households in that city. Therefore, one of the first problems facing a researcher who wishes to estimate a population value from a sample value is how to determine the population involved.

Sampling method refers to the way that observations are selected from a population to be in the sample for a sample survey. The reason for conducting a sample survey is to estimate the value of some attribute of a population.

Population parameter - A population parameter is the true value of a population attribute.

Sample statistics -A sample statistic is an estimate, based on sample data, of a population parameter.

3.2 Sample

A sample is thus a part of the population under study selected so that inferences can be drawn from it about the population. Sample can also be referred to a collection of observation on a certain variable. The number of observations included in the sample is called the sample size. When the data serves as the basis for inferences is comprised of a subset of the population, that subset is called a sample. It is cheaper and quicker to use samples to obtain information about a population than to take a census.

Sampling on the other hand is the process of selecting individuals (single items) from a population. An experiment which generates data for use by the statistician is often referred to as sampling, with the data so generated being called a sample (of data). The reason for sampling is that it would be impossible to study (at the very least too costly) all unemployed individuals, in order to explain the cause of unemployment variations or, or all members of the voting population in order to say something about the outcome of a general election. We therefore select a number of them in some way (not all of them), analyse the data on this sub-set, and then (hopefully) conclude something useful about the population of interest in general. The initial process of selection is called sampling and the conclusion drawn (about the general population from which the sample was drawn) constitutes statistical inference.

SELF ASSESSMENT EXERCISE

Differentiate between population and sample

3.3 Sampling Theory

Sampling theory is a study of relationships existing between a population and samples drawn from the population. Sampling theory is also useful in determining whether the observed differences between two samples are due to chance variation or whether they are really significant. The main objective of the theory is the development of method of drawing conclusion about the population (unknown) from the information provided by a sample.

In order to facilitate the study of population and sample, statisticians have introduced various descriptive measures that are various characteristics values that describe the important features of the sample or the population. The most important of these characteristics are the mean, variance and the standard deviation. To distinguish between sample and population, statisticians use the term parameter for the basic descriptive measure of population while statistics is usually used for the basic descriptive measure of a sample.

3.3.1 Sampling Techniques

In modern sampling theory, a basic distinction is made between probability and non-probability sampling. The distinguishing characteristic of probability sampling is that for each sampling unit of the population, you can specify the probability that the unit will be included in the sample. In the simplest case, all the units have the same probability of being included in the sample. In non probability sampling, there is no assurance that every unit has some chance of being included.

A well – designed sample ensures that if a study were to be repeated on a number of different samples drawn from a given population, the findings from each sample would not differ from the population parameters by more than a specified amount. A probability sample design makes it possible for researchers to estimate the extent to which the findings based on one sample are likely to differ from what they would have found by studying the entire population. When a researcher is using a probability sample design, it is possible for him or her to estimate the population‘s parameters on the basis of the sample statistics calculated.

3.3.2 Non-probability Sample Designs

With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen. Non-probability sampling methods offer two potential advantages - convenience and cost. The main disadvantage is that non-probability sampling methods do not allow you to estimate the extent to which sample statistics are likely to differ from the population parameters. Only probability sampling methods permit that kind of analysis. Four major designs utilizing non-probability samples have been employed by social scientists: convenience samples, purposive samples, voluntary samples and quota samples.

i. **Convenience sampling:** Researchers obtain a convenience sample by selecting whatever sampling units are conveniently available. Thus a University professor may select students in a class; or a researcher may take the first 200 people encountered on the street who are willing to be interviewed. The researcher has no way of estimating the representativeness of convenience sample, and therefore cannot estimate the population‘s parameters.
ii. **Purposive sampling:** With purposive samples (occasionally referred to as judgment samples), researchers select sampling units subjectively in an attempt to obtain a sample that appears to be representative of the population. In order words, the chance that a particular sampling unit will selected for the sample depends on the subjective judgment of the researcher. At times, the main reason for selecting a unit in purposive sampling is the possession of pre-determined characteristic(s) which may be

different from that the main population. For example, in a study of demand preference for cigarette brands in a city, researcher will need to select smokers purposively.

iii. **Voluntary sample.** A voluntary sample is made up of people who self-select into the survey. Often, these folks have a strong interest in the main topic of the survey. Suppose, for example, that a news show asks viewers to participate in an on-line poll. This would be a volunteer sample. The sample is chosen by the viewers, not by the survey administrator.

iv. **Quota sampling:** The chief aim of quota sample is to select a sample that is as similar as possible to the sampling population. For example, if it is known that the population has equal numbers of males and females, the researcher selects an equal numbers of males and females in the sample. In quota sampling, interviewers are assigned quota groups characterized by specific variables such as gender, age, place of residence, and ethnicity.

3.3.3 Probability Sample Designs

With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample. Four common designs of probability samples are simple random sampling, systematic sampling, stratified sampling, and cluster sampling. The key benefit of probability sampling methods is that they guarantee that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid.

i. **Simple random sampling:** is the basic probability sampling design, and it is incorporated into all the more elaborate probability sampling designs. Simple random sampling is a procedure that gives each of the total sampling units of the population an equal and known non zero probability of being selected. For example, when you toss a perfect coin, the probability that you will get a head or a tail is equal and known (50 percent), and each subsequent outcome is independent of the previous outcomes.

ii. **Systematic Sampling:** With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first k elements on the population list. Thereafter, we select every k^{th} element on the list.This method is different from simple random sampling since every possible sample of n elements is not equally likely.Systematic sampling is more convenient than simple random sampling. Systematic samples are also more amenable for use with very large populations or when large samples are to be selected.

iii. **Stratified Sampling**: Researchers use this method, primarily to ensure that different groups of population are adequately represented in the sample. This is to increase their level of accuracy when estimating parameters. Furthermore, all other things being equal, stratified sampling considerably reduces the cost of execution. The underlying idea in stratified sampling is to use available information on the

population —to divide it into groups such that the elements within each group are more alike than are the elements in the population as a whole. That is, you create a set of homogeneous samples based on the variables you are interested in studying. If a series of homogenous groups can be sampled in such a way when the samples are combined they constitute a sample of a more heterogeneous population, you will increase the accuracy of your parameter estimates.

iv. **Cluster sampling:** it is frequently used in large-scale studies because it is the least expensive sample design. Cluster sampling involves first selecting large groupings, called clusters, and then selecting the sampling units from the clusters. The clusters are selected by a simple random sample or a stratified sample. Depending on the research problem, researchers can include all the sampling units in these clusters in the sample or make a selection within the clusters using simple or stratified sampling procedures.

3.4 Sampling Distribution of Parameter Estimate

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.

The variance of a population is defined as the expected value of the squared deviations of the value of x from their expected mean value. This shows the various ways in which the various value of random variable x is distributed around their expected mean values. The smaller the variance, the closer and cluster of the values of x around the population mean.

The standard deviation of a population is defined as the square root of the population variance. The standard deviation is a measure that describes how dispersed the values of x is around the population mean.

Suppose we have a finite population and we draw all possible simple random samples of size *n* without replacement or with replacement. For each sample we calculate a statistic (sample mean $\bar{x}$ or proportion $\hat{p}$, etc.). All possible values of the statistic make a probability distribution which is called the sampling distribution. The number of all possible samples is usually very large and obviously the number of statistics

(any function of the sample) will be equal to the number of samples if one and only one statistic is calculated from each sample. In fact, in practical situations, the sampling distribution has a very large number of values. The shape of the sampling distribution depends upon the size of the sample, the nature of the population and the statistic which is calculated from all possible simple random samples. Some of the most well-known sampling distributions are:

(1) Binomial distribution

(2) Normal distribution

(3) t-distribution

(4) Chi-square distribution

3.5 Basic Descriptive Measure of Population and Sample

	Population parameters	**Symbol**	**Formula**	**Sample statistics**	**Symbol**	**Formula**
I.	Population mean	μ_x	$\frac{x_1 + x_2 + x_3 + \ldots x_N}{N} = \frac{\Sigma X}{N}$	Sample mean	$\bar{x}$	$\frac{x_1 + x_2 + x_3 + \ldots x_n}{n}$
II.	Population variance	$\sigma^2_{\bar{x}}$	$\frac{\Sigma(x-\mu)^2}{N}$ $= \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$	Sample variance	S_x^2	$\frac{\Sigma(x-\bar{x})^2}{n-1}$
III.	Population standard deviation	σ_x	$\sqrt{\frac{\Sigma(x-\mu)^2}{N}}$	Sample standard deviation	S_x	$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

Note: a sample is just an estimate of a large population. If we take another random sample and make the same calculation, the result will be different. As it turns out, dividing by n-1 instead of n gives you a better estimate of variance of the larger population. But if the sample is taken with replacement then variance and the standard deviation of the sample is same as that of the population.

4.0 CONCLUSION

The unit has exposed us to some definitions that with help in further studies. However, we have differentiated between population parameters and statistical parameters. Also, during our study we saw the different sampling techniques which help researchers in their surveys in order to draw inferences in their work.

5.0 SUMMARY

In this unit, we have discussed the definition of population, sample, sample distribution theory, so also estimation of parameter estimate and sample statistics had been attempted.

6.0 TUTOR-MARKED ASSIGNMENT

An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand. Is this an example of a simple random sample?

a) Yes, because each buyer in the sample was randomly sampled.
b) Yes, because each buyer in the sample had an equal chance of being sampled.
c) Yes, because car buyers of every brand were equally represented in the sample.
d) No, because every possible 400-buyer sample did not have an equal chance of being chosen.
e) No, because the population consisted of purchasers of four different brands of car.

7.0 REFERENCES/FURTHER READING

Adedayo, O. A. (2006). Understanding Statistics. Akoka, Yaba: JASPublishers.

Dominick, S. & Derrick P. (2011). Statistics and Econometrics. (2nd ed.) New York: Mcgraw Hill..

Edward, E. L.(1983). Methods of Statistical Analysis in Economics andBusiness. Boston: Houghton Mifflin Company.

Esan, F. O. &Okafor, R. O.(2010). Basis Statistical Method. Lagos:Toniichristo Concept.

Koutsoyianis, A. (2003). Theory of Econometrics. (2nd ed.). London:Palgrav Publishers Ltd. (formerly Macmillan Press Ltd).

Murray, R. S. & Larry, J. S. (1998). Statistics. (3rd ed.). New York:Mcgraw Hills.

Olufolabo, O. O. &Talabi, C. O. (2002). Principles and Practice ofStatistics. Shomolu,Lagos: HASFEM Nig Enterprises.

Oyesiku, O. K. &Omitogun, O. (1999). Statistics for Social andManagement Sciences. Lagos: Higher Education Books Publisher.

UNIT 2 & 3 SAMPLING DISTRIBUTION OF THE MEAN ($\bar{X}$) AND SAMPLING DISTRIBUTION OF DIFFERENCE OF TWO MEANS AND SUM ($\bar{X}$-$\bar{X}$)

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Sampling Distribution of the Mean ($\bar{X}$)
    3.2 Properties of Sample Distribution of the Mean ($\bar{X}$)
    3.3 Sampling Distribution of Difference of Two Means and Sum ($\bar{x}$-$\bar{x}$)
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

This unit is an extension of unit one of this module. In this unit we are looking at sampling distribution of sample mean. As it has been said before now that, the term statistics is usually used in describing the features of a sample. The basic statistic of a sample corresponding to the parameters of the population are sample mean usually denoted by $\bar{X}$, sample variance denoted by $S_x^2$and sample standard deviation denoted by S_x.Sample mean is defined as the average value in the sample it is denotedby $\bar{X}$. The sample arithmetic mean is calculated by adding up the observation of the sample and then dividing by the total number of observations.

2.0 OBJECTIVES

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

- Calculate sampling distribution of sample mean
- Know the properties of Sample distribution of $(\bar{X})$
- Explain the Sampling Distribution of Difference of Two Means and Sum ($\bar{x}$-$\bar{x}$)

3.0 MAIN CONTENT

3.1 Sampling Distribution of the Mean($\bar{X}$)

The probability distribution of all possible values calculated from all possible simple random samples is called the sampling distribution of $\bar{X}$. In brief, we shall call it the distribution of $\bar{X}$. The mean of this distribution is called the expected value and is written as $E(\bar{X})$ or $\mu_{\bar{x}}$. The standard deviation (standard error) of this distribution is denoted by S.E.$(\bar{X})$ or $\sigma_{\bar{x}}$ and the variance of is denoted by $Var(\bar{X})$ or $\sigma^2_{\bar{x}}$. The distribution of has some important properties:

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

$$E(\bar{X}) = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \ldots \bar{x}_n}{n} = \mu$$

Therefore this means that the expectation of the sample mean is the same as the population mean. The formula 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 $\sigma^2_{\bar{x}}$

$$= \frac{\sigma^2}{n} \ with\ replacement\ and\ \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\ OR\ \left(\frac{\sigma}{\sqrt{n}}\right)^2\left(\frac{N-n}{N-1}\right)\ without\ replacement.$$

$$If\ the\ population\ variance\ is\ not\ known, we\ make\ use\ of\ the\ estimate = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

Given that X_1, X_2, X_3 … X_n is a random sample of size n from normal population with mean$\bar{x}$ and variance σ^2 $i.e.(X \sim N\ (\mu, \sigma^2)$. Therefore $Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N\ (0,1)$

This is a general case whereby sampling is specifically taken from a normal distribution

Illustration:

Draw all possible samples of size 2 without replacement from a population consisting of 3, 6, 9, 12, 15. Form the sampling distribution of sample means and verify the results.

i. $E(\bar{X})=\mu$

ii. $Var\ (\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

Solution:

We have population values 3, 6, 9, 12, 15, population size $N=5$ and sample size $n=2$. Thus, the number of possible samples which can be drawn without replacement is 10 using the combination formula: ${}_n^N C = \frac{N!}{n!\times(N-n)!}$. Alternatively we can

Sample No.	Sample Values	Sample Mean $\bar{X}$
1	3,6	4.5
2	3,9	6
3	3,12	7.5
4	3, 15	9
5	6,9	7.5
6	6,12	9
7	6,15	10.5
8	9,12	10.5
9	9,15	12
10	12,15	13.5

The sampling distribution of the sample mean $\bar{X}$ and its mean and standard deviation are

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$

4.5	1	$\frac{1}{10}$	$\frac{4.5}{10}$	$\frac{20.25}{10}$
6	1	$\frac{1}{10}$	$\frac{6}{10}$	$\frac{36}{10}$
7.5	2	$\frac{2}{10}$	$\frac{15}{10}$	$\frac{112.5}{10}$
9	2	$\frac{2}{10}$	$\frac{18}{10}$	$\frac{162}{10}$
10.5	2	$\frac{2}{10}$	$\frac{21}{10}$	$\frac{220.5}{10}$
12	1	$\frac{1}{10}$	$\frac{12}{10}$	$\frac{144}{10}$
13.5	1	$\frac{1}{10}$	$\frac{13.5}{10}$	$\frac{182.25}{10}$
Total	10	1	$\frac{90}{10}$	$\frac{877.5}{10}$

$$E(\bar{X}) = \Sigma \bar{X} f(\bar{X}) = \frac{90}{10} = 9$$

$$Var\ (\bar{X}) = \Sigma \bar{X}^2 f(\bar{X}) - [\Sigma \bar{X} f(\bar{X})]^2 = \frac{877.5}{10} - \left[\frac{90}{10}\right]^2 = 87.75 - 81 = 6.75$$

The mean and variance of the population are:

X	3	6	9	12	15	$\Sigma X = 45$
X^2	9	36	81	144	225	$\Sigma X^2 = 495$

$$\mu = \frac{\Sigma X}{N} == \frac{45}{5} = 9 \sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{495}{5} - \left(\frac{45}{5}\right)^2 = 99 - 81 = 18$$

Verification……

i. $E(\bar{X}) = \mu = 9$

ii. $Var\ (\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{18}{2}\left(\frac{5-2}{5-1}\right) = 6.75$

3.2 Properties of Sample Distribution of $(\bar{X})$

- One important property of the distribution of $(\bar{X})$ is that it is a normal distribution when the size of the sample is large. When the sample size *n* is more than *30*, we call it a large sample size. The shape of the population distribution does not matter. The population may be normal or non-normal, the distribution of $(\bar{X})$ is normal for $n>30$, but this is true when the number of samples is very large. As the distribution of random variable $(\bar{X})$ is normal, $(\bar{X})$ can be transformed into a standard normal variable *Z*, where $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. The distribution of $(\bar{X})$ has a t-distribution when the population is normal and $n \leq 30$.
- The mean of the distribution is equal to the mean of the population. This relation is true for small as well as large sample sizes in sampling without replacement and with replacement.
- The standard error (standard deviation) is related to the standard deviation of the population through the relations:

  i. $S.E(\bar{X}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ This is true when population is infinite, which means *N* is very large or the sampling is done with replacement from a finite or infinite population.

ii. $S.E(\bar{X}) = \sigma_{\bar{x}} = \frac{\sigma^2}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$This is true when sampling is without replacement from a finite population. The above two equations between and are true both for small as well as large sample sizes.

SELF ASSESSMENT EXERCISE

List and explain the properties of sample distribution $(\bar{X})$

3.3 Sampling Distribution of Difference of Two Means and Sum $(\bar{x} - \bar{x})$

If two independent random samples of sizes n_1 and n_2 are selected from 2 different population of size N_1 and N_2 with population means μ_1 and $\mu_1$respectively and population variance σ_1^2 and σ_2^2 respectively, then the sampling distribution of the difference of two means $(\bar{x_1} - \bar{x_2})$ = $\mu_{p1} + \mu_{p2}$ and standard deviation of the sample distribution is written as:

$$\sigma_{x_1 - x_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} - \frac{{\sigma_2}^2}{n_2}}$$

Also the sampling distribution of sum of means is as defined below:$\mu_{p1+p2} = \mu_{p1} + \mu_{p2}$and the standard deviation ${\sigma^2}_{p_1 - p_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} - \frac{{\sigma_2}^2}{n_2}}$

Illustration

Given that p_1 = (30, 50) and p_2 = (40, 70) for a sample drawn from each other show that

1. $\mu_{p1+p2} = \mu_{p1} + \mu_{p2}$
2. $\mu_{p1-p2} = \mu_{p1} - \mu_{p2}$and
3. ${\sigma^2}_{p1+p2} = {\sigma^2}_{p1} + {\sigma^2}_{p2}$

Solution

Sampling sum Possible sample combination = (30, 40), (30, 70) (50, 40) (50, 70)

1. Sample sum = 30 + 40 = 70; 30 + 70 = 100; 50 + 40 = 90; 50 + 70 = 120

$$\therefore \mu_{p1+p2} = \frac{70 + 100 + 90 + 70 + 120}{4} = 95$$

Now considering the 1st population p_1 (30, 50) $\mu_{p1} = \frac{30+50}{2} = 40$

Considering the 2nd population p_2 (40, 70) $\mu_{p2} = \frac{40+70}{2} = 55$

$\mu_{p1} + \mu_{p2} = 55 + 40 = 95$

$\therefore \mu_{p1+p2} = \mu_{p1} + \mu_{p2} = 95$

2. Sample difference = $30 - 40 = -10;\ 30 - 70 = -40;\ 50 - 40 = 10;\ 50 - 70 = -20$

$\therefore \mu_{p1-p2} = \frac{-10 - 40 + 10 - 20}{4} = -15$

Now considering the 1st population p_1 (30, 50) $\mu_{p1} = \frac{30+50}{2} = 40$

Considering the 2nd population p_2 (40, 70) $\mu_{p2} = \frac{40+70}{2} = 55$

$\mu_{p1} - \mu_{p2} = 40 - 55 = -15$

$\therefore \mu_{p1-p2} = \mu_{p1} - \mu_{p2} = -15$

3. σ^2_{p1+p2} = variance of 70, 10, 90 & 120.

Note that the population mean = 95

$$\sigma^2_{p1+p2} = \frac{\sum(x - \bar{x})^2}{n}$$

$$\sigma^2_{p1+p2} = \frac{(70-95)^2 + (100-95)^2 + (90-95)^2 + (120-95)^2}{4}$$

$$\sigma^2_{p1+p2} = \frac{-25^2 + -5^2 + 5^2 + 25^2}{4} = \frac{625 + 25 + 25 + 625}{4} = \frac{1300}{4} = 325$$

Let's consider the population independently

σ^2_{p1} = variance of (30, 50), $\mu_{p1} = 40$

$$\sigma^2_{p1} = \frac{(30-40)^2 + (50-40)^2}{2} = \frac{100 + 100}{2} = 100$$

σ^2_{p2} = variance of (40, 70), $\mu_{p2} = 55$

$$\sigma^2_{p1} = \frac{(40-55)^2 + (70-55)^2}{2} = \frac{225 + 225}{2} = 223$$

$\therefore \sigma^2_{p1+p2} = \sigma^2_{p1} + \sigma^2_{p2} = 325$

4.0 CONCLUSION

In this unit, it has been established that given a random sample of X_1, X_2, ….X_n with population mean μ and standard variance. We can verify that

i. $E(\bar{X})=\mu$

ii. $Var\ (\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

5.0 SUMMARY

In this unit, we have discussed the definition of sample distribution of sample mean, so also it has been proved from our calculation that the mean of sample must always equal to the population mean it's representing and that the variance of the population and sample estimate are equal.

6.0 TUTOR-MARKED ASSIGNMENT

1. If random samples of size 2 are drawn without replacement from the population consisting of four numbers 4, 5, 5, 7. Find the sample mean$(\bar{X})$ for each sample and make a sampling distribution of $(\bar{X})$. Calculate the mean and standard deviation of this sampling distribution. Compare your calculations with the population parameters.
2. 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.
3. Given the following population p1 = (10, 20) p2 = (30,40) show that
   i. $\mu_{p1+p2} = \mu_{p1} + \mu_{p2}$
   ii. $\mu_{p1-p2} = \mu_{p1} - \mu_{p2}$and
   iii. $\sigma^2_{p1+p2} = \sigma^2_{p1} + \sigma^2_{p2}$

7.0 REFERENCES/FURTHER READING

Adedayo, O. A. (2006). Understanding Statistics. Akoka, Yaba: JAS Publishers.

Dominick, S. & Derrick P. (2011). Statistics and Econometrics. (2nd ed.) New York: Mcgraw Hill..

Edward, E. L.(1983). Methods of Statistical Analysis in Economics and Business. Boston: Houghton Mifflin Company.

Esan, F. O. &Okafor, R. O.(2010). Basis Statistical Method. Lagos: Toniichristo Concept.

Koutsoyianis, A. (2003). Theory of Econometrics. (2nd ed.). London: Palgrav Publishers Ltd. (formerly Macmillan Press Ltd).

Murray, R. S. & Larry, J. S. (1998). Statistics. (3rd ed.). New York: Mcgraw Hills.

Olufolabo, O. O. &Talabi, C. O. (2002). Principles and Practice of Statistics. Shomolu,Lagos: HASFEM Nig Enterprises.

Oyesiku, O. K. &Omitogun, O. (1999). Statistics for Social and Management Sciences. Lagos: Higher Education Books Publisher.

UNIT 4 SAMPLING DISTRIBUTION OF PROPORTION

CONTENTS

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

1.0 INTRODUCTION

This unit is an extension of unit one of this module. In this unit we are going to look at sampling distribution of proportion, sampling distribution of sum and difference and standard error. Since this unit is an offshoot of the unit one of this module, most of the statistical term used in unit one will be implied here.

2.0 OBJECTIVE

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

- Sampling distribution of proportion
- Sampling distribution of sum
- Sampling distribution of difference and
- Standard error

3.0 MAIN CONTENT

3.1 Sampling Distribution of Proportion Defined

To this point, the discussion in this section has been solely concerned with the behaviour of the mean in random sampling. Another population parameter with which researchers are commonly concerned is the population proportion. Suppose a particular characteristic of a population is being investigated, and the proportion of members of the population with this characteristic is to be determined. Ordinarily the proportion of cases in the sample which take on this characteristic will be used to make statements concerning the population proportion.

If the sample is a random sample, and is reasonably large, then the sampling distribution of the sample proportion can be determined. This is based on the binomial distribution, and is a simple extension of the normal approximation to the binomial probability distribution.

Let p be the proportion of members of the population having the characteristic being investigated. This can be stated in terms of the binomial by defining this characteristic as a success. Any member of the population which does not have this characteristic can be considered as a failure.

Samples are usually embedded in a population, each time attribute is sampled, the concept of proportion is coming in the estimation here is concentrating on the proportion of the population that has a peculiar characteristics. This sampling distribution is like that of binomial distribution, where an event is divided into been a success represented with p or been a failure represented with *q* or *1 – p*.

Given an infinite population consisting of sample size *n*, the sampling distribution of proportion is said to have a mean of *np* and variance $Var(\hat{p}) = \frac{p(1-p)}{n} = \frac{pq}{n}$

It is to be noted at this juncture that the sample proportion is also an unbiased estimator of the population proportion i.e. $\Sigma(\hat{p})$ = P

p = the proportion of the population having some characteristic. Sample proportion ($\hat{p}$) provides an estimate of p. $0 \leq \hat{p} \leq 1$ and has a binomial distribution, but can be approximated by a normal distribution when *n* is large. Standardize to a Z value with the formula $Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

Illustration 1

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

Solution:

The 150 tosses is a sample from infinite population of all possible tosses of the coin. However, since the coin is fair $From\ the\ above\ the\ prob(head) = ½ = p$

$Prob(not\ obtaining\ ahead) = ½ = q = 1 - p$

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

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

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

$$\therefore P(57 \leq \hat{p} \leq 87) = P\left(\frac{57-75}{6.12} \leq Z \leq \frac{87-75}{6.12}\right) = P(-2.94 \leq Z \leq 1.96) = 0.5596 - 0.0053 = 0.5543.$$

SELF ASSESSMENT EXERCISE

The sampling distribution of proportion is important in statistics. Discuss

3.2 Standard Error

Standard error usually represented by S.E. is defined as the square root of the population variance written as

$\sqrt{var\ p}$ where $Var(p) = \frac{pq}{n} = \frac{p(1-p)}{n}$ $\therefore S.E = \sqrt{\frac{p(1-p)}{n}}$

From our illustration above $S.E = \sqrt{\frac{0.5(1-0.5)}{150}} = 0.04$

4.0 CONCLUSION

During the course of our discussion of this unit we have talked about;

- Sampling distribution of proportion
- Standard error

5.0 SUMMARY

In the course of our discussion we defined the mean of a sampling distribution of proportion as *np*. i.e. mean = *np*, $variance\ (p) = \frac{p(1-p)}{n}$, $\sigma(p) = \sqrt{npq}$ and $S.E = \sqrt{\frac{p(1-p)}{n}}$

6.0 TUTOR-MARKED ASSIGNMENT

If the true proportion of voters who support Proposition A is p = 0.4, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?

7.0 REFERENCES/FURTHER READING

Adedayo, O. A. (2006). Understanding Statistics. Yaba, Lagos: JAS Publishers.

Esan, F. O. &Okafor, R. O. (2010). Basis Statistical Method (Revised edition). Lagos: Toniichristo Concept.

Murray, R.S. & Larry, J. S. (1998). (Schaum Outlines Series). Statistics. (3rd ed.). New York: Mcgraw Hills.

Olufolabo, O.O. &Talabi, C. O.(2002). Principles and Practice of Statistics. Shomolu Lagos: HASFEM Nig Enterprises.

Oyesiku, O. K. &Omitogun, O. (1999). Statistics for Social and Management Sciences. Lagos: Higher Education Books Publisher.

MODULE 2 ESTIMATION AND STATISTICAL TEST OF SIGNIFICANCE

Unit 1: Estimation: Point and Interval Estimation

Unit 2: Z-test and T-test

Unit 3: ANOVA/F-test

Unit 4: Chi-Square

UNIT 1: ESTIMATION: POINT AND INTERVAL ESTIMATION

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Estimation
    3.2 Point Estimation
    3.3 Interval Estimation
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

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

2.0 OBJECTIVES

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

- know the difference between point estimation and interval estimation
- know the properties of a good estimator
- know how to calculate the point and interval estimation.

3.0 MAIN CONTENT

3.1 Estimation

From our previous knowledge we are now able to distinguish between a population and a sample, a parameter and a statistic, and random and non-random sampling. We also have a little knowledge of some sampling distributions. The difficulty in dealing with populations is that we usually do not know the value of their parameters. Even if we could find these values, it is not practical in terms of time, money and reliability.

This section is devoted to the study of the most suitable values of these parameters based on the random samples from the given population. These values, which are functions of sample observations and derived on the basis of a certain criterion, are called the "estimates". The formulae with which these estimates are obtained are called "estimators". The method by which we extract information about the population on the basis of samples is called "estimation".

What makes estimation challenging is the problem of determining the most probable values of the parameters of probability distributions. For example, estimate the average height of individuals in a

- **a.** The average height of the individuals is 65 inches, or
- **b.** It is most probable that the average height is between 62 and 67 inches

We therefore have two kinds of estimates: An estimate of the type given in (a) above which is known as a "Point Estimate", and an estimate of the type given in (b) above which is known as an "Interval Estimate". Point estimation is the opposite of interval estimation. It produces a single value while the latter produces a range of values. A point estimator is a statistic utilized to estimate the value of an unknown parameter of a population. It uses sample data when calculating a single statistic that will be the best estimate of the unknown parameter of the population.

On the other hand, interval estimation uses sample data to calculate the interval of the possible values of an unknown parameter of a population.

The interval of the parameter is selected in a way that it falls within a 95% or higher probability, also known as the confidence interval. The confidence interval is used to indicate how reliable an estimate is, and it is calculated from the observed data. The endpoints of the intervals are referred to as the upper and lower confidence limits.

3.2 Point Estimation

In the point estimation procedure we make an attempt to compute a numerical value from sample observations, which could be taken as an approximation to the parameter. The estimators, which are also referred to as

statistics (plural of statistic), since they are based on observations which are random variables themselves. A number of estimation methods like method of least square, method of maximum likelihood, method of moments, etc., are available with some specific properties.

- **Method of Least Square**

The method of least square is specifically used in regression analysis to estimate the regression coefficients. To understand the technique of estimation let us consider the following simple example. Please note that a formal treatment of the least square method which involves the inclusion of the disturbance term has been avoided for simplicity.

Suppose that consumption expenditure Y is linearly related to only one variable, family income X

This can be written mathematically as $Y = a + bX$

In economics, this relation is known as a consumption function, where a is a measure of the consumption expenditure at zero level of income and is a measure of the marginal propensity to consume, i.e., it gives a measure of how much will be consumed from each additional unit of income. The consumption function is in the parametric form, specifying a different relationship for different values of the parameters (a and b**).** The parameters (a , b) are not known and need to be estimated on the basis of a sample. A random sample of n households is drawn from the population under study. The information about consumption and income is recorded as follows for each of these households.

Consumption Expenditure	**Family Income**
Y_1	X_1
Y_1	X_2
. . .	. . .
Y_n	X_n

On the basis of these sample observations we wish to estimate the consumption function. Let the estimating equation be $\hat{Y} = \hat{a} + \hat{b}X$, where $\hat{Y}$, $\hat{a}$ and $\hat{b}$ are the estimates of Y, a and b respectively.

Since $\hat{Y}$ is an estimate of Y, it will be very lucky on our part to have $\hat{Y}$ equal to Y; otherwise they will be different. The difference between an estimate value $\hat{Y}$ and the observed value Y is denoted by e, which is usually termed “residual”, “deviation” or “error term”. This residual may be positive or negative.

$$e = Y - \hat{Y}$$

$$e = Y - \hat{a} - \hat{b}X$$

The smaller the residuals are, the closer the estimating equation be $\hat{Y} = \hat{a} + \hat{b}X$ is to the original model $Y = a + bX$. Hence, to have a closer estimating equation for $Y = a + bX$ we should minimize the residuals. The residuals are minimized according to the following principle, which states that:

Note: *Those values of $\hat{a}$ and $\hat{b}$ should be chosen which minimize the sum of squared residual". This principle is known as the "principle of least squares.*

Thus, the sum of the squared residual may be written as $\sum e^2 = \sum(Y - \hat{a} - \hat{b}X)^2$

In order to minimize the quantity $\sum e^2$, we will use the technique of differential calculus. Hence, differentiating $\sum e^2 = \sum(Y - \hat{a} - \hat{b}X)^2$ with respect to $\hat{a}$ and $\hat{b}$ equating the resulting derivatives to zero.

$$\frac{\partial \sum e^2}{\partial \hat{a}} = -2\sum(Y - \hat{a} - \hat{b}X) = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

$$\frac{\partial \sum e^2}{\partial \hat{b}} = -2\sum X(Y - \hat{a} - \hat{b}X) = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

Simplifying the above equations, we have

$$\sum Y = n\hat{a} + \hat{b}\sum X \ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

$$\sum XY = \hat{a}\sum X + \hat{b}\sum X^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots 4$$

These two equations are called the "Normal Equation" in which if we substitute the values $\sum XY$, $\sum Y, \sum X$ and $\sum X^2$ and from our sample observations, the two estimates and of the unknown parameters $\hat{a}$ and $\hat{b}$ can be determined by solving the simultaneous equations.

Our estimates:

$\hat{a} = \bar{Y} - \hat{b}\bar{X}$ by dividing thruogh by n in eq. 3

$\sum XY - (\bar{Y} - \hat{b}\bar{X})\sum X = \hat{b}\sum X^2$ substituting for $\hat{a}$ in eq. 4

$$\sum XY - \bar{Y}\sum X = \hat{b}\left(\sum X^2 - \bar{X}\sum X\right)$$

Recall $\sum X = N\bar{X}$

$$\sum XY - N\overline{XY} = \hat{b}\left(\sum X^2 - N\overline{X}^2\right)$$

$$\hat{b}\sum (X-\overline{X})^2 = \sum (X-\overline{X})(Y-\overline{Y})$$

$$\hat{b} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{Cov(x,y)}{Var(x)}$$

- **Maximum Likelihood Estimation**

The maximum likelihood estimator method of point estimation attempts to find the unknown parameters that maximize the likelihood function. It takes a known model and uses the values to compare data sets and find the most suitable match for the data. For example, a researcher may be interested in knowing the average weight of babies born prematurely. Since it would be impossible to measure all babies born prematurely in the population, the researcher can take a sample from one location. Since the weight of pre-term babies follows a normal distribution, the researcher can use the maximum likelihood estimator to find the average weight of the entire population of pre-term babies based on the sample data.

The maximum likelihood estimation method is a very rigorous statistical method of estimation. The word likelihood has the same meaning as the word probability. The method of maximum likelihood is not restricted to a specific type of analysis like the least square method; rather its application is universal provided the probability distribution of the population is known. Using this method we are able to obtain an estimate of the parameter which is most likely to be true (i.e., it has the maximum probability to be true). The method of determining maximum likelihood estimates is briefly outlined in the following steps.

Formulate the likelihood function (L). The likelihood function is the joint probability distribution of a sample of n values of random variables.

If the likelihood function (L) is in exponential form, it will be much more convenient if we write it in the logarithmic form, i.e., find $L_n L$

Maximize L or$L_n L$ with respect to the parameter whose estimate(s) is (are) desired using the technique of differential calculus.

Illustration

Find the maximum likelihood estimate of the parameter in a normal population assuming it is known.

***Solution*:**

Let the random sample $X_1, X_2, X_3, \ldots X_n$ be drawn from a normal population, and each of the X_i will be normally distributed, i.e.

$$P(X_1) = \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_1 - \mu)^2\right]$$

$$P(X_2) = \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_2 - \mu)^2\right]$$

$$P(X_3) = \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_3 - \mu)^2\right]$$

… … … … … …

$$P(X_n) = \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_n - \mu)^2\right]$$

The likelihood function will be the joint distribution (product) of all these density functions. Therefore,

$$L = \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_1 - \mu)^2\right]$$

$$\times \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_2 - \mu)^2\right]$$

$$\times \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_3 - \mu)^2\right]$$

… … … … … …

$$\times = \frac{1}{\sqrt{2\pi\sigma}^2} exp\left[-\frac{1}{2\sigma^2}(X_n - \mu)^2\right]$$

$$L = \frac{1}{\sqrt{2\pi\sigma^2}^n} exp\left[-\frac{1}{2\sigma^2}(X_i - \mu)^2\right]$$

To simplify the process of maximization, take the logarithm of both sides. Therefore,

$$L_n L = \frac{n}{2} L_n(2\pi) - \frac{n}{2} L_n(\sigma^2) - \frac{1}{2\sigma^2} \sum (X - \mu)^2$$

Now the necessary condition for maximization is that the first derivative with respect to μ sshould be zero. Therefore,

$$\frac{\partial L_n L}{\partial \mu} = 0 - 0\frac{1}{2\sigma^2}(-2)\sum(X-\mu) = \mathbf{0}$$

$$or \ \sum(X-\mu) = 0$$

$$or \quad \sum X - \sum \mu = 0$$

$$or \quad \sum X - n\mu = 0$$

Therefore,

$$n\mu = \sum X$$

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

Hence $\bar{X}$ (the sample mean) is an estimator of the population mean

- **Method of Moments**

The method of moments of estimating parameters was introduced in 1887 by Russian mathematician Pafnuty Chebyshev. It starts by taking known facts about a population and then applying the facts to a sample of the population. The first step is to derive equations that relate the population moments to the unknown parameters.

The next step is to draw a sample of the population to be used to estimate the population moments. The equations derived in step one are then solved using the sample mean of the population moments. It produces the best estimate of the unknown population parameters.

In short, the method of moments involves equating sample moments with theoretical moments. So, let's start by making sure we recall the definitions of theoretical moments, as well as learn the definitions of sample moments.

Definitions

1. $E(X^k) is\ the$ kth (Theoritical) moment of the distribution about a origin, $for\ k = 1,2, \dots, n$
2. $E[(X-\mu)^k]$ is the kth (theoretical) moment of the distribution about the mean, for k = 1, 2, ... n
3. $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ is the k^{th} sample moment, for $k = 1,2, \dots$ n

4. $M_k^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ is the k^{th} sample moment about the mean, for $k = 1, 2, \ldots$ n

One Form of the Method

The basic idea behind this form of the method is to:

a) Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ to the first theoretical moment $E(X)$.

b) Equate the second sample moment about the origin $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ to the second theoretical moment $E(X^2)$.

c) Continue equating sample moments about the origin, M_k, with the corresponding theoretical moments $E(X^k), k = 3, 4, \ldots$ until you have as many equations as you have parameters.

d) Solve for the parameters.

The resulting values are called method of moments estimators. It seems reasonable that this method would provide good estimates, since the empirical distribution converges in some sense to the probability distribution.

Therefore, the corresponding moments should be about equal.

Illustration 1:

Let $X1, X2, \ldots, Xn$ be Bernoulli random variables with parameter *p*. What is the method of moments estimator of *p*?

Solution

Here, the first theoretical moment about the origin is:

$$E(X_i) = p$$

We have just one parameter for which we are trying to derive the method of moments estimator. Therefore, we need just one equation. Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

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

Now, we just have to solve for *p*. In this case, the equation is already solved for *p*. We just need to put a hat (^) on the parameter to make it clear that it is an estimator. We can also subscript the estimator with an "MM" to indicate that the estimator is the method of moments estimator:

$$\bar{p}_{mm} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

So, in this case, the method of moments estimator is the same as the maximum likelihood estimator, namely, the sample proportion.

Illustration 2:

Let $X1, X2, \ldots, Xn$ be normal random variables with mean μ and variance σ^2. What are the method of moments estimators of the mean μ and variance σ^2?

Solution

The first and second theoretical moments about the origin are:

$$E(X_i) = \mu \text{ and } E(X_i^2) = \sigma^2 + \mu^2$$

(Incidentally, in case it's not obvious, that second moment can be derived from manipulating the shortcut formula for the variance). In this case, we have two parameters for which we are trying to derive method of moments estimators. Therefore, we need two equations here. Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

And, equating the second theoretical moment about the origin with the corresponding sample moment, we get:

$$E(X^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Now, the first equation tells us that the method of moments estimator for the mean μ is the sample mean:

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

And, substituting the sample mean in for μ in the second equation and solving for σ^2, we get that the method of moments estimator for the variance σ^2 is:

$$\bar{\sigma}^2_{mm} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

This can be rewritten as:

$$\bar{\sigma}^2_{mm} = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Again, for this example, the method of moments estimators are the same as the maximum likelihood estimators.

In some cases, rather than using the sample moments about the origin, it is easier to use the sample moments about the mean. Doing so provides us with an alternative form of the method of moments.

The basic idea behind this form of the method is to:

1. Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ to the first theoretical moment $E(X)$.
2. Equate the second sample moment about the mean $M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to the second theoretical moment about the mean $E[(X - \bar{X})^2]$.
3. Continue equating sample moments about the mean M_k^* with the corresponding theoretical moments about the mean $[(X - \mu)^k]$, *k* = 3, 4, ... until you have as many equations as you have parameters.
4. Solve for the parameters.

Illustration 3

Let X_1, X_2, …X_n be gamma random variables with parameters α and θ, so that the probability density function is:

$$f(x_i) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}$$

For $x > 0$. Therefore, the likelihood function:

$$L(\alpha, \theta) = \left(\frac{1}{\Gamma(\alpha)\theta^{\alpha}}\right)^n (x_1 x_2 \ldots x_n)^{\alpha-1} \exp\left[-\frac{1}{\theta}\sum x_i\right]$$

It is difficult to differentiate because of the gamma function Γ(α). So, rather than finding the maximum likelihood estimators, what are the method of moments estimators of α and θ?

Solution.

The first theoretical moment about the origin is: $E(X_i) = \alpha\theta$ And the second theoretical moment about the mean is: $Var(X_i) = E(X_i - \mu)^2 = \alpha\theta^2$

Again, since we have two parameters for which we are trying to derive method of moments estimators, we need two equations. Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

$$E(X) = \alpha\theta = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

And, equating the second theoretical moment about the mean with the corresponding sample moment, we get:

$$Var(X) = \alpha\theta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now, we just have to solve for the two parameters α and θ. Let's start by solving for α in the first equation $(E(X))$. Doing so, we get:

$$\alpha = \frac{\bar{X}}{\theta}$$

Now, substituting $\alpha = \frac{\bar{X}}{\theta}$ into the second equation $[Var(X)]$ we get:

$$\alpha\theta^2 = \left(\frac{\bar{X}}{\theta}\right)\theta^2 = \bar{X}\theta = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now, solving for θ in that last equation, and putting on its hat, we get that the method of moment estimator for θ is:

$$\hat{\theta}_{mm} = \frac{1}{n\bar{X}}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

And, substituting that value of θ back into the equation we have for α, and putting on its hat, we get that the method of moment estimator for α is:

$$\hat{\alpha}_{mm} = \frac{\bar{X}}{\hat{\theta}_{mm}} = \frac{\bar{X}}{\frac{1}{n\bar{X}}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

SELF ASSESSMENT EXERCISE

Explain the maximum likelihood estimation

3.2 Properties of Good Point Estimators

There are essentially three criteria which we use to select good estimators. The problem that arises, of course, is that a particular estimator may be better than another under one criterion but worse than that other estimator under another criterion.

1. **Unbiasedness:** An estimator is unbiased if the mean of its sampling distribution is equal to the population characteristic to be estimated. That is, S is an unbiased estimator of θ if $E(S) = \theta$. If the estimate is biased, the bias equals $B = E(S) - \theta$. The median, for example, is a biased estimator of the population mean when the probability distribution of the population being sampled is skewed. The estimator

   $$\hat{s}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

   turns out to be a biased estimator of σ^2 while the estimator

   $$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

   is unbiased. This explains why we have been using s^2 rather than $\hat{s}^2$. Unbiasedness in point estimators refers to the tendency of sampling errors to balance out over all possible samples. For any one sample, the sample estimate will almost surely differ from the population parameter. An estimator may still be desirable even if it is biased when the bias is not large because it may have other desirable properties.
2. **Consistency:** An estimator is a consistent estimator of a population characteristic θ if the larger the sample size the more likely it is that the estimate will be close to θ. For example in the shoe-pair testing example above, $\bar{X}$ is a consistent estimator of μ because its sampling distribution tightens around μ = .2 as n increases. More formally, S is a consistent estimator of population characteristic θ if for any small positive value ϵ, $\lim_{n\to\infty}(P(|S - \theta| < \epsilon) = 1.)$
3. **Efficiency:** The efficiency of an unbiased estimator is measured by the variance of its sampling distribution. If two estimators based on the same sample size are both unbiased, the one with the smaller variance is said to have greater relative efficiency than the other. Thus, S_1 is relatively more efficient than S_2 in estimating θ if $\sigma^2(S_1) < \sigma^2(S_2)$ and $E(S_1) = E(S_2) = \theta$ For example, the sample mean and sample median are both unbiased estimators of the mean of a normally distributed population but the mean is a relatively more efficient estimator because at any given sample size its variance is smaller.

3.3 Interval Estimation

From the discussion on point estimation we know that $\bar{X}$ is the best possible estimator of the population mean μ, which is a fixed usually unknown parameter. It would be extremely lucky to have a sample which has a mean $\bar{X}$ exactly equal to the population mean μ, so in most cases it will be a little higher or a little lower.

A point estimate of any parameter might be misleading. We, therefore, try to determine two values within which the true value of the parameter is expected to fall instead of one point estimate. We can also attach a certain degree of true values of a parameter which are called confidence limits (lower and upper confidence limits), and the two together is called a confidence interval. The word ‘confidence” refers to probability. We can determine

an interval estimate of any parameter to any degree of confidence but usually it is estimated with 90, 95 or 99 percent confidence Thus, if we determine a 95 percent confidence interval estimate, we understand that the probability that the interval contains the true parameter is 0.95, or in other words, out of 100 possible intervals 95 of the intervals are certain to contain the true parameter.

Remembering that is a standard normal variate consulting the normal table.

$$P\,(-1.80 < Z < 2.19) = 0.95$$

$$P\,(-2.36 < Z < 1.74) = 0.95$$

We can in fact construct thousands of such intervals each having a probability equal to 0.95. Let us now look at the range (length) of each of these intervals:

$$R1 = 1.96 - (-1.96) = 3.92$$

$$R2 = 2.19 - (-1.80) = 3.99$$

In general we can write $P\,(a < Z < b) = 0.95$ with a range equal to$(b - a)$. To get an interval as precise as possible one would prefer an interval with the lowest value of$(b - a)$. Hence, in the above mentioned four statements, the first interval is more precise than the other three. In the forthcoming discussion we will try to construct confidence intervals with the least possible length.

To provide an indication of the precision of a point estimate we combine it with an interval estimate. An interval estimate of the population mean μ would consist of two bounds within which μ is estimated to lie:

$$L \leq \mu \leq U$$

where L is the lower bound and U is the upper bound. This interval gives an indication of the degree of precision of the estimation process. To obtain an estimate of how far the sample mean is likely to deviate from the population mean — i.e., how tightly it is distributed around the population mean — we use our estimate of the variance of the sample mean

$$S_{\bar{x}}^2 = \frac{S^2}{n}$$

This enables us to say that if the sample is large enough, $\bar{X}$ will lie within a distance of ±2s of μ with probability 0.95. Take, for example, a trade-association problem where a random sample of 225 firms was selected to estimate the mean number of hourly paid employees in member firms. Suppose the estimators $\bar{x}$ of μ and s of σ yield point estimates $\bar{x}$= 8.31 and s = 4.80. Since the sample size is quite large we can

reasonably expect that in roughly 95 percent of such samples the sample mean will fall within $\frac{2s}{\sqrt{n}} = 9.60/15$ = 0.64 paid employees of μ in either direction. It would thus seem reasonable that by starting with the sample mean 8.31 and adding and subtracting 0.64 we should obtain an interval [7.67 — 8.95] which is likely to include μ.

If we take many large samples and calculate intervals extending two standard deviations of the sample mean on either side of that sample mean for each sample using the estimates of $\bar{X}$ and $S_{\bar{x}}$ obtained, about 95% of these intervals will bracket μ. The probability that any interval so obtained will bracket μ is roughly 0.95 (actually 0.9548).

More formally, consider an interval estimate $L \leq \mu \leq U$ with a specific probability (1 − α) of bracketing μ. The probability that a correct interval estimate (i.e., one that actually brackets μ) will be obtained is called a confidence coefficient and is denoted by (1 − α). The interval $L \leq \mu \leq U$ is called a confidence interval and the limits L and U are called the lower and upper confidence limits, respectively. The numerical confidence coefficient is often expressed as a percentage, yielding the $100\ (1-\alpha)\%$ confidence interval.

The confidence limits U and L for the population mean μ with approximate confidence coefficient (1 − α) when the random sample is reasonably large are

$$\bar{X} \pm z\frac{s}{\sqrt{n}}$$

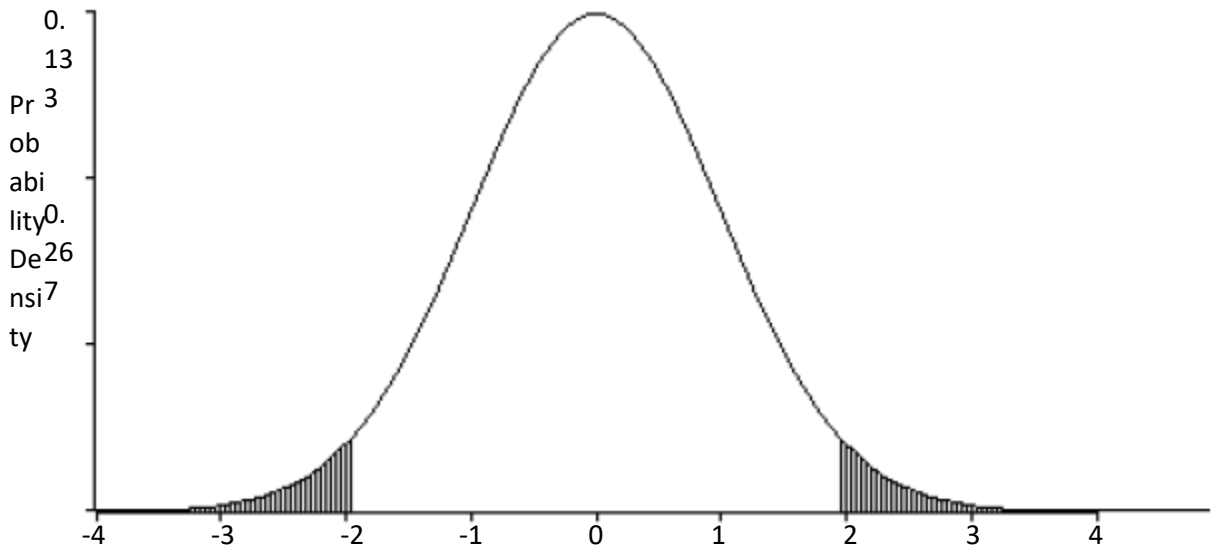
Where $z = z\ (1-\alpha/2)$ is the $100\ (1-\alpha/2)$ percentile of the standard normal distribution. The $100\ (1-\alpha)$ percent confidence interval for μ is

$$\bar{X} - z\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{s}{\sqrt{n}}$$

Note that the confidence interval does not imply that there is a probability (1 − α) that μ will take a value between the upper and lower bounds. The parameter μ is not a variable — it is fixed where it is. Rather, there is a probability (1 − α) that the interval will bracket the fixed value of μ. The limits $-z(1-\alpha/2)$ and $z\ (1-\alpha/2)$ are given by the innermost edges of the shaded areas on the left and right sides of the Figure below. The shaded areas each contain a probability weight equal to α/2. So for a 95% confidence interval these areas each represent the probability weight (1 − 0.95)/2 = 0.05/2 = 0.025 and the sum of these areas represents the probability weight 0.05. The area under the probability density function between the two shaded areas represents the probability weight 0.95. Note also that the probability (1−α) is

chosen in advance of taking the sample. The actual confidence interval calculated once the sample is taken may or may not bracket μ. If it does, the confidence interval is said to be correct.

What confidence coefficient should be chosen? This question hinges on how much risk of obtaining an incorrect interval one wishes to bear. In the trade-association problem above the 90, 95, and 99 percent confidence intervals are



The areas (1−α) and α/2 (shaded) for a standard normal probability distribution with α = 0.05.

$(1-\alpha)$	$(1-\alpha/2)$	z	$S_{\bar{x}}$	$zS_{\bar{x}}$	$\bar{X}$	$\bar{X}-zS_{\bar{x}}$	$\bar{X}+zS_{\bar{x}}$
0.90	0.950	1.645	0.32	0.5264	8.31	7.78	8.84
0.95	0.975	1.960	0.32	0.6272	8.31	7.68	8.94
0.99	0.995	2.576	0.32	0.8243	8.31	7.48	9.13

Note that greater confidence in our results requires that the confidence interval be larger — as (1 − α) gets bigger, α/2 gets smaller and z must increase. We could, of course, narrow the confidence interval at every given level of confidence by increasing the sample size and thereby reducing $\frac{s}{\sqrt{n}}$.

Confidence Intervals with Small Samples (n<30)

In making all the above calculations we standardised the sampling distribution of $\bar{X}$, obtaining

$$z = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

and then calculated limits for μ based on values for z in the table of standard normal probabilities. We used s as an estimator of σ. Had we known σ the standardised value would have been

$$z = \frac{(\bar{X} - \mu)}{s/\sqrt{n}} = \frac{-\mu}{\sigma/\sqrt{n}} + \frac{1}{\sigma/\sqrt{n}}\bar{X}$$

Statistical theory tells us that when the population is normally distributed $\bar{X}$ is normally distributed because it is a linear function of the normally distributed X_i. Then the standardised value z is also normally distributed because it is a linear function of the normally distributed variable$\bar{X}$. But when we use s as an estimator of σ the above expression for z becomes

$$z = \frac{-\mu}{s/\sqrt{n}} + \frac{1}{s/\sqrt{n}}\bar{X}$$

Whereas the divisor $\frac{\sigma}{\sqrt{n}}$is a constant,$\frac{s}{\sqrt{n}}$ is a random variable. Thisimmediately raises the question of the normality of z.It turns out that the variable

$$\frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

is distributed according to the t-distribution, which approximates the normal distribution when the sample size is large. The t-distribution is symmetrical about zero like the standardised normal distribution but is flatter, being less peaked in the middle and extending out beyond the standard normal distribution in the tails. An example is presented in the Figure below. The t-distribution has one parameter, v, and equal to the degrees of freedom, which equals the sample size minus unity in the case at hand. It has mean zero and variance $v/(v - 2)$ with $v > 2$.

Because the t-distribution approximates the normal distribution when the sample size is large and because the Central Limit Theorem implies that $\bar{X}$ is approximately normally distributed for large samples, we could use z = $(\bar{X} - \mu)/s_{\bar{x}}$ to calculate our confidence intervals in the previous examples.

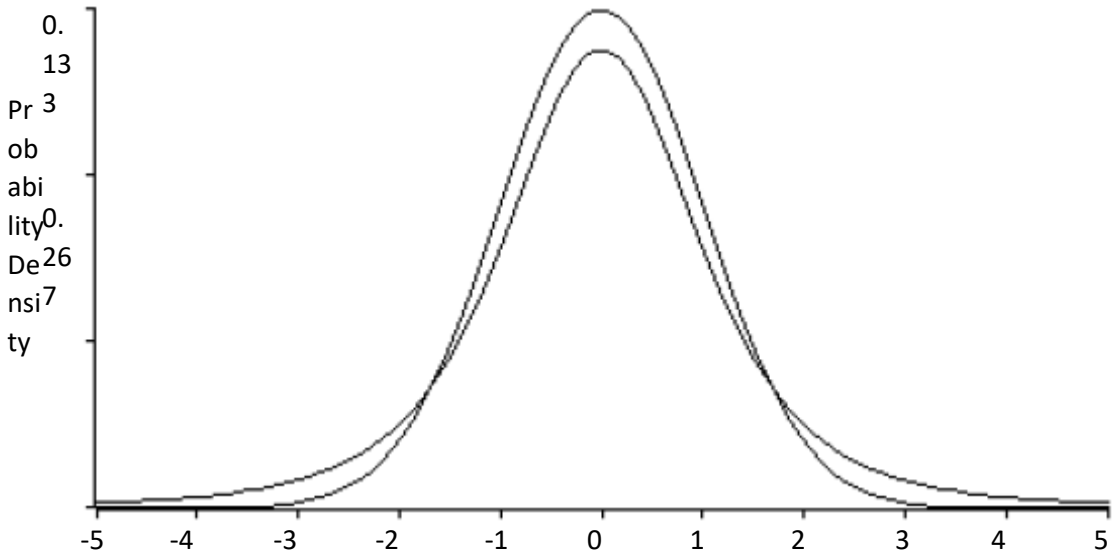
When the sample size is small, however, we must recognize that $(\bar{X} - \mu)/s_{\bar{x}}$ is actually distributed according to the t-distribution with parameter $v = n - 1$ for samples of size n drawn from a normal population. We calculate the confidence interval using the same procedure as in the large sample case except that we now set

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

and use the appropriate percentile from the t-distribution instead of from the normal distribution.

More formally, we can state that the confidence limits for μ with confidence coefficient (1 − α), when the sample is small and the population is normally distributed or the departure from normality is not too marked, are

$\bar{X} \pm tS_{\bar{x}}$



A t-distribution compared to the standard normal distribution. The t-distribution is the flatter one with the longer tails.

Where $t = t(1 - \alpha/2;\ n - 1)$. Expressing t in this way means that the value of t chosen will be the one with degrees of freedom n − 1 and percentile of the $distribution\ 100(1 - \alpha/2)$.

Illustration

Suppose that the mean operating costs in cents per mile from a random sample of 9 vehicles (in a large fleet) turns out to be 26.8 and a value of s equal to 2.5966 is obtained. The standard deviation of the mean is thus$s/3 = 0.8655$. We want to estimate μ, the mean operating costs of the fleet. For a 90% confidence interval, $t\ (0.95;\ 8) = 1.860$.

This implies a confidence interval of

$$26.80 \pm (1.8860)(.8655)$$

or $\quad 25.19 \leq \mu \leq 28.41.$

Had the normal distribution been used, z would have been 1.645, yielding a confidence interval of

$$26.80 \pm 1.4237$$

or

$$25.38 \leq \mu \leq 28.22.$$

Inappropriate use of the normal distribution would give us a narrower interval and a degree of 'false confidence'. Notice that the use of the t-distribution requires that the population be normal or nearly so. If the population is non-normal and n is large we can use z and the standard normal distribution. What do we do if the population is non-normal and the sample size is small? In this case we "cross our fingers" and use the t-distribution and allow that the confidence coefficient is now only approximately 1 − α. This assumes that the t-distribution is robust — i.e., applies approximately for many other populations besides normal ones.

Essentially we are arguing, and there is disagreement among statisticians about this, that the distribution of $\frac{(\bar{X}-\mu)}{S_{\bar{X}}}$ is better approximated by the t-distribution than the normal distribution when the population is non-normal and the sample size is small.

One-Sided Confidence Intervals

Sometimes we are interested in an upper or lower bound to some population parameter. For example, we might be interested in the upper limit of fuel consumption of trucks in a fleet. One-sided confidence intervals are constructed the same as two-sided intervals except that all the risk that the interval will not bracket μ, given by α, is placed on one side. We would thus set a single lower confidence interval at $\bar{X} - z(1 - \alpha)S_{\bar{x}}$ instead of $\bar{X} - z(1 - \alpha/2)S_{\bar{x}}$. A single upper-confidence interval is set in similar fashion. Of course, for small samples we would use *t* instead of *z*.

4. CONCLUSION

This unit has looked at Estimation: Point and Interval Estimation, and we were exposed to the fact that point estimation is not sufficient and there is need to use the interval estimation when such cases arise. However, using interval estimation one needs to be careful when drawing conclusion on a sample. It is therefore advised that the standardised normal distribution (z-test) should be used for n>30 and the t-test distribution for n<30.

5. SUMMARY

In this unit we examined various aspect of Estimation, properties of a good estimator and how to take decision based on the confident interval.

6. TUTOR MARKED ASSIGNMENT

1. Let X_1, X_2, …Xn are normal random variables with mean μ and variance σ2. What are the method of moments estimators of the mean μ and variance σ2?

2. Health insurers and the federal government are both putting pressure on hospitals to shorten the average length of stay of their patients. In 1993 the average length of stay for men in the United States was 6.5 days and the average for women was 5.6 days (Statistical Abstract of the United States: 1995). A random sample of 20 hospitals in one state had a mean length of stay for women in 1996 of 3.6 days and a standard deviation of 1.2 days.

   a. Use a 90% confidence interval to estimate the population mean length of stay for women in the state's hospitals in 1996.

   b. Interpret the interval in terms of this application.

   c. What is meant by the phrase '90% confidence interval'?

7. REFERENCES/FURTHER READINGS

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

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

UNIT 2: Z-TEST AND T-TEST

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Z-test
    3.2 T-test
    3.3 Uses of T-test
    3.4 Assumptions for Student‘s test
    3.5 Unpaired and Paired Two-Sample T-Tests
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

Based on our discussion in Unit 5, we can then go on to discuss z-test and t-test in details. However, in this unit, we shall proceed to carry out series calculation on how to do the unpaired and paired two sample t tests**.**

2.0 OBJECTIVES

The objective of this unit is to introduce z-test and students to t-distribution and emphasize its application in statistics. At the end of this unit, you should be able to;

2.1 know the unpaired and paired two sample t tests

2.2 Understand the t test formula

3.0 MAIN CONTENT

3.1 Z-test

A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Z-test tests the mean of a distribution in which we already know the population variance σ^2. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level in confidence interval, the Z-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more

convenient than the Student's t-test which has separate and different critical values for each sample size (for different sample size, it would have different degree of freedom, which may determine the value of the critical values). Therefore, many statistical tests can be conveniently performed as approximate Z-tests if the sample size is large or the population variance is known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large ($n < 30$), the Student's t-test may be more appropriate.

How to perform a Z test when T is a statistic that is approximately normally distributed under the null hypothesis is as follows:

First, estimate the expected value μ of T under the null hypothesis, and obtain an estimate s of the standard deviation of T.

Second, determine the properties of T: one tailed or two tailed.

For Null hypothesis H_0: $\mu \geq \mu_0$ vs alternative hypothesis H_1: $\mu < \mu_0$, it is upper/left-tailed (one tailed).

For Null hypothesis H_0: $\mu \leq \mu_0$ vs alternative hypothesis H_1: $\mu > \mu_0$, it is lower/right-tailed (one tailed).

For Null hypothesis H_0: $\mu = \mu_0$ vs alternative hypothesis H_1: $\mu \neq \mu_0$, it is two-tailed.

Third, calculate the standard score:

$$z = \frac{(\bar{X} - \mu)}{S} \qquad \textit{where } S \textit{ is the standard error}$$

3.2 T-test

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.

3.3 Uses of T-Test

T-test is used for the following:

(a) A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.

(b) A two-sample location test of the null hypothesis that the means of two populations are equal. All such tests are usually called Student's tests. It should be noted that that name should only be used if the variances of the two populations are also assumed to be equal: the form of the test used when this assumption is dropped is sometimes called WELCH'S TEST. These tests are often referred to as "unpaired" or "independent samples" t-tests, as they are typically when the statistical units underlying the two samples being compared are non-overlapping.

(c) A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, assuming we measure the size of a breast cancer patient's tumour before and after a treatment. If the treatment is effective, we can only expect the tumour size for many of the patients to be smaller following the treatment and it is called a paired or repeated measures t-test.

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

3.4 Assumptions for Student's test

The assumptions of a t-test are as follow:

a. z follows a standard normal distribution under the null hypothesis.

b. S^2 follows a λ^2 distribution (chi-square) with p degrees of freedom under the null hypothesis, where p is positive constant.

c. z and s are independent. However, in a specific type of t-test, these conditions are consequences of the population being studied, and of the way in which the data are sampled. For instance, the t-test comparing the means of two independent samples, the following assumptions should be met.

d. Each of the two populations being compared should follow a normal distribution, and this can be tested using a normality test, or it can be assessed graphically using a normal quartile plot.

e. If using Student's original definition of the t-test, the two populations being compared should have the same variance (using F-test or assessable graphically using a $Q-Q$ plot), but if the sample size in the two groups being compared are equal, student's original t-test is highly the best to the presence of unequal variances.

f. The data used to carry out the test should be sampled independently from the two populations being compared.

SELF ASSESSMENT EXERCISE

List the uses and assumptions of the student's t test

3.5 Unpaired and Paired Two-Sample T-Tests

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

3.5.1 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.

Illustration

It is assumed that the mean systolic blood pressure is $\mu = 120$ mm Hg. In the Honolulu Heart Study, a sample of $n = 100$ people had an average systolic blood pressure of 130.1 mm Hg with a standard deviation of 21.21 mm Hg. Is the group significantly different (with respect to systolic blood pressure!) from the regular population?

Solution

The null hypothesis is H_0: $\mu = 120$, and because there is no specific direction implied, the alternative hypothesis is H_a: $\mu \neq 120$. In general, we know that if the data are normally distributed, then:

$$T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$$

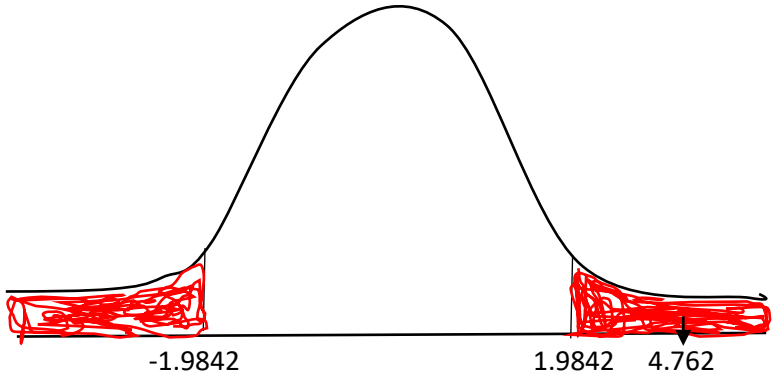
follows a t-distribution with $n-1$ degrees of freedom. Therefore, it seems reasonable to use the test statistic:

$$T = \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}}$$

for testing the null hypothesis $H_0 : \mu = \mu_0$ against any of the possible alternative hypotheses $H_1 : \mu \neq \mu_0$, $H_1 : \mu < \mu_0$, and $H_1 : \mu > \mu_0$. For the example in hand, the value of the test statistic is:

$$t = \frac{(130.1 - 120)}{21.21/\sqrt{100}} = 4.762$$

The critical region approach tells us to reject the null hypothesis at the $\alpha = 0.05$ level if $t \geq t_{0.025,99} = 1.9842$ *or if* $t \leq t_{0.025,99} = -1.9842$. Therefore, we reject the null hypothesis *because* $t = 4.762 > 1.9842$, and therefore falls in the rejection region:



Again, as always, we draw the same conclusion by using the P-value approach. The P-value approach tells us to reject the null hypothesis at the $\alpha = 0.05$ level if the P-value $\leq \alpha = 0.05$. In this case, the P-valueis $2 \times P(T_{99} > 4.762) < 2 \times P(T_{99} > 1.9842) = 2(0.025) = 0.05$:

3.5.2 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".

The T test formula is given as:

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

Where $\bar{x_1}$ = mean of first set values

$\bar{x_2}$ = mean of second set values

S_1 = standard deviation of first set of values

S_2 = standard deviation of second set of values

n_1 = total number of values in first set

n_2 = total number of values in second set

The formulae for standard deviation is given by:

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

Where x = values given

$\bar{x}$= mean

n= total number values

In the previous lesson, we learn how to compare the means of two independent populations, but there may be occasions in which we are interested in comparing the means of two dependent populations. For example, suppose a researcher is interested in determining whether the mean IQ of the population of first-born twins differs from the mean IQ of the population of second-born twins. She identifies a random

sample of n pairs of twins, and measures X, the IQ of the first-born twin, and Y, the IQ of the second-born twin. In that case, she's interested in determining whether:

$\mu_X = \mu_Y$ or equivalently if: $\mu_X - \mu_Y = 0$

Now, the population of first-born twins is not independent of the population of second-born twins. Since all of our distributional theory requires the independence of measurements, we're rather stuck. There's a way out though... we can "remove" the dependence between X and Y by subtracting the two measurements X_i and Y_i for each pair of twins i, that is, by considering the independent measurements

$D_i = X_i - Y_i$

Then, our null hypothesis involves just a single mean, which we'll denote μ_D, the mean of the differences:

$H_0 = \mu_D = \mu_X - \mu_Y = 0$

And then our hard work is done! We can just use the t-test for a mean for conducting the hypothesis test... it's just that, in this situation, our measurements are differences d_i whose mean is $\bar{d}$ and standard deviation is s_D. That is, when testing the null hypothesis $H_0: \mu_D = \mu_0$ against any of the alternative hypotheses $H_1: \mu_D \neq \mu_0, H_1: \mu_D > \mu_0 \text{ and } H_1: \mu_D < \mu_0$, we compare the test statistic:

$$t = \frac{(\bar{d} - \mu_0)}{s_D/\sqrt{n}}$$

to a t-distribution with n−1 degrees of freedom. Let's take a look at an example!

Illustration

Blood samples from n = 10 people were sent to each of two laboratories (Lab 1 and Lab 2) for cholesterol determinations. The resulting data are summarized here: $\bar{X}_1 = 260.6, \bar{X}_2 = 275, \bar{d} = -14.4 \text{ and } s_d = 6.77$. Is there a statistically significant difference at α = 0.01 level, say, in the (population) mean cholesterol levels reported by Lab 1 and Lab 2?

Solution:

The null hypothesis is $H_0: \mu_D = \mu_0$, and the alternative hypothesis is $H_1: \mu_D \neq \mu_0$. The value of the test statistic is: $t = \frac{(-14.4-0)}{6.77/\sqrt{10}} = -6.73$

The critical region approach tells us to reject the null hypothesis at the $\alpha = 0.01$ level if $t > t_{0.01,9} = 3.25$ or if $t < t_{0.01,9} = -3.25$. Therefore, we reject the null hypothesis because $t = -6.73 < -3.25$, and therefore falls in the rejection region.

4.0 CONCLUSION

In this unit, we were exposed to know when we can use the z-test and t-test and we distinguished between the t calculated of the mean and the student's t test. It is worth noting that the paired t-test can be used to reduce the effects of confounding factors in an observational study.

5.0 SUMMARY

This unit, as vividly takes a look at two-sample t-tests for a difference in mean involve independent samples and overlapping samples. We also understood that 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.

6.0 TUTOR MARKED ASSIGNMENT

1. Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 11.8kg and standard deviation is 0.15kg. Does the sample mean differ significantly from the intended weight of 12kg, α=0.05 *(Hint: You are given that for d.f =9, $t_{0.05} = 2.26$)*
2. Boys of a certain age are known to have a mean weight of $\mu = 85$ pounds. A complaint is made that the boys living in a municipal children's home are underfed. As one bit of evidence, $n = 25$ boys (of the same age) are weighed and found to have a mean weight of = 80.94 pounds. It is known that the population standard deviation σ is 11.6 pounds (the unrealistic part of this example!). Based on the available data, what should be concluded concerning the complaint?

7.0 REFERENCES/FURTHER READING

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

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

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

UNIT 3: ANOVA/F-TEST

CONTENTS

1.0 Objectives
2.0 Introduction
3.0 Main Content
3.1 Logic of Analysis of Variance
3.2 Assumption and Steps Involved in Analysis of Variance
3.3 Computation
3.4 Applications of the F-distribution
3.5 For testing equality of population variances
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

A detailed knowledge and understanding of introductory statistics is assumed, it is also expected that students would have familiarised themselves with hypothesis testing.

2.0 OBJECTIVES

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

- calculate the total sum of square
- state sum of square between groups
- explain sum of square within the group
- describe mean square.
- Know the various examples of f tests Statistics
- Understand formulae and analysis of f tests Statistics

3.0 MAIN CONTENT

7.1 Logic of Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) models study the relationship between a dependent variable and one or more independent variables within the same framework as do linear regression models but from a different perspective. The null hypothesis (H_o) tested in the case of ANOVA is that the means of the population

from which the sample is drawn are all equal i.e. $H_o;\ \mu_1 = \mu_2 = \mu_3 = \ \ldots\ldots\ldots\ = \mu_n$ while the alternative hypothesis says that H_o taken as a whole is not true i.e. $H_1;\ \mu_1 \neq \mu_2 \neq \mu_3$.

It is to be noted that each time ANOVA is used, all we are trying to do is to analyse or test the variances in order to test the null hypothesis about the means (i.e. $H_o;\ \mu_1 = \mu_2 = \mu_3$). The ANOVA procedure is based on mathematical theory that the independent sample data can be made to yield two independent estimate of the population variance namely;

(i) Within group variance (or error) this is variance estimate which deals with how different each of the values in a given sample is from other values in the same group.

(ii) Between group variance this is estimate that deals with how the means of the various samples differs from each other.

3.0 Assumptions of ANOVA

(i) Observations are independent and value of any of observation should not be related to the value of another observation.

(ii) Homogeneity of sample variance, it should be assumed that the variance is equal for all treatment populations.

(iii)The values in the population are normally distributed.

7.2 Steps Involved in Anova Analysis

i. Estimate the population variance from the variance between sample means (MSA)

ii. Estimate the population variance from the variance within the samples (MSE)

iii. Compute the fisher ratio. This is given as $F = \frac{MSA}{MSE}\ i.e. F = \frac{Variance\ of\ between\ the\ sample\ mean}{Variance\ of\ within\ the\ sample}$

iv. Compute the various degree of freedom i.e. the degree of freedomfor between, within and total groups.
Degree of freedom for the sum between group is given as C – 1
Degree of freedom within group is writer as (r – 1) c
Total degree of freedom as r – 1
Where c = no of samples
R = no of observations

v. The next thing is to obtain the critical value of F statistics using the F-table in the table, we have the horizontal row which is for degree of freedom of the sum between group numerator. While, the vertical column is meant for within group, check the between degree of freedom along the horizontal axis and within group along vertical axis. This can be checked at either at 0.05 (5%) level of significance or 0.01(1%) level of significance.

vi. Compare the F- statistic value with the critical value if the calculated value is less than the tabulated value, accept the null hypothesis (H_o) and concluded that the difference is not significant. If the calculated value is greater the critical value reject H_o and accept Hi the alternative hypothesis and conclude that the difference is significant.

vii. The result is expected to be summarized on an ANOVA table.

SELF ASSESSMENT EXERCISE

Discuss the assumptions and steps involved in the ANOVA Analysis

Analysis of Variance Table

Sources of variation	Sum of squares	Degree of freedom	Mean Square	F-ratio
Between the means (examples by Factor A)	$SSA = r\sum(\bar{x}_j - \bar{\bar{x}})^2$	$C - 1$	$MSA = \frac{SSA}{C-1}$	$\frac{MSA}{MSE}$
Within the sample (error or unexplained)	$SSE = \sum\sum(x_{ij} - \bar{x}_{ij})^2$	$(r-1)c$	$MSE = \frac{SSE}{(r-1)c}$	-
Total	$SSE = \sum\sum(x_{ij} - \bar{\bar{x}})^2 = \text{SSA} + \text{SSE}$	$rc - 1$	-	-

Where $\bar{x}_j$=mean of sample j composed of r observations $= \frac{\Sigma x_{ij}}{r}$

$\bar{\bar{x}}$=$grandmeanofallcsamples$= $\frac{\Sigma_i\Sigma_j x_{ij}}{rc}$

SSA = Sum of square explained by factor $A = r\Sigma(\bar{x} - \bar{\bar{x}})^2$

SSE = Sum of square of error unexplained by factor $A = \Sigma\Sigma(x_{ij} - \bar{x})^2$

SST = Total Sum of squares = $SSA + SSE = \sum\sum(x_{ij} - \bar{\bar{x}})^2$

Where c = no of samples

r = no of observations in each sample

Illustration

The information below relates to quantities of plastic produced by a plastic industry in 3 sections (morning, afternoon and evening) for 5 weeks. The production data are normally distributed with equal variance.

Table showing production of a plastic industry

Weeks	Morning (X_1)	Afternoon (X_2)	Evening (X_3)
1	85	77	90
2	83	81	92
3	79	75	84
4	81	82	82
5	82	80	87

Is there any significant difference due to production session? Test at 5% level of significance.

Solution

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

$$H_1;\ \mu_1 \neq \mu_2 \neq \mu_3$$

Note let the quantities produced in morning be represented by X_1, afternoon X_2, evening X_3.where r = number of weeks

$$\sum X_1 = 410; \sum X_2 = 395; \sum X_3 = 435$$

$$\bar{X}_1 = \frac{\sum X_1}{r} = \frac{410}{5} = 82;\ \bar{X}_2 = \frac{\sum X_2}{r} = \frac{395}{5} = 79;\ \bar{X}_3 = \frac{\sum X_3}{r} = \frac{435}{5} = 87;$$

$$\bar{X} = \frac{410 + 395 + 435}{5 \times 3} = \frac{1240}{15} \cong 82.67$$

SSA = $5[(82 - 82.67)^2 + (79 - 82.67)^2 + (87 - 82.67)^2]$

$= 5[(-0.67)^2 + (-3.67)^2 + (4.33)^2]$

$= 5(0.4489 + 13.4689 + 18.7489)$

$= 5(32.667)$

$= 163.3335$

SSE $\sum\sum(x_{ij} - \bar{x}_{ij})^2$

$= (85 - 82)^2 + (83 - 82)^2 + (79 - 82)^2 + (81 - 82)^2 + (82 - 82)^2 + (77 - 79)^2 + (81 - 79)^2 + (75 - 79)^2 + (82 - 79)^2 + (80 - 79)^2 + (90 - 87)^2 + (92 - 87)^2 + (84 - 87)^2 + (82 - 87)^2 + (87 - 87)^2$

$= (3)^2 + (1)^2 + (-3)^2 + (-1)^2 + 0^2 + (-2)^2 + (2)^2 + (-4)^2 (3)^2 + (1)^2 + (3)^2 + (5)^2 + (-3)^2 + (-5)^2 + 0$

$= 9 + 1 + 9 + 1 + 0 + 4 + 4 + 16 + 9 + 1 + 9 + 25 + 9 + 25 + 0$

$= 122$

SST $= (85 - 82.67)^2 + (83 - 82.67)^2 + (79 - 82.67)^2 + (82 - 82.67)^2 + (77 - 82.67)^2 + (81 - 82.67)^2 + (75 - 82.67)^2 + (82 - 82.67)^2 + (80 - 82.67)^2 + (90 - 82.67)^2 + (92 - 82.67)^2 + (84 - 82.67)^2 + (82 - 82.67)^2 + (87 - 82.67)^2$

$= (2.33)^2 + (0.33)^2 + (-3.67)^2 + (1.67)^2 + (0.67)^2 + (-5.67)^2 + (1.67)^2 + (-7.67)^2 + (0.67)^2 + (2.67)^2 + (7.33)^2 + (9.33)^2 + (1.33)^2 + (0.67)^2 + (4.33)^2$

$= 5.4289 + 0.1089 + 13.4689 + 2.7889 + 0.4489 + 32.1489 + 58.8289 + 2.7889 + 0.4489 + 7.1289 + 53.7289 + 87.0489 + 1.7689 + 0.4489 + 18.7489$

$= 285.3335$

One-Way Analysis of Variance Table

Sources of variation	**Sum of squares**	**Degree of freedom**	**Mean square**	**F-ratio**
Explained variation (between column)	SSA=163.3335	3-1 =2	$MSA = \frac{163.335}{2} = 81.66675$	$\frac{81.66675}{10.167} = 8.0325$
Unexplained variation or error (within column)	SSE=122	(5 – 1)3 = (4)3 = 12	$MSE = \frac{122}{12} = 10.167$	
Total	285.3335	rc – 1 = 14	-	

$F_{0.05(2,12)}$ = 3.88 (Critical value)

Decision

We reject H_o because $F_{cal} > F_{tab}$ which implies that there is significant difference between the mean of production sessions.

7.3 F-TEST

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.

Examples OF F-Tests

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

(a) Looking at the situation where the hypothesis that the means of given set of normally distributed populations that all the parameters having the same standard deviation are equal. We can call this the best known F test commonly used in statistical test and it plays a key role in the analysis of variance (ANOVA).
(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.

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:

$$F = \frac{explained\ Variance}{unexpected\ variance}\ or\ \frac{between-group\ variability}{within-group\ variability}$$

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, $y\bar{y}$denotes the overall mean of the data and k is the number of groups.

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

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

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

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

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

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

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

7.4 Two- Way Analysis of Variance

This is an extension of one- way ANOVA, the difference between them is that, here, we can test for two (2) null hypothesis, one for factor A and the other for factor B.

For two way analysis, the set of observation involved are classified into two (2) factors or criteria; treatment factor or criteria and block or homogenous factor or criteria. As we have discussed in one factor- analysis of variance, the total variation is divided or splitted into 3 components.

- Variation between treatment
- Variation between blocks and
- Residual or error variation

Two-way classification table

		Treatment (Factor A)	
		1 2 3 …… t	Total
Block factor B	1	Y_{11} Y_{12} Y_{13} …… Y_{1j}	B_1
	2 3 4 5 ⋮ ⋮ ⋮	Y_{21} Y_{22} Y_{23} …… Y_{2t}	B_2
	B	Yb1 Yb2 Yb3 …… Ybt	B_b

The Formulas

Column means is given by $\frac{\sum x_{ij}}{r}$

Row means of given $\frac{\sum x_{ij}}{c}$

Grand mean is given by $\bar{\bar{x}} = \Sigma\frac{x_{ij}}{r} = \Sigma\frac{x_{ij}}{c}$

The subscripted dot signifies that more than one factor is under consideration.

SST = $\sum\sum (\bar{x}_{ij} - \bar{\bar{x}})^2$

SSA = $r\sum (\bar{x}_{ij} - \bar{\bar{x}})^2$ between column variation

SSB = $c\sum (\bar{x}_{ij} - \bar{\bar{x}})^2$ between row variation

SSE = SST – SSA – SSB

Degree of freedom of SSA = c – 1

Degree of freedom of SSB = r – 1

Degree of freedom of SSE = (r-1) (c – 1)

Degree of freedom of SST = rc – 1

MEAN SQUARE

MSA = $\frac{SSA}{c-1}$

MSB = $\frac{SSB}{r-1}$

MSE = $\frac{SSE}{(r-1)(c-1)}$

F- STATISTICS

F-ratio for factor A = $\frac{MSA}{MSE}$F-ratio for factor $\frac{MSB}{MSE}$It is to be noted that; two (2) separate null hypothesis is considered.

- H_o; There is no difference between mean of treatment
- H_o; There is no difference between mean of block.

Illustration

Samples taken involving two (2) interactive factors A & B in a two analysis of variance experience gives the result below:

Table showing interactive factors A and B

	Treatment A			
Block (B)	22	11	10	5
	13	10	8	6
	7	9	6	2

You are carry out a 2-way analysis of variance at 0.05 level of significance?

Solution

Hypothesis

1. $H_0; \mu_1 = \mu_2 = \mu_3 = \mu_4;\ H_1; \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$
2. $H_0; \mu_1 = \mu_2 = \mu_3 = \mu_4;\ H_1; \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Two-Way Classification Table

	Treatment A				**Total**	**Sample mean**
Block B	22 13 7	11 10 9	10 8 6	5 6 1	48 37 23	$\bar{x_1}$ = 12 $\bar{x_2}$ = 9.25 $\bar{x_3}$ = 5.75
Total	42	30	24	12	108	$\Sigma\bar{x}i$ = 27
Sample mean	42/3 $x_{.1}$ = 14	30/3 $x_{.2}$ = 10	24/3 $x_{.3}$ = 8	12/3 x.4 = 4		$\bar{\bar{x}}$ = 9

$$\textbf{SST} = \sum\sum(\bar{x}_{ij} - \bar{\bar{x}})^2$$

=$(22 - 9)^2 = (13)^2 = 169$; $(11 - 9)^2 = (2)^2 = 4$; $(10 - 9)^2 = (1)^2 = 1$; $(5 - 9)^2 = (-4)^2 = 16$;

$(13 - 9)^2 = (4)^2 = 16$; $(10 - 9)^2 = (1)^2 = 1$; $(8 - 9)^2 = (-1)^2 = 1$; $(6 - 9)^2 = (-3)^2 = 9$;

$(7 - 9)^2 = (-2)^2 = 4$; $(9 - 9)^2 = (0)^2 = 0$; $(6 - 9)^2 = (-3)^2 = 9$; $(1 - 9)^2 = (-8)^2 = 64$

= 189 = 5 = 11 = 89

:. SST = 189 + 5 + 11 + 89 = 294

SSA $= r\sum(\bar{x}_{ij} - \bar{\bar{x}})^2$

$= 3\,[(14 - 9)^2 + (10 - 9)^2 + (8 - 9)^2 + (4-9)^2]$

$= 3\,[5^2 + (1)^2 + (-1)^2 + (-5)^2$

$= 3\,(25 + 1 + 1 + 25)$

$= 3\,(52)$

$= 156$

SSB $= c\sum(\bar{x}_{ij} - \bar{\bar{x}})^2$

$= 4[(12 - 9)^2 + (9.25 - 9)^2 + (5.75 - 9)^2]$
$= 4\,[(3)^2 + (0.25)^2 + (-3.25)^2]$
$= 4\,(9 + 0.0625 + 10.5625)$
$= 4\,(19.625)$
$= 78.5$

SSE = SST – SSA – SSB

$= 294\text{-}156 - 78.5$

$= 59.5$

Degree of Freedom

SSA = c – 1 = 4 -1 = 3
SSB = r – 1 = 3 – 1 = 2
SSE = (r-1) (c – 1) = (3 – 1) (4 – 1) = (2) 3 = 6
SST = rc – 1 = (4 x 3) – 1 = 12 – 1 = 11

Mean Square

MSA = $\frac{SSA}{c-1} = \frac{156}{3} = 52$

MSB = $\frac{SSB}{r-1} = \frac{78.5}{2} = 39.25$

MSE = $\frac{SSE}{(r-1)(c-1)} = \frac{59.5}{6} \cong 9.92$

F-ratio for factor A = $\frac{MSA}{MSE} = \frac{52}{9.92} = 5.24$

F-ratio for factor $\frac{MSB}{MSE} = \frac{39.25}{9.92} = 3.97$

Sources of variation	**Sum of squares**	**Degree of freedom**	**Mean square**	**F ratio**
Explained variation by factor A (between column)	SSA=156	C – 1 = 3	MSA = 52	$\frac{MSA}{MSE} = 5.24$
Explained variation by factor B (between rows)	SSB=78.5	r – 1 = 2	MSB = 39.25	$\frac{MSB}{MSE} = 3.97$
Unexplained variation or error	SSE=59.5	(r – 1)(c-1) = 6	MSE= 9.92	-
Total	294	11	-	-

Decision Criteria

1. Factor A Critical Value$F_{3,6}$= 4.76Because $F_{cal.}$ >F_{tab}. Reject H_o and accept H_1 meaning that the mean of factor A are not equal.
2. Factor B Critical Value$F_{2,6}$= 5.14. Since $F_{cal.}$ <F_{tab}. Accept H_o and reject H_1 conclude that the mean of factor B are all equal**.**

4.0 CONCLUSION

In the course of our study of one-way analysis of variance and two-way analysis of variance you must have learnt about;Explained variation, Unexplained variation, Total variation, Sum of square of Factor A, Sum of square of Factor B, Sum of square of the error term, Mean square of Factor A, Mean square of Factor B, F-ratio of both Factor A and Factor B, and Sum of Square of total variation.

5.0 SUMMARY

In the course of our discussion of one-way analysis of variation the following definitions were inferred

$SSA = r\sum(\bar{x}_j - \bar{\bar{x}})^2$; $MSA = \frac{SSA}{C-1}$; $SSE = \sum\sum(x_{ij} - \bar{x}_{ij})^2$; $MSE = \frac{SSE}{(r-1)c}$;

$SSE = \sum\sum(x_{ij} - \bar{\bar{x}})^2 = \text{SSA} + \text{SSE}$ as well as that of two-way analysis of variation.

6.0 TUTOR-MARKED ASSIGNMENT

A physiologist was interested in learning whether smoking history and different types of stress tests influence the timing of a subject's maximum oxygen uptake, as measured in minutes. The researcher classified a subject's smoking history as either heavy smoking, moderate smoking, or non-smoking. He was interested in seeing the effects of three different types of stress tests — a test performed on a bicycle, a test on a treadmill, and a test on steps. The physiologist recruited 9 non-smokers, 9 moderate smokers, and 9 heavy smokers to participate in his experiment, for a total of n = 27 subjects. He then randomly assigned each of his recruited subjects to undergo one of the three types of stress test. Here are his resulting data:

	Test		
Smoking History	**Bicycle (1)**	**Treadmill (2)**	**Step Test (3)**
	12.8, 13.5, 11.2	16.2, 18.1, 17.8	22.6, 19.3, 18.9
	10.9, 11.1, 9.8	15.5, 13.8, 16.2	20.1, 21.0, 15.9
	8.7, 9.2, 7.5	14.7, 13.2, 8.1	16.2, 16.1, 17.8

Is there sufficient evidence at the α = 0.05 level to conclude that smoking history has an effect on the time to maximum oxygen uptake? Is there sufficient evidence at the α = 0.05 level to conclude that the type of stress test has an effect on the time to maximum oxygen uptake? And, is there evidence of an interaction between smoking history and the type of stress test?

7.0 REFERENCES/FURTHER READIDING

Adedayo, O. A. (2006). Understanding Statistics. Yaba, Lagos: JAS Publishers.

Esan, F. O. &Okafor, R. O. (2010). Basis Statistical Method (Revised edition). Lagos: Toniichristo Concept.

Murray, R.S. & Larry, J. S. (1998). (Schaum Outlines Series). Statistics. (3rd ed.). New York: Mcgraw Hills.

Olufolabo, O.O. &Talabi, C. O.(2002). Principles and Practice of Statistics. Shomolu Lagos: HASFEM Nig Enterprises.

UNIT 4: CHI-SQUARE

CONTENTS

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

1.0 INTRODUCTION

A Chi-square can said to be a measurement of how expectations are compared to results. The 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) is a statistical hypothesis test where the sampling distribution of the test statistic, is a chi-squared distribution when the null hypothesis H_0 is true.

2.0 OBJECTIVES

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

1. Know the examples of chi-square distribution –
2. Understand the application of chi-square analysis

3.0 MAIN CONTENT

3.1 Examples of Chi-Square Distribution

1. Pearson's Chi-Square Test

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

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

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

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

2. Discrete Uniform Distribution

The formula for the discrete Uniform Distribution is given as:

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

When N observations are divided by n cells, but the degree of freedom reduction gives p = 1 because the observed frequencies O_i are contrasted 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 t 6. The numbers of times you rolled the die does not determine the number of degree of freedom. More so, the chi-square test statistic is as follows:

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

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

O_i= observed frequency

E_i = Expected (theoretical) frequency supported by null hypothesis.

n = Number of cells in table

And the degree of freedom is $(r - i)(n-1)$where r is the numbers of rows and nn 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. **McNemar'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.

3.2 Application of Chi-Square Analysis

3.2.1 Application on Type of 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.

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

3.2.2 2×2 Application of Contingency Table

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

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

General notation for a 2×2 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:

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

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 tutorial test on a group of female students and you hypothesized that the female students attending the tutorial would show increase in class performance compared to those that did not attending the tutorials. Assuming you conduct the study and collected the following data:

Null Hypothesis ($\boldsymbol{H}_0$): The female students whose class performance increased are independent of tutorials.

Alternative Hypothesis ($\boldsymbol{H_1}$): The Female students whose class performance increased are associated with tutorials.

	Class Performance Increased	**No Class Performance Increased**	Total
Attended	46	18	64
Did not attend	40	31	71
Total	86	49	135

Using the formulae specified above:

$$X^2 = \frac{(135)\left((46 \times 31) - (18 \times 40)\right)^2}{(64)(86)(49)(71)} = \frac{67288860}{19148416} = 3.51$$

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 DDD = $(c-1)(R-1)$.

∴Df = $(2-1)(2-1)=1$.

But the chi-square statistic (X^2=3.51)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 otherwise

But in the care of our example, the chi-square calculated value is 3.51 and chi-square tabulated is 3.841, therefore, the chi-square tabulated (3.841) is greater than the chi-square calculated (3.77) then conclude that the female students whose class performance increased are independent of tutorials.

SELF ASSESSMENT EXERCISE

Discuss any five (5) examples of chi square distribution

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

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

When to Use the Chi-Square Goodness of Fit Test

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

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

3.2.4 Chi-Square Test of Homogeneity

This lesson explains how to conduct a chi-square test of homogeneity. The test is applied to a single categorical variable from two different populations. It is used to determine whether frequency counts are distributed identically across different populations. The test for homogeneity is a method, based on the chi-square statistic, for testing whether two or more multinomial distributions are equal.

When to Use Chi-Square Test for Homogeneity

- The test procedure described in this lesson is appropriate when the following conditions are met:
- For each population, the sampling method is simple random sampling.
- The variable under study is categorical.
- If sample data are displayed in a contingency table (Populations x Category levels), the expected frequency count for each cell of the table is at least 5.

3.2.5 Chi-Square Test for Independence

This lesson explains how to conduct a chi-square test for independence. The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

When to Use Chi-Square Test for Independence

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

• The sampling method is simple random sampling.

• The variables under study are each categorical.

• If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

Illustration

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

	Republican	Democrat	Independent	Row total
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

Solution

H_0: Gender and voting preferences are independent.

H_1: Gender and voting preferences are not independent.

Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = \frac{(n_r * n_c)}{n}$$

	Republican	Democrat	Independent	Row total
Male	$E_{1,1}$	$E_{1,2}$	$E_{1,3}$	E_{r1}
Female	$E_{2,1}$	$E_{2,2}$	$E_{2,3}$	E_{r2}
Column total	E_{c1}	E_{c2}	E_{c3}	1

$$E1,1 = \frac{(400 * 450)}{1000} = 180;\ E1,2 = \frac{(400 * 450)}{1000} = 180;\ E1,3 = \frac{(400 * 100)}{1000} = 40$$

$$E2,1 = \frac{(600 * 450)}{1000} = 270;\ E2,2 = \frac{(600 * 450)}{1000} = 270;\ E2,3 = \frac{(600 * 100)}{1000} = 60$$

$$X^2 = \sum_{i=r,c}^{n} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

$$X^2 = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} + \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60}$$

$$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, nr is the number of observations from level r of gender, n_c is the number of observations from level c of voting preference, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count when gender is level r and voting preference is level c, and $O_{r,c}$ is the observed frequency count when gender is level r voting preference is level c. The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

We use the Chi-Square Distribution table $X^2=5.99_{0.05,2.}$ Since the $X^2=5.99_{0.05,2}$ is less than $X^2_{calculated}=16.2$ we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

Note: Specifically, the approach is appropriate because the sampling method was simple random sampling, the variables under study were categorical, and the expected frequency count was at least 5 in each cell of the contingency table.

4.0 CONCLUSION

Chi-square analysis is one of the statistical test that measure how the expectations is being compared to the results and different examples of chi-square were examined such as Pearson's chi-square test, Yate's correction for continuity, Cochran – Mantel-Haenzel Statistics, Mac Nemar's test and turkey's test of additivity In conclusion, chi-squared analysis has very wide applications which include test of independence of attributes; test of goodness fit; test of equality of population proportion and to test if population has a specified variance among others. This powerful statistical tool is useful in business and economic decision making.

5.0 SUMMARY

In this unit, we have examined the concept of chi-square and its scope. We also look at its methodology and applications. It has been emphasized that it is not just an ordinary statistical exercise but a practical tool for solving day-to-day business and economic problems. The unit has vividly examined and discussed the analysis of chi-square statistic and the applications to several calculations such as $2x2$ contingency table, categorical and numerical analysis and the chi-square of Goodness of fit for one sample. Therefore I belief at this point you must have really understood the use and calculation of chi-square statistic.

6.0 TUTOR MARKED ASSIGMENT

1. The head of a surgery department at a university medical centre was concerned that surgical residents in training applied unnecessary blood transfusions at a different rate than the more experienced attending physicians. Therefore, he ordered a study of the 49 Attending Physicians and 71 Residents in Training with privileges at the hospital. For each of the 120 surgeons, the number of blood transfusions prescribed unnecessarily in a one-year period was recorded. Based on the number recorded, a surgeon was identified as either prescribing unnecessary blood transfusions Frequently, Occasionally, Rarely, or Never. Here's a summary table (or "contingency table") of the resulting data:

Physician	Frequent	Occasionally	Rarely	Never	Total
Attending	2 (4.1%)	3 (6.1%)	31 (63.3%)	13 (26.5%)	**49**
Resident	15 (21.1%)	28 (39.4%)	23 (32.4%)	5 (7.0%)	**71**
Total	17	31	54	18	**120**

Are attending physicians and residents in training distributed equally among the various unnecessary blood transfusion categories?

2. Is age independent of the desire to ride a bicycle? A random sample of 395 people were surveyed. Each person was asked their interest in riding a bicycle (Variable A) and their age (Variable B). The data that resulted from the survey is summarized in the following table:

		Variable B (Age)				
	OBSERVED	18-24	25-34	35-49	50-64	Total
Variable A	Yes	60	54	46	41	201
	No	40	44	53	57	194
	Total	100	98	99	98	395

Is there evidence to conclude, at the 0.05 level, that the desire to ride a bicycle depends on age?

7.0 REFERENCES/FURTHER READING

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

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

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

MODULE 3: TEST OF HYPOTHESIS AND REGRESSION ANALYSIS

Unit 1: Hypothesis Testing

Unit 2: Simple Regression

Unit 3: Multiple Regression

Unit 4: Time Series

UNIT 1: HYPOTHESIS TESTING

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 Statistical Hypothesis
    3.2 Definitions
    3.3 Type 1 and Type 2 Errors
    3.4 The Criterion of Significance
        3.4.1 One-Tailed and Two-Tailed Tests
        3.4.2 Procedures for Carrying Out Tests for Hypothesis
        3.4.3 Statistical Test for Mean of a Single Population when Population Variance is known.
        3.4.4 Testing Difference between Two Means of Independent Samples
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assessment
7.0 References/Further Readings

1.0 INTRODUCTION

One of the uses of Statistics is to make a decisive decision. However, in some cases the decisions may be on a population based on results obtained from selected samples of the said given population. 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.

2.0 OBJECTIVES

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

- know the meaning of hypotheses testing
- understand type 1 and type 2 errors
- understand the procedures for carrying out test of Hypothesis.
- know how to calculate statistical test for mean of a single population
- be able to differentiate between one-tailed and two-tailed tests and calculate one-tailed and two-tailed tests
- understand the test of difference between two means of independent samples.

3.0 MAIN CONTENTS

3.1 STATISTICAL HYPOTHESIS

Despite the fact that population parameters are usually unknown, we now know how to estimate population means, variances and proportions. The population mean is the most important in this course, and we have seen how to obtain both point and interval estimators and estimates of this parameter, whether or not sampling takes place from a normal distribution. It is very common for an investigator to have some sort of preconception or expectation (in the ordinary sense of the word) about the value of a parameter, say the mean, μ: Statisticians usually talk about having a hypothesis about the value of the parameter. However, we already have some theory or hypothesis about what the population parameters are and we need to use our sample statistics to determine whether or not it is reasonable to conclude that the theory or hypothesis is correct. Statistical procedures used to do this are called ***statistical tests***.

A ***statistical hypothesis test*** is a method of making decisions using data from a scientific study. In statistics, a result is interpreted as being statistically significant if it has been predicted as unlikely to have occurred

by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "***test of significance***" was coined by statistician Ronald Fisher. These tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance; this can help to decide whether results contain enough information to cast doubt on conventional wisdom, given that conventional wisdom has been used to establish the null hypothesis. The critical region of a hypothesis test is the set of all outcomes which cause the null hypothesis to be rejected in favour of the alternative hypothesis.

Statistical hypothesis testing is sometimes called confirmatory data analysis, in contrast to exploratory data analysis, which may not have pre-specified hypotheses. Statistics are helpful in analysing most collections of data. This is equally true of hypothesis testing which can justify conclusions even when no scientific theory exists.

3.2 DEFINITIONS

Hypothesis Testing:A (parametric) hypothesis is a statement about one or more population parameters. Given a random sample from a certain population, is the sample evidence enough to disregard a particular belief about that population? (E.g.: the value of a parameter). This hypothesis can be tested using a hypothesis test. A hypothesis test consists of:

1. Two complementary hypotheses: the null hypothesis and the alternative hypothesis, denoted H_0 and H_1 respectively.
2. A decision rule that specifies for which sample values the null hypothesis is not rejected ('accepted'), and for which sample values the null hypothesis is rejected in favour of the alternative hypothesis.

The set of sample values for which H_0 is rejected is called the rejection or critical region. The complement of the critical region is called the acceptance region (where H_0 is accepted).

A hypothesis that if true completely specifies the population distribution, is called a simple hypothesis; one that does not is called a composite hypothesis.

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 ($\boldsymbol{H_1}$) ∴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. Null Hypothesis (***H*₀**): There is no difference in their mean scores of male and female students in Edo University. ($\mu_1 = \mu_2$).

   Alternative Hypothesis (***H*₁**)**:**∴**There is a difference in their mean scores, ($\mu_1 \neq \mu_2$).

   So this is a non-directional hypothesis. More so, after setting up the null and alternative hypotheses statistical test are carried out to justify whether to reject the null hypothesis or the alternative hypothesis using a level of significance (1% or 5% level of significance).

3.3 HYPOTHESIS ERRORS (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 is when an invigilator who is supposed to raise alarm when a student cheats it the exams hall, he must therefore decide between:

H_0: The student is not cheating and

H_1: The student is cheating and the student should be expelled.

A false alarm by the invigilator 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 the student is not cheating 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 alpha risk but the probability of committing a type 2

error is denoted by the Greek letter beta (β)and it's called beta risk. The probability of correctly rejecting (H_0) when it is false is called the power of the statistical test and it's denoted by $1-\beta$ while the probability of correctly accepting H_0 is equal to $1-\alpha$.

We wanted to minimize our chance of making a Type 1 error! In general, we denote α = P(Type 1 error) = the "significance level of the test." Obviously, we want to minimize α. Therefore, typical α values are 0.01, 0.05, and 0.10.

In general, we denote β = P(Type 2 error). Just as we want to minimize α = P(Type I error), we want to minimize β = P(Type 2 error). Typical β values are 0.05, 0.10, and 0.20. We will discuss more in the criterion of significance.

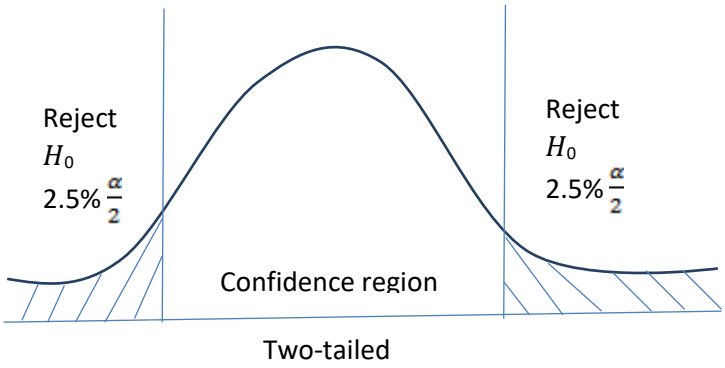
3.4 THE CRITERION OF SIGNIFICANCE

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, a 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.

3.4.1 One-Tailed and Two-Tailed Tests

- **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.
- **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

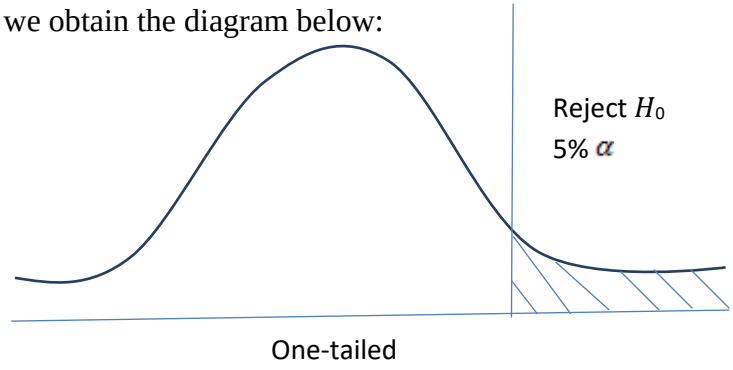
(sides) of a normal distribution, although the test can be used in other non-normal distributions. The normal curve is one of the most popular models used in statistical tests of hypothesis. For a non-directional hypothesis, a two-tailed test is used when finding the critical region. If the 0.05 level 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:



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

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

For the 0.05 level, we obtain the diagram below:



SELF ASSESSMENT EXERCISE

Differentiate between type I and type II errors

3.4.2 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.

3.4.3 Statistical Test for Mean of a Single Population when Population Variance is known.

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

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

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.

Illustration

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

Solution:

Using the formula $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{52-49}{\frac{10}{\sqrt{25}}} = 1.5$

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 H_0. Which implies that the population mean is not greater than 49.

Illustration 2:

A midwife claims that the mean weights of babies delivered at her maternity clinic is 3.5kg. A statistician takes a sample of 10 babies and obtains the following weights: 2.8, 2.5, 3.2, 3.5, 3.7, 2.7, 4.0, 4.5, 3.9, 3.6. Test the midwife's claim at 0.05 level of significance.

Solution:

H_0: μ=3.5, and $_1$: $\mu \neq$3.5.

The first step here is to find the mean of the sample and the standard deviation using formula

$$\bar{x}=\frac{\sum x}{N}=\frac{2.8+\ 2.5+\ 3.2+\ 3.5+\ 3.7+\ 2.7+\ 4.0+\ 4.5+\ 3.9+\ 3.6}{10}=3.44$$

We then obtain mean = 3.44, using the formula for standard deviation $\sigma=\frac{\Sigma(x-\bar{x})^2}{n-1}=0.6$

It should be noted that the distribution of $(\bar{X})$ *has a t-distribution when the population is normal and* $n\leq30$. $t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$

$$t=\frac{3.44-3.5}{0.6/\sqrt{10}}=0.32$$

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. Which implies that the mean weights of babies delivered at her maternity clinic is 3.5kg.

3.4.4 Testing Difference between Two Means of Independent Samples

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:

$$t = \frac{(\bar{X} - \bar{Y})}{\frac{\sqrt{(N_X - 1)S_X^2 - (N_Y - 1)S_Y^2}}{N_X + N_Y - 2}\left(\frac{1}{N_X} + \frac{1}{N_Y}\right)}$$

Where the degrees of freedom is$N_X + N_Y - 2$. Note that $\bar{X}$ *and* $\bar{Y}$are the respective sample means of the two groups. S_X and S_Y are the standard deviation, N_X and N_Y are the sample size of the two groups. The above formula is used when the population variance is known.

$$\frac{(N_X - 1)S_X^2 - (N_Y - 1)S_Y^2}{N_X + N_Y - 2} \; is\ the\ pooled\ sample\ varience$$

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X + \mu_Y)}{\sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}}} \; when\ \sigma\ is\ known$$

at_r distribution where r, the adjusted degrees of freedom is determined by the equation:

$$r = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{(S_X^2/n)^2}{n_X - 1} + \frac{(S_Y^2/n)^2}{n_Y - 1}} \; two\ tail - test\ df$$

If r doesn't equal an integer, as it usually doesn't, then we take the integer portion of r. That is, we use [r] if necessary.

With that now being recalled, if we're interested in testing the null hypothesis: H_0: $\mu_X = \mu_Y$ *or* $\mu_X - \mu_Y = 0$

Illustration

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

Solution:

H_0:$\mu_X = \mu_Y$ of no difference and H_1:$\mu_X \neq \mu_Y$ of difference

Let N_X=22, N_Y=20, S_X=12, S_Y=10, $\bar{X}$=81, $\bar{Y}$=78, Using the formula: $t = \frac{(\bar{X}-\bar{Y})}{\frac{\sqrt{(N_X-1)S_X^2-(N_Y-1)S_Y^2}}{N_X+N_Y-2}\left(\frac{1}{N_X}+\frac{1}{N_Y}\right)}$

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

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.

4.0 CONCLUSION

Hypothesis testing has been seen as the way forward to justify the truth and false of a research situation or problem. So at this end, I believe you must have gained enough on how we formulate hypothesis. This unit has look at the statistical test for hypothesis, and we can conclude that hypothesis testing is a process by which an analyst tests a statistical hypothesis. The methodology employed by the analyst depends on the nature of the data used, and the goals of the analysis. The goal is to either accept or reject the null hypothesis. We also looked at one and two tailed test. However, the unit analyses that in statistical significance testing, a one-tailed test or two-tailed test are alternative ways of computing the statistical significance of a data set in terms of a test statistic, depending on whether only one direction is considered extreme (and unlikely) or both directions are considered equally likely.

5.0 SUMMARY

This unit has vividly taken a look at the meaning of hypothesis and how it is used in solving some research problems. However, type 1 and type 2 errors were also examined to know whether we have accepted or rejected what we do not need to accept or reject. One and two tailed test has be discussed extremely in this unit and their graphs was also examined. Therefore we can conclude that one-tailed tests are used for asymmetric distributions that have a single tail, such as the chi-squared distribution, which are common in measuring goodness-of-fit, or for one side of a distribution that has two tails, such as the normal distribution, which is common in estimating location; this corresponds to specifying a

direction. Two-tailed tests are only applicable when there are two tails, such as in the normal distribution, and correspond to considering either direction significant. However, we have examine various aspect of statistical test for hypothesis such as procedure for carrying out tests for hypothesis, statistical test for mean of a single population when population variance is known and unknown and interval estimation for mean of a single population. Therefore, it is at this point that I believe that you must have learnt a lot from the unit.

6.0 TUTOR MARKED ASSIGNMENT

A psychologist was interested in exploring whether or not male and female college students have different driving behaviours. There were a number of ways that she could quantify driving behaviours. She opted to focus on the fastest speed ever driven by an individual. Therefore, the particular statistical question she framed was as follows: Is the mean fastest speed driven by male college students different than the mean fastest speed driven by female college students? She conducted a survey of a random n = 34 male college students and a random m = 29 female college students. The results of her survey shows that the mean fastest speed driven by male and college students are 105.5 and 90.9 and their standard deviation are 20.1 and 12.2 respectively.

Is there sufficient evidence at $\alpha = 0.05$, level to conclude that the mean fastest speed driven by male college students differs from the mean fastest speed driven by female college students?

7.0 REFERENCES/FURTHER READINGS

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

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

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.

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

UNIT 2: SIMPLE REGRESSION

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Parameter Estimation Strategies

3.2 Assumptions Of The Linear Stochastic Regression Model

3.3 Ordinary Least Square Estimators

3.4 Numerical Estimation of Parameters

3.5 Coefficient of Determination (R^2)and The Regression Line

3.6 Statistical Significance of b_i Using Standard Error Test.

3.7 Z test of Statistical Significance of OLS Estimates

3.8 T- test of Significance of bi

4.0 Summary

5.0 Conclusion

6.0 Tutor-Marked Assignment

7.0 References/Further Reading

1.0 INTRODUCTION

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as Regression Analysis.

Linear model shows the relationship between two variables. In this relationship, one variable is depending on the other variable. The model consists of the independent variables and the constant term, with their respective coefficient and we need to estimate the parameters of the model in order to know the magnitude of their relationship.

Consider a familiar supply function of the form $Y_i = b_o + b_1 x$…… (1)

This function shows the positive linear relationship between quantity supply, Y and price of the commodity, X. The dependent variable in this model is the quantity supply, denoted by Y, while the independent variable (explanatory variable) is the price, X. This is a two variable case with two parameters representing the

intercept and the slope of the function. This supply-price relationship, Y = f(x) is a one way causation between the variables Y and X: price is the cause of changes in the quantity supply, but not the other way round. From the above equation (…1), the parameters are b_o and b_1, and we need to obtain numerical value of these parameters.

The left hand variable Y is variously referred to as the endogenous variables, the regressand, the dependent variable or the explained variable. Similarly, the right hand variable X is variously described as exogenous variables, the regressor, the independent variable or the explanatory variable.

2.0 OBJECTIVE

The objective of this unit is to expose learners to understanding and importance of the basic idea of regression analysis being to predict the behaviour of one variable (the predict) based on fluctuations in one or more related variables (the predictors). That, It is possible to estimate the behaviour of the predict using multiple predictors (multiple regression), and it is also possible to predict the behaviour of the predict and based on nonlinear relationships with the predictors. At the end of this unit you should be able to:

1. Discuss parameter estimation procedures.
2. Discuss the assumptions of the stochastic variable.
3. Discuss the assumptions of the explanatory variables.
4. Attempt algebraic estimation of simple regression
5. Coefficient of determination, r^2,
6. Standard error test
7. Z and t-statistic (test)

3.0 MAIN CONTENT

3.1 PARAMETER ESTIMATION STRATEGIES

The parameters of this model are to be estimated using ordinary least square (OLS) method. We shall employ this method for a start due to the following reasons.

i. The computational procedure using this method is easy and straight forward.
ii. The mechanics of the OLS method are simple to understand.
iii. This method always produces satisfactory results.
iv. The parameter estimates using the O.L.S. method are best, Linear and unbiased. This makes the estimates to be more accurate compared with the estimates obtained using other methods.
v. The OLS method is an essential component of most econometric techniques.

Note that the model $Y_i = b_o + b_1 x.$ implies an exact relationship between Y and X that is, all the variation in Y is due to changes in X only, and no other factor(s) responsible for the change. When this is represented on a graph, the pairs of observation (Y and X) would all lie on a straight line

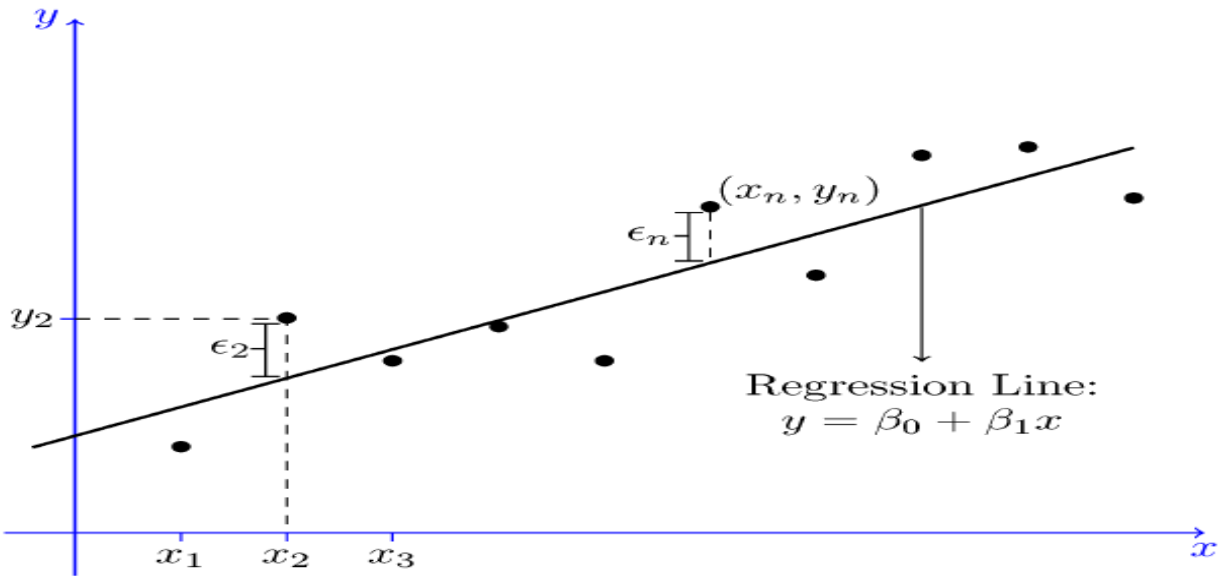
Ideally, if we gather observations on the quantity actually supplied in the market at various prices and plot them on a diagram, we will notice that they do not really lie on a straight line.

3.1.1 Sources of Deviations in Parameter Model Estimation

There are deviations of observations from the line. These deviations are attributable to the following factors:

i. Omission of variable(s) from the function on ground that some of these variables may not be known to be relevant.
ii. Random behaviour of human beings. Human reactions at times are unpredictable and may cause deviation from the normal behavioural pattern depicted by the line.
iii. Imperfect specification of the mathematical form of the model. A linear model, for instance, may mistakenly be formulated as a non-linear model. It is also possible that some equations might have been left out in the model.
iv. Error of aggregation – usually, in model specification, we use aggregate data in which we add magnitudes relating to individuals whose behaviour differs. The additions and approximations could lead to the existence of errors in econometric models.
v. Error of measurement – this error arises in the course of data collection, especially in the methods used in the collection of data. Data on the same subject collected from Central Bank of Nigeria and National Bureau of Statistics could vary in magnitude and units of measurements. Therefore when you use different sources you could get different results.

Line of Regression: This is the line which gives the best estimate of one variable for any given value of the other variable. Therefore, the line of regression of y on x is the line which gives the best estimates for the value of y for any specified value of x.



Regression line is the line that best represents the data points

The term best fit is interpreted in accordance with the principle of least squares which involves minimising the sum of the squares of the residuals or the errors of the estimates i.e, the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit.

It should be noted that several lines can be drawn from the same set of pairs of observations plotted in the form of a scattered diagram, but the best fit line gives the best estimate of the dependent variable for any given level of independent variable.

3.1.2 The Uses of Random variable in Models

The inclusion of a random variable usually denoted by U, into the econometric function help in overcoming the above stated sources of errors. The U's is variously termed the error term, the random disturbance term, or the stochastic term.

This is so called because its introduction into the system disturbs the exact relationship which is assumed to exist between Y and the X. Thus, the variation in Y could be explained in terms of explanatory variable X and the random disturbance term U.

That is $Y_i = b_o + b_1 x + U_i$…… (ii)

Where Y = variation in Y; $b_o + b_1 x$ = systematic variation, U_i = random variation. Simply put, variation in Y = explained variation plus unexplained variation. Thus, $Y_i = b_o + b_1 x + U_i$is the true relationship that connects the variable Y and X and this is our regression model which we need to

estimate its parameters using OLS method. To achieve this, we need observations on X, Y and U. However, U is not observed directly like any other variables, thus, the following assumptions hold:

3.2 ASSUMPTIONS OF THE LINEAR STOCHASTIC REGRESSION MODEL

3.2.1 Assumptions With Respect To Random Variable

In respect to the random variable U, the following assumptions apply to any given econometric model that is used in prediction of any economic phenomenon:

(i) U_i is a random variable, this means that the value which U_i takes in any one period depends on chance. Such values may be positive, negative or zero. For this assumption to hold, the omitted variables should be numerous and should change in different directions.

(ii) The mean value of "U" in any particular period is zero. That is, E (u_i) denoted by U is zero. By this assumption, we may express our regression as $Y_i = b_o + b_1 x.$

(iii) The variance of ui is constant in each period. That is, Var (U_i) = E (U_i)^2 = □(U_i)^2 = □^2U which is constant. This implies that for all values of x, the U's will show the same dispersion about their mean. Violation of this assumption makes the U's heteroscedastic.

(iv) U has a normal distribution. That is, a bell shaped symmetrical distribution about their zero mean. Thus, U = N(0, 1).

(v) The covariance of U_i and U_j = 0. $i \neq j$. This assumes the absence of autocorrelation among the U_i. In this respect, the value of U_i in one period is not related to its value in another period.

3.2.2 Assumptions in terms of the relationship between 'u' and the explanatory variables.

The following assumptions also hold when you conduct a regression analysis in terms of the relationship between the explanatory variable and the stochastic variable:

i. U_i and X do not covary. This means that there is no correlation between the disturbance term and the explanatory variable. Therefore, cov. X_u = 0.

ii. The explanatory variables are measured without error. This is because the U_i absorbs any error of omission in the model.

3.2.3 Assumptions in relation to the explanatory variable(s) alone.

The following assumptions are made.

(i) The explanatory variables are not linearly correlated. That is, there is absence of multicollinearity among the explanatory variables. This means that cov. $X_iX_j = 0$. $i \neq j$ (This assumption applies to multiple regression model).

(ii) The explanatory variables are correctly aggregated. It is assumed that the correct procedures for such aggregate explanatory variables are used.

(iii) The coefficients of the relationships to be estimated are assumed to have a unique mathematical form. That is, the variables are easily identified.

(iv) The relationships to be estimated are correctly specified.

3.3 ORDINARY LEAST SQUARE ESTIMATORS

The ordinary least square (OLS) is among the best method used in obtaining estimates of the parameter $\hat{b}_1$and $\hat{b}_0$. The use of OLS method in estimating economic relationship is based on the fact that the estimates of the parameters have some optimal properties.

Generally, in choosing a particular estimation method, one should aim at such method that gives an estimate, which is close to the value of the true population parameter; the variation of which (if at all it exists) will be within only a small range around the true parameter.

For any estimation method, the goodness of the estimator is judged on the basis of the following desirable properties.

(i) Unbiasedness: - An estimator is unbiased if its mean is zero i.e. $E\ (\hat{b}_1) - b_1 = 0$.In such case, the unbiased estimator changes to the true value of the parameter as the number of samples increases. An unbiased estimator always gives, on the average, the true value of the parameter. The cases of biased and unbiased estimator of the true value are as illustrated diagrammatically.

(ii) Minimum Variance – An estimator is best if it has the smallest variance compared with any other estimate obtained using other methods. By minimum variance, we mean that the values of the parameter b_i clusters very closely around the true parameter b.

(iii) Efficient estimator: An estimator is efficient when it combines the property of unbiasedness with minimum variance property. Symbolically bi is efficient if the following two conditions are fulfilled.

$(a)\ E(b) = b$: $(b)\ E[\hat{b} - E(\hat{b})]^2 < E\ [(b*)]^2$

Where b* is another unbiased estimate of the true b. This means that in the class of unbiased estimators, such estimator has a minimum variance.

(iv) Linear Estimator: An estimator is linear if it is a linear combination of the given sample data. Thus, with the sample observation Y_1, Y_2. …, Y_n, a linear estimator takes the form: $K_1Y_1 + K_2Y_2 + \ldots + K_nY_n$, where the k_i s are some constants.

(v) BLUE (Best, Linear, Unbiased, Estimator): This property is abbreviated to BLUE meaning that the estimator is best (having minimum variance) linear, and unbiased as compared with all other linear unbiased estimators. Thus all the properties (i-iv) are included in the BLUE property.

(vi) Minimum mean square error (MSE) Estimators. This property combines unbiasedness and the minimum variance properties. An estimator, therefore, is a minimum MSE estimator if it has the smallest mean square error, defined as the expected value of the squared difference of the estimator around the true population parameter b. that is, $MSE\left(\hat{b}\right) = \left(\hat{b} - b\right)^2$

(vii) Sufficiency: This property implies that the estimator uses all the available information a sample contains about the true parameter. For this property to hold, the estimator should accommodate all the observations of the sample, and should not give room for any additional information in connection with the true population parameter.

The ordinary least square method (OLS) satisfies the above stated properties. For this reason, the method seems to be the best and most widely used of all the estimation methods. In a nut shell, the OLS has the BLUE (best, linear, unbiased properties among the class of linear and unbiased estimators). The linearity property as previously discussed implies that the parameter estimates are linear functions of the observed Y_i. That is, the estimators b_o and b_1 includes the variable Y and X in the first power.

Thus, $\hat{b}_1$ = f(Y). This property enables one to compute the values of the parameter estimates with ease.

The unbiased property of the OLS estimates implies that the expected value of the estimates parameter is equal to the true value of the parameter, that is, E ($\hat{b}_1$ = b).

The importance of this property lies in the fact that for large samples, the parameter estimates obtained will on the average give a true value of the *b's.*

The minimum variance property becomes desirable when combined with unbiasedness. The importance of this property is obvious when we want to apply the standard tests of significance for b_0 and b_i, and to construct confidence intervals for these estimates. Because of the minimum variance they have, their respective confidence intervals will be narrower than for other estimates obtained using any other econometric procedures.

The smaller confidence interval obtained is interpreted to mean in effect that we are extracting more information from our sample than we would be, if we were to use any other methods which yielded the same unbiased estimates.

3.4 NUMERICAL ESTIMATION OF PARAMETERS

The following procedures are used in finding numerical values of the parameters b_o and b_i.

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

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

$$\hat{b} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{n\sum XY - \sum \bar{X} \sum \bar{Y}}{n\sum X^2 - \sum (X)^2} = \frac{Cov(x, y)}{Var(x)} = \frac{\sum xy}{\sum x^2}$$

See working in UNIT 5 ESTIMATION

Where $x = (X - \bar{X})$ *and* $y = (Y - \bar{Y})$ are the deviations of the variables from their respective means and the summation is over i = 1 . . . n. and $b_0 = \hat{a}$ and $b_1 = \hat{b}$

Illustration

Given the following data on the supply of commodity W, find the estimated supply function (Table 1).

No	Y_i (Quantity)	X_i (Price)
1	64	8
2	68	10
3	44	6
4	48	9
5	50	6
6	65	10
7	45	7
8	56	8

The given expression for b_0 and b_1 lead us to reproduce the above Table as seen in table 2.

	Y	X	X^2	XY	y=Y-	x=X-	y^2	x^2	xy	$\hat{Y}$	e	e^2

					$\bar{Y}$	$\bar{X}$						
1	64	8	64	512	9	0	81	0	0	55	9	81
2	68	10	100	680	13	2	169	4	26	64	4	16
3	44	6	36	264	-11	-2	121	4	22	46	-2	4
4	48	9	81	432	-7	1	49	1	-7	59.5	-11.5	132.25
5	50	6	36	300	-5	-2	25	4	10	46	4	16
6	65	10	100	650	10	2	100	4	20	64	4	16
7	45	7	49	315	-10	-1	100	1	10	50.5	-5.5	30.25
8	56	8	64	448	1	0	1	0	0	55	1	1
n=8	**ΣY= 440**	**ΣX= 64**	**ΣX^2= 530**	**ΣXY= 3601**	**Σy=0**	**Σx=0**	**=646**	**Σx^2=18**	**Σxy =81**			**296.5**

$$\bar{Y} = \frac{\Sigma y}{n} = 55. \quad X = \frac{\Sigma x}{n} = 8$$

Therefore, using upper case letters,

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum x^2 - \sum (X)^2} = \frac{8\,(3601) - (64)(440)}{8(530) - (64)^2} = 4.5$$

$$b_0 = \bar{Y} - b_1\bar{X} = 55 - 4.5(8) = 19$$

The regression equation is $\hat{Y} = 19 + 4.5x + U_i$

SELF ASSESSMENT EXERCISE

Explain the properties that are used to judge a good estimator

3.5 R^2 and the Regression Line

The coefficient of determination, r^2, is used in determining goodness of fit of the regression line obtained using the OLS method. That is, it is used in testing the explanatory power of the linear regression of Y on X.

Thus, in order to determine the degree to which the explanatory variable is able to explain the variation in the dependent variable Y, the r^2 provides a useful guide.

If we measure the dispersion of observations around the regression line some may be closer to the line while others may be far away from it. Our argument is that the closer these observed values are to the line,

the better the goodness of fit. On the basis of this, we may turn out to state that a change in the dependent variable, Y, is explained by changes in the explanatory (independent) variable X. To know precisely the extent to which the total variation in Y is explained by the independent variable, X, we compute the value of r^2 as the ratio of explained variation to total variation. That is, $r^2 = \frac{Explained\ variation}{Total\ variation} = \frac{\Sigma \hat{y}^2}{\Sigma y^2}$

However, the coefficient of determination, r^2 determines the proportion of the variation in Y which is explained by variation in X. If for instance, $r^2 = 0.75$, it means that 75% of the variation in Y is due to the variation in X, while 25% of the variation is explained by the disturbance term, u. Thus, the regression line gives a good fit to the observed data.

If, however, $r^2 = 0.45$, it means that only 45% of the variation in Y is as a result of variation in X, while 55% of the variation is due to the disturbance term. This is a poor indication and the regression line does not give a good fit to the observed data. Generally, if r^2 is 0.5 and above, it shows a good fit while a value of r^2 less than 0.5 shows a bad fit.

Note that the r^2 is much of relevance when the estimated model is used for forecasting. Note also that the value of $r^2 = 0$ signifies that the independent variable cannot explain any change in the dependent variable, hence variation in the independent variable has no effect on the dependent variable.

3.6 STATISTICAL SIGNIFICANCE OF b_i USING STANDARD ERROR TEST.

To use this test, it is important to know the mean and variance of the parameter estimates $\hat{b}_0$ and $\hat{b}_1$. It has been established that the mean and variance of $\hat{b}_0$ and $\hat{b}_1$ are respectively.

$$E(\hat{b}_0) = b_o ; Var\hat{b}_0 = E(\hat{b}_0 - b_o)^2 = \delta_u^2 \frac{\Sigma x_i^2}{n \Sigma x_i^2}$$

$$E(\hat{b}_1) = b_1 ; Var\hat{b}_1 = E(\hat{b}_1 - b_1)^2 = \frac{\delta_u^2}{n \Sigma x_i^2}$$

The standard error test enables us to determine the degree of confidence in the validity of the estimate that is, from the test, we are able to know whether the estimates b_o and b_1 are significantly different from zero. The test is mainly useful when the purpose of the research is the explanatory (analysis) of economic phenomena and the estimation of reliable values.

We formally start by stating the null hypothesis, H_0: $\hat{b}_1 = 0$,

Against the alternative hypothesis H_1: $\hat{b}_1 \neq 0$

The standard error for the parameter estimates b_0 and b_1 are respectively computed as shown:

1. S ($\hat{b}_0$) = $\sqrt{\text{Var } \hat{b}_0}$ = $\sqrt{\delta_u^2 \frac{\sum x^2}{n \sum x^2}}$

   But $\delta_u^2 = \frac{\sum e_i^2}{n-2}$

   S ($\hat{b}_0$) = $\sqrt{\delta_u^2 \frac{\sum x^2}{n \sum x^2}} = \left(\frac{\sum e^2}{n-2}\right)\left(\frac{\sum x^2}{n \sum x^2}\right)$

2. S ($\hat{b}_i$) = Var $\hat{b}_1$ = $\frac{\delta_u^2}{n \sum x_i^2} = \left(\frac{\sum e^2}{n-2}\right)\left(\frac{1}{\sum x^2}\right)$

When the numerical values for the s ($\hat{b}_0$) and s ($\hat{b}_1$) are each compared with the numerical values of $\hat{b}_0$and $\hat{b}_1$the following decision rules apply.

i. If $S\hat{b}_1 < \hat{b}_1/2$, we reject the null hypothesis and conclude that b_1 is statistically significant.

ii. If on the other hand, $S\hat{b}_1 > \hat{b}_1/2$ we accept the null hypothesis that the true population parameter $b_i = 0$. This concludes that the estimate is not statistically significant. Therefore, change in X has no effect on the values of Y.

The acceptance of the null hypothesis has economic implication. Thus, the acceptance of the null hypothesis $\hat{b}_1 = 0$ implies that the explanatory variable to which this estimate relates does not influence the dependent variable Y, and should not be included in the function. This situation renders the relationship between Y and X, hence the regression equation is parallel to axis of the explanatory variable X in the case, $Y = b_o + 0X_i$ or Y = bo. The zero slope indicates that no relationship exist between Y and X.

Similarly, if the null hypothesis b0 = 0 is accepted, on the basis that s($\hat{b}_0$) > ½ b_0, it implies that the intercept of this regression line is zero. Therefore, the line passes through the origin. In this case, the relationship between Y and X will be $Y = 0 + b_1 X_i, or\ Y = b_1 X.$

Illustration

Refer to the example 1 on the regression analysis, the r^2 and the standard errors S ($\hat{b}_0$), s ($\hat{b}_1$) are as computed:

Solution

From the computed values of the Table.

$\Sigma xy = 81$, $\Sigma x^2 = 18$ $\Sigma Y^2 = 646$, $\Sigma e^2 = 296.5$, n = 8; K = 2 Where K = degrees of freedom.

$$R^2 = \frac{\sum (XY)^2}{\sum X^2 \sum Y^2} = \frac{6561}{11628} = 0.56$$

Hence, X is a fairly good predicator of Y about 56% of the variation in Y is explained by variation in the explanatory variable X, while 44% is due to disturbance term u. Thus, the regression line fairly gives a good fit to the observed data. On the statistical significance of the parameter estimate we compute:

$$S(\hat{b}_0) = \left(\frac{\sum e^2}{n-2}\right)\left(\frac{\sum x^2}{n\sum x^2}\right) = \frac{296.5\times 18}{6\times 8\times 18} = 6.177$$

$$S(\hat{b}_i) = \left(\frac{\sum e^2}{n-2}\right)\left(\frac{1}{n\sum x^2}\right) = \left(\frac{296.5}{8-2}\right)\left(\frac{1}{18}\right) = 2.75$$

Thus, the result of our regression may be presented formally as:

$$\hat{Y} = 19 + 4{,}5x$$

$$(6.177)\ (2.75), R^2 = 0.56, n = 8$$

From the above estimated result,

$\hat{b}_0$ =19; $\hat{b}_1$=4.5; S ($\hat{b}_0$) = 6.177; S($\hat{b}_1$) =2.75

Therefore, ½ $\hat{b}_0$and we accept the null hypothesis that H_o: $b_o = 0$.

This shows that $\hat{b}_0$is not statistically significant. Similarly s($\hat{b}_1$) < ½$\hat{b}_1$and we reject the null hypothesis. Thus, the estimate $\hat{b}_1$is statistically significant. Note: the acceptance and meaningful interpretation of any econometric result entails a combination of high r^2 and low standard errors.

3.7 Z TEST OF STATISTICAL SIGNIFICANCE OF OLS ESTIMATES

The z test is employed when the sample size is sufficiently large (i.e$n \geq 30$). It could be applied whether the population variance is known or not. The z test is applied using the formula, $Z_i^* = \frac{\hat{b}_1}{s\hat{b}_1}$ where the z* is the calculated zi which is compared with the z table (theoretical value of z) at a given level of significance, say 5%. If $-z < z* < +z$ at 0.025, we accept the null hypothesis that H_0:$b_1 = 0$; and conclude that the estimate is not statistically significant. If however $z* > z$, then we accepts that H_1: $\hat{b}0 \neq 0$ and the estimate is statistically significant.

Illustration

Consider $\hat{Y} = 19 + 4.5x$

$$(6.177)\ (2.75), R^2 = 0.56, n = 8$$

To conduct a z test for the estimates $\hat{b}0$, and $\hat{b}1$ at 5% level of significance we proceed as follows:

H_0: $\hat{b}_1 = 0$ (Null hypothesis)

H_1 :$\hat{b}_1 \neq 0$ (Alternative hypothesis)

$Z^* = \frac{\hat{b}_0}{s\hat{b}_0} = \frac{19}{6.177} = 3.08$(for b_0)

$Z^* = \frac{\hat{b}_1}{s\hat{b}_1} = \frac{4.5}{2.75} = 1.64$ (for b1)

Z table at 5% (i.e 0.025 since it is a two tail test) level of significance = 1.96

*b*0, Z* > Z Table; for b1, Z1* < Z Table

Therefore, the null hypothesis is rejected, and concluded that estimate b_1 is not statistically significant at 0.05 level while b_0 is statistically significant.

3.8 T- TEST OF SIGNIFICANCE of bi

The students test is used when the sample size is small (i.e. n < 30) provided that the population of the parameter follows a normal distribution. With this in view, and taking degrees of freedom into consideration, we need to compare this with the theoretical t, at given level of significance, say 5%.

Our null and alternative hypothesis are respectively formulated thus

H_0: $b_i = 0$

H_i: $b_i \neq 0$

Following a normal distribution, our t is computed as follows:

$t = \frac{\hat{b}_1}{s\hat{b}_1}$

As stated previously, the empirical t (t) value is compared with the table t with n - k degrees of freedom, given a 5% level of significance.

Decision rule: if – t*0.025 < t < t0.025 (with n - k degree of freedom), we accept the null hypothesis, and conclude that our estimate b_1 is not statistically significant at 0.05 level of significance.

If however, t* > + t0.025, we reject the null hypothesis, and accept the alternative hypothesis. This concludes that the estimate b_1 is statistically significant.

Illustration

Given a sample size of n=18, the model was estimated to be:

$$\hat{Y} = 23 + 7.5X$$

$$(10.2)\ (1.4)$$

We wish to test the reliability of the estimates $\hat{b}_0$ and $\hat{b}_1$respectively. From the estimated model,

$\hat{b}_0$=23, S($\hat{b}_0$) 10.2; $\hat{b}_1$ =7.5. S($\hat{b}_1$) =1.4

Therefore,

$$t* = \frac{\hat{b}_0}{S(\hat{b}_0)} = \frac{23}{10.2} = 2.25\ (for\ b_0)$$

$$t* = \frac{\hat{b}_1}{S(\hat{b}_1)} = \frac{7.5}{1.4} = 5.4\ (for\ b_1)$$

The critical values of t for (n – k) or 18 -2 = 16 d.f. are:

$$-t_{0.025} = -2.12 \text{ and } +t = 2.12.$$

For the estimate $\hat{b}_0$, we can see that t*>t we reject the null hypothesis and conclude that the estimate b_0 is statistically significant.

In the case of the estimate $\hat{b}_1$since t*>t we also reject the null hypothesis and conclude that b_1 is statistically significant at 5% significance level.

4.0 CONCLUSION

In this unit you have learn how to determine and interpret coefficient of multiple determination R^2, the Z test and the T test for evaluating an estimated model. Evaluating the model helps in the assessment of reliability, how the model estimates can be used for policy and applications. Simple linear regression was examined in this unit and we got to know that it is the relationship between the dependent and independent variable. However, various calculation of the straight line graph was also examined in this unit that calculation of regression equation $y = \beta_0 + \beta_1 X_1$ of y on x and x on y respectively.

5.0 SUMMARY

This unit has done justice to the simple linear regression analysis and the meaning and the uses of simple linear regression analysis were discussed in detail and also the calculation of the straight line regression equation and the following were derived in the unit. R^2, T and Z tests are important components of econometric test that are used in evaluating the strength of econometric variables. A good and continuous learning of these methods will enable you to conduct an effective and useful econometric research.

6.0 TUTOR MARKED ASSIGNMENT

The following data relate to the model $Y_i = \alpha + \beta X_i + \epsilon_i$ where the Xi are assumed non-stochastic and the $\epsilon_i$are assumed to be independently identically normally distributed with zero mean and constant variance.

i	Yi	Xi
1	21	10
2	18	9
3	17	8
4	24	11
5	20	11
6	20	10
7	22	12
8	21	11
9	17	9
10	20	9

a) Calculate the regression estimates of α and β.

b) R^2 and Calculate a 95% confidence interval for β.

7.0 REFERENCES/ FURTHER READING

Berndt, Ernst R.(1991) The practice of Econometrics: Classic and contemporary, Addison-Wesley,

Damodar, N. G., Dawn, C. P., and Sangetha, G. (2012): Basic Econometrics. Tata McGraw Hill Education Private Ltd. New Delhi, India.

Davidson, James (2000). Econometric Theory, Blackwell Publishers, Oxford, U.K..

Dominick, S. and Derrick, R. (2011): Statistics and econometric (Schaum outline) (2nd edition) McGraw Hill, New York.

Gujarati, D.N. (2003. Basic Econometrics. Tata Mc-Graw – Hill PublishingCompany Ltd New-Delhi.

Koutsoyiannis, A. (1977) Theory of Econometrics An Introductory ExpositionEconometric Methods Macmillan

Oyesiku, O.K. and Omitogun, O. (1999): Statistics for Social and Management Sciences: Higher Education Books Publishers Lagos.

Oyesiku, O.O., Abosede, A.J., Kajola, S.O, and Napoleon, S.G.(1999): Basics of Operation research. CESAP Ogun State University. Ago-Iwoye, Ogun State.

Wooldridge, J. M. (2009) Introductory Econometrics A modern Approach, Cengage Learning Singapore 4th Edition

UNIT 3: MULTIPLE REGRESSION

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
  3.1 Multiple regression
  3.2 Assumptions of multiple regression
  3.3 Estimation of multiple regression parameters
  3.4 Multiple Correlation Coefficient (R) Coefficient of Determination (R^2) Defined
  3.5 Test of Significance
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/ Further Readings

1.0 INTRODUCTION

For your success in this course of study it is required that you have a thorough knowledge of simple regression model, hypothesis testing among others. For example, if we want to develop a model to estimate the quantity of bread demanded we can expect the latter to depend, at the very minimum, on the price of bread, on the prices of at least some substitutes and on real income.

Regression equation is an expression by which you may calculate a typical value of a dependent variable say Y, on the basis of the values of independent variable(s). While simple regression analysis is useful for many purposes, the assumption that the dependent variable Y depends on only one independent variable is very restrictive.

Multiple regression model attempts to expose the relative and combine importance of the independent variables on dependent variables.

Multiple regression models is one among the commonly used tools in research for the understandings of functional relationship among multi-dimensional variables. The model attempts to expose the relative and combine effect of the independent variable on the dependent variable.

2.0 OBJECTIVE

At the end of our discussion on multiple regression you should be to;

(i) Regress the independent variable on the dependent variable

(ii) Understand parameter estimates involved

(iii) You should know how to calculate the values of b_o, b_1, b_2, … b_n

(iv) Test of significance: Coefficient of multiple determinations and Test of overall significance of the regression

3.0 MAIN CONTENT

3.1 Multiple Regression and Assumptions Defined

Multiple regression analysis is usually used for testing hypothesis about the relationship between a dependent variable Y and two or more independent variable X and for prediction or forecasting. Three variable linear regression models is usually written as:

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \cdots\cdots + \beta_k XK_i + \mu_i$$

Where Y = dependent variable

β_0 = intercept

β_1, β_2, β_k = partial correlation coefficient or regression coefficient

□ = error term or residuals

Where i = 1 . . . n and the μ_i are independently normally distributed with mean zero and constant variance σ^2. Actually, we can often get away with less restrictive assumptions about the μ_i, namely

$$E\{\mu_i\} = 0$$

and

$$E\{\mu_i \mu_j\} = 0, i \neq j$$

$$E\{\mu_i \mu_j\} = \sigma^2, i = j.$$

This says that the μ_i must be independently distributed with constant variance but not necessarily normally distributed. Our problem is to estimate the parameters β_k, k = 0 . . .K, and σ and to establish confidence intervals and conduct appropriate statistical tests with respect to these parameters.

3.2Assumptions of Multiple Regressions

Multiple regression models has the following assumptions

i. Randomness
ii. Normality
iii. Measurement error
iv. Independent of □ and x_s
v. Correct specification of model
vi. Multi-colinearity
vii. Homoscedascity
viii. Linearity

3.3 Estimation of the Parameters of the Multiple Regression (b_o, b_1 …b_n)

For the purpose calculation and because of the parameters involved deviation method of calculating regression will be used. The parameters involve are define as stated below:

$$\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}_1 - \hat{b}_2\bar{X}_2$$

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Illustration

Given the following data estimate the multiple regression equation using Real GDP as dependent variable and Private consumption or Government consumption as independent variable

	1990	**1991**	**1992**	**1993**	**1994**	**1995**
Gov. Con. (₦b)X_2	35	98	83	76	93	77
Private Con. (₦b)X_1	35	22	27	16	28	46

RGDP(₦b) Y	20	15	17	16	27	35

Find the least square regression equation of Real GDP, Private consumption and Government consumption.

Solution

Years	**Y**	**X_1**	**X_2**	**y= Y-$\bar{Y}$**	**x_1= X_1-$\bar{X}_1$**	**x_2= X_2-$\bar{X}_2$**	**x_1x_2**	**x_1^2**	**x_2^2**	**x_1y**	**x_2y**	**y^2**
1990	20	35	35	-1.67	6	-42	-252	36	1764	-10.00	70.00	2.78
1991	15	22	98	-6.67	-7	21	-147	49	441	46.67	-140.00	44.44
1992	17	27	83	-4.67	-2	6	-12	4	36	9.33	-28.00	21.78
1993	16	16	76	-5.67	-13	-1	13	169	1	73.67	5.67	32.11
1994	27	28	93	5.33	-1	16	-16	1	256	-5.33	85.33	28.44
1995	35	46	77	13.33	17	0	0	289	0	226.67	0.00	177.78
n = 6	**130**	**174**	**462**	**0**	**0**	**0**	**-414**	**548**	**2498**	**341**	**-7**	**307**

$$\hat{b}_1 = \frac{(341)\,(2498) - (-7)(-414)}{(548)\,(2498) - (-414)^2} = \frac{848920}{1192512} \cong 0.712$$

$$\hat{b}_2 = \frac{(-7)\,(548) - (341)\,(-414)}{(548)\,(2498) - (-414)^2} = \frac{137338}{1192512} \cong 0.115$$

$$\hat{b}_0 = 21.7 - 0.712(29) - 0.115(77) = -7.803$$

The regression of Y on X_1 and X_2 is as written below

$$\hat{Y} = -7.803 + 0.712X_1 + 0.115X_2$$

The equation above is the multiple regression of value of Real GDP, private consumption and government consumption.

SELF ASSESSMENT EXERCISE

List the assumptions of the multiple regression models

3.4 Multiple Correlation Coefficient (R) Coefficient of Determination (R^2) Defined

Multiple correlation coefficients represented by R measures the degree of linear association between two or more variables. Say variable Y and the entire explanatory variable jointly. Coefficient of determination (R^2) is defined as the proportion of the total variation in Y explained by the multiple regression of Y on X_1 and X_2. It measures goodness of fit of the regression equation. In a three variable model we are always interested in knowing the proportion of the variation in Y explained by each of the explanatory variable X_1 and X_2.

$$R^2 = \frac{\hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2}{\sum y^2}$$

The value of R^2 lies between 0 and 1, if it is 1, the fitted regression line explains 100% of the variation in Y, on the other hand, if it is 0, the model does not explain any of the variation in Y. typically, however, R^2 lies between these two extremes values. The fit is said to be better, the closer R^2 is to 1.

Illustration

From the above example find R^2

Solution

$$R^2 = \frac{0.712(341) + 0.115(-7)}{307} \cong 0.79$$

This implies that the explanatory variable (x_1 and x_2) can only account for 79% variation in variable Y i.e. both x_1 and x_2 contributes 78.8% to the explanation of the variation in Y.

3.5 Test of Significance

The overall significance of the regression can be tested with the ratio of the explained to the unexplained variance. This follows an F-distribution with k – 1 and n – k degree of freedom, where n is the number of observations and k is the number of parameters estimated.

The joint hypothesis can be tested by the analysis of variance (ANOVA).

The F-statistics or F-ratio for the test of significance can be written as:

$$H_o: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$$

$$F_{k-1,n-k} = \frac{\sum y^2 - \sum e^2/(k-1)}{\sum e^2/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Also ANOVA table can as well be used for test of significance.

Source of variation	**Sum of squares**	**DF**	**Mean sum of square**
Due to Regression (ESS)	$\hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2$	2	$\frac{\hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2}{2}$
Due to Residual (RSS)	$\sum \mu_i^2$	n – 3	$\sigma^2 = \frac{\sum \mu_i^2}{n-3}$
Total	$\sum y_i^2$	n – 1	

Note:

$$\sum y_i^2 = \hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2 + \sum \mu_i^2$$

$$TSS = ESS + RSS$$

$$\left(\frac{\frac{\hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2}{2}}{\sum \mu_i^2/(n-3)} \right)$$

F-Statistics Table

Null hypothesis H_0	Alternative Hypothesis H_1	Criteria Region Reject H_0 if
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\frac{S_1^2}{S_2^2} >$ F$\alpha, ndf, ddf (ftab)$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$\frac{S_1^2}{S_2^2} >$ F$\alpha/2$, ndf, ddf (f_{tab}) or <$F_{1-}\alpha/2$, ndf, ddf

If the calculated F-ratio (Fc) exceeds the tabular value of F (F_{tab}) at the specified level of significance and degree of freedom, the hypothesis is accepted that the regression parameters are not all equal to zero and that R^2 is significantly different from zero.

If the null hypothesis is true, it gives identical estimates of true σ^2. This statement should not be surprising because if there's a trivial relationship between y and x_1 and x_2 the source of variation in Y will be due to the random forces usually represented by e_i or $\square_1$. If however, the null hypothesis is false, that is x_1 and x_2 actually influence Y; the equality will not hold. Here, the ESS will be relatively larger than the RSS taking due account of their respective degree of freedom. Therefore, the F-.ratio provides a test of the null hypothesis that the true slope coefficients are simultaneously zero.

DECISION CRITERIA; If the F-ratio calculated exceeds the critical F-value from the table at the α percent level of significance we reject H_o; otherwise do not reject it. Alternatively if the F_{-cal} of the observed F is sufficiently low accept H_o.

Illustration

From our previous example above it is glaring that calculation of Σe^2 is required, so there's need to generate a new table apart from the one generated in unit one of this modules as to get the value for our Σe^2.

Solution

Test of Significance Table

Years	**Y**	**X_1**	**X_2**	$\hat{Y}$	**e**	**e^2**
1990	20	35	35	21.14	-1.14	1.30
1991	15	22	98	19.13	-4.13	17.07
1992	17	27	83	20.97	-3.97	15.73
1993	16	16	76	12.33	3.67	13.48
1994	27	28	93	22.83	4.17	17.41
1995	35	46	77	33.80	1.20	1.43
n = 6	**130**	**174**	**462**			**66.41**

Recall $\hat{Y} = -7.803 + 0.712X_1 + 0.115X_2$

$$F_{3-1,6-3} = \frac{307 - 66.41/(3-1)}{66.41/(6-3)} \cong 5.43$$

OR

ANOVA Table for 3-Variance

Sources of variation	**Sum of squares**	**DF**	**MSS**
ESS	241.987	2	120.295
RSS	66.41	3	22.14
Total	308.397	5	

$$MSS = \frac{\hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2}{2} = \frac{(0.712)(341) + (0.115)(-7)}{2} = 120.295$$

$$ESS = \hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2 = 241.987$$

$$\text{Fcal} = \frac{\frac{\hat{b}_1 \sum y x_1 + \hat{b}_2 \sum y x_2}{2}}{\sum \mu_i^2 / (n-3)} = \frac{120.295}{22.14} = 5.43$$

OR

$$\text{Fcal} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0.783/(3-1)}{(1-0.783)/(6-3)} = 5.41$$

Note: F_{cal}in the three method might not b constant in each due to approximation. However, from our three methods it can be seen that $F_{cal}$is approximately 5.4

4.0 CONCLUSION

In the course of our discussion on multiple regression you have learnt about

- Definition of multiple regression
- Assumptions of multiple regression
- Regression coefficients
- Estimation of Multiple regression equation

5.0 SUMMARY

In our discussion of this unit we were exposed to the assumptions of multiple regression analysis and we estimated the regression line, coefficient of determination and the overall test of significance using three different approaches and we draw a conclusion on the model.

6.0 TUTOR MARKED ASSIGNMENT

A shoe store owner estimated the following regression equation to explain sales as a function of the size of investment in inventories (X_1) and advertising expenditures (X_2). The sample consisted of 10 stores. All variables are measured in thousands of Naira.

$$\hat{Y} = 29.1270 + .5906X_1 + .4980X_2$$

The estimated R^2 was .92448, $\Sigma(Y_i - \bar{Y})^2 = 6,724.125$, and the standarddeviations of the coefficients of X_1 and X_2 obtained from the regression were.0813 and .0567 respectively.

a) Find the sum of squared residuals
b) Can we conclude that sales are dependent to a significant degree onthe size of stores' inventory investments?
c) Can we conclude that advertising expenditures have a significant effecton sales?
d) Can we conclude that the regression has uncovered a significant overallrelationship between the two independent variables and sales?
e) What do we mean by the term 'significant' in b), c) and d) above?

7.0 REFERENCES

Damordar, N. G., Dawn, C. P. and Sangeetha, G. (2012): Basic Econometrics (5th edition) Tata McGraw Hill Education Private Limited. New Delhi, India.

Davidson, James (2000). Econometric Theory, Blackwell Publishers, Oxford, U.K..

Dominick, S. and Derrick, R. (2011): Statistics and Econometrics (Schaum outline) (2nd edition). McGraw Hill, New York.

Dutta, M. (1975) Econometric Methods, South-Western Publishing Company, Cincinnati.

Greene, William H. (2000) Econometric Analysis, 4th ed., Prentice Hall, Engle-wood cliffs, N. J.

Goldberger, Arhur S. (1998) Introductory Econometrics, Harvard University Press.

Hill, Carter, William Griffiths, and George Judge (2001) Undergraduate Econometrics, John Wiley & Sons, New York.

Maddala, G, S, (2001) Introduction to Econometrics, John Wiley & Sons, 3d ed., New York.

Mills, T.C. (1993) The Econometric Modelling of Financial Time Series, Cambridge University Press.

Mukherjee, Chandan, Howard white, and Mare Wuyts (1998) Econometrics and Data Analysis for Developing Countries, Routledge, New York.

Oyesiku O.K. and Omitogun O. (1999): Statistics for Social and Management Science (2nd edition). Higher Education Books Publisher, Lagos.

Theil, Henry (1971) Principles of Econometrics, John Wiley & Sons, New York.

Wonnacott, R.J., and T. H. Wonnacott (1979) Econometrics, 2d ed., John Wiley & Sons, New York.

Wooldridge, Jeffrey M. (2000) Introductory Econometrics, South-Western College Publishing.

UNIT 4: TIME SERIES

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Time Series defined

3.2 Component of time series

3.3 Measurement of Trend

3.4 Worked example

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/ Further Readings

1.0 INTRODUCTION

In all the social sciences, and particularly economics and business, the problem of how condition changes with the passage of time is of utmost importance. For study of such problems, the appropriate kind of statistical information consist of data in the form of time series, figures which shows the magnitude of a phenomenon month after month or year after year. The proper methods for treating such data and thus summarizing the experience which they represent are indispensable part of the practicing statistician equipment.

2.0 OBJECTIVE

At the of this unit, you should be able to

- Understand or define time series
- Understand component part of time series
- Understand methods of estimating time series
- Estimation and graphical representation of the trend

3.0 MAIN CONTENT

3.1 Time Series Defined

Time series refers to sequence of observations that gives information on how data has been behaving in the past. You might wonder why we should spend so much effort constructing series showing what has happened in the past. This is history and should we not rather be looking to the future? As you know the twentieth century is age of planning: government plans the economy for many years ahead; public corporation plan output and investment; most state plan to keep the rate of inflation down to an acceptable level.

Good planning is usually based on information and this is where the time series comes into its own. It provides information about the way in which economic and social variable have been behaving in the recent past, and provides an analysis of that behaviour that planner cannot ignore. Naturally, if we are looking into the future, there is certain assumption we have to make, the most important of which is that the behavioural pattern that we have found in the past could continue into the future. In looking to the future there are certain pattern that we assume will continue and it is to help in the determination of these pattern that we undertake the analysis of the time series.

Time series is usually ordered in time or space. Time series is denoted by sequence (Yt) where Yt is the observed value at time t. Essentially, time series is usually applied to economic and business problems whose purpose of analyses data is to permit a forecast to the future both in the long term and short term. It may be used as an essential aid to planning. Example of time series data are volume of sales, the character and magnitude of its cost of production etc. population figure, price level, demand of a commodity.

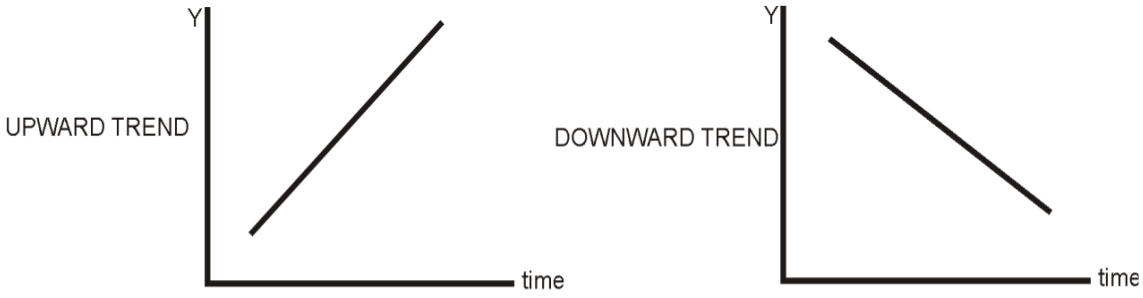
3.2 Components of Time Series

The nature or variation or type of changes in times series can be categorise into:

- Secular trend or long term movement
- Seasonal variation
- Cyclical variation
- Irregular or residual variation

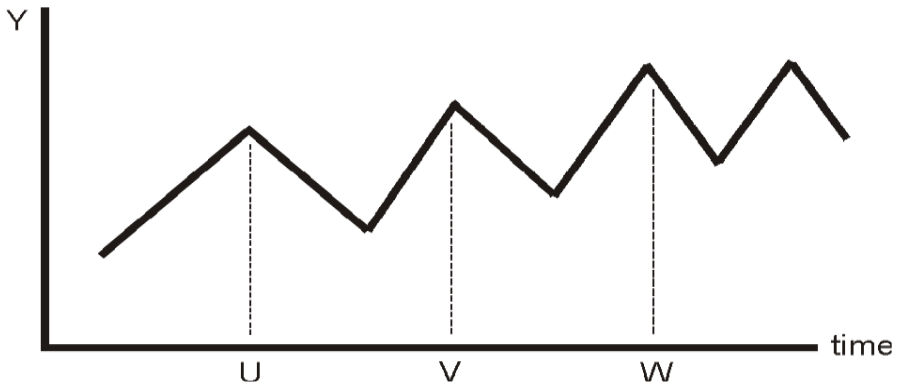
3.2.1 Secular Trend

This refers to the general direction in which the graph of time series appears to be going over a long period of time. This explains the growth or decline of a time series over a long period. Time series is said to contain a trend if the mean or average of series changes systematically with time. The trend could be upward or downward, this could take any of the shape below.



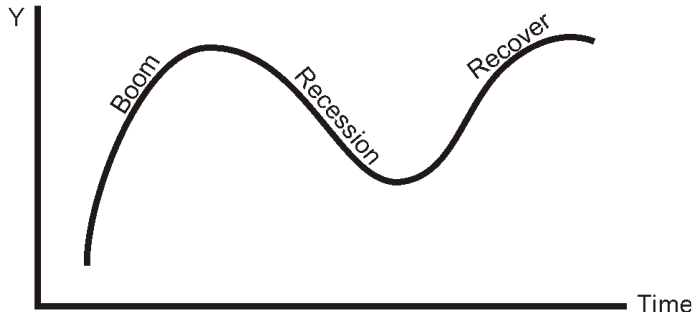
3.2.2 Seasonal Variation

This refers to short term fluctuation or changes that occur at regular intervals less than a year. It is usually brought about by climatic and social factor(s), it is usually because of an event occurring at a particular period of the year. Examples of these are sale of card during valentine period, sale of chicken during x-mas, new year or any festive period(s).



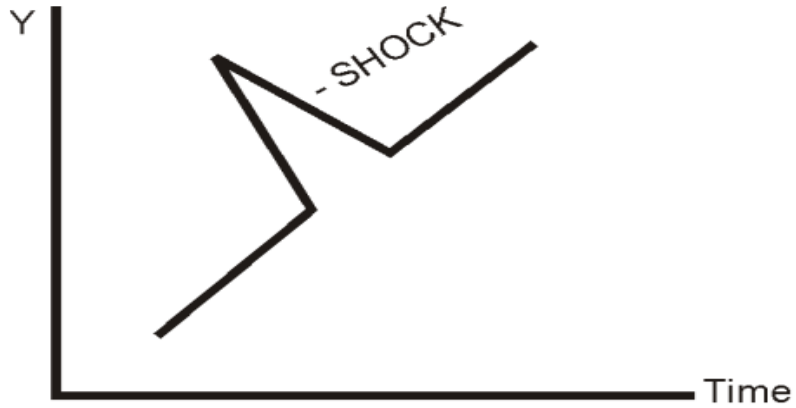
3.2.3 Cyclical Variation

This refers to long term variations about the trend usually caused by disruption in services or socio-economic activities, cyclical variations are commonly associated with economic cycles, successive boom and slumps in the economy. A good example of this is business cycle.



3.2.4 Irregular Variation

This refers to time series movement that are not definite this is usually caused by unusual or unexpected and unpredictable events such as strike, war, flood, disasters. Here, there's no definite behavioural pattern.



SELF ASSESSMENT EXERCISE

Discuss the components of time series?

3.3 Measurement of Trend

Basically trend values of a time series can be estimated by any of the following methods:

- Free hand
- Least square method
- Moving average and
- Semi average method

3.3.1 **Free Hand Method**

This method involves the drawing a scattered diagram of the values with time as the independent variable on the x-axis and then drawing the trend line by eye. This method is condemned because it is subjective and inaccurate method of obtaining a Trend line.

3.3.2 **Least Square Method**

This method is a statistical technique usually used in calculating the line of best fit or line of goodness that measures the goodness of fit of the curve, this is usually independent of human judgments, it makes an assumption that the trend line is a straight one. The least square formular is given as;

Y = a + bx + e

Where a = intercept

b = slope of the curve

e = error term

Illustration

Given the 7weeks information below about the sales of a company

Wk	**Sales**
1	15
2	25
3	38
4	32
5	40
6	37
7	50

Estimate the regression line $Y = a + \hat{b}x$ and forecast the sales of the 10th and 12th week.

Solution

Let X represents the weeks

Y represent the sales value

Least Square Method Table of Analysis

X	**Y**	**XY**	**X^2**
1	15	15	1
2	25	50	4
3	38	114	9
4	32	128	16
5	40	200	25
6	37	222	36
7	50	350	49
ΣX= 28	**ΣY=237**	**ΣXY=1079**	**ΣX^2=140**

$$\widehat{\boldsymbol{b}} = \frac{7(1079) - 28\ (237}{7\ (140) - 28^2} = 4.67857$$

$$\boldsymbol{a} = \overline{\boldsymbol{Y}} - \widehat{\boldsymbol{b}}\overline{\boldsymbol{X}} = 33.857 - (4.67857)4 = 15.1427$$

The trend equation will be:

$$Y = 15.1427 + 4.6785x$$

This trend equation can be used in forecasting into future sales of the company, for example future sales value for the 10th and 12th week can be known by simply substituting the week's value into the trend equation.

i.e. for the 10th week we have;

$$Y = 15.1427 + 4.6785(10)$$

$$Y = 15.1427 + 46.785$$

$$Y = 61 - 9277$$

For the 12th week

$$Y = 15.1427 + 4.6785(12)$$

$$Y = 15.1427 + 56.142$$

$$Y = 71.2847$$

$$Y = 71$$

$$Y \cong 62$$

3.3.3 Moving Average Method

A moving average is a simple arithmetic mean. We select a group of figures at the start of the series e.g. 3,4,5,7 and average them to obtain our first trend figure. Then you drop the first figure and include the next item in the series to obtain a new group. The average of this group gives the second trend figure. You continue to do this until all figures in the series is exhausted.

There is no doubt that the trend eliminates the large scale fluctuations found in the original series moving average smoothing is a smoothing technique used to make the long-term trend of a time series cleared.

Illustration

The table below contained information about the actual sales of a company, Prepare a 3 month moving average forecast.

Month	**Sale (units)**
Jan	35
Feb	34
Mar	36
April	31
May	28
June	30
July	27
August	26
Sept	31
Oct	35
Nov	37
Dec	39

Solution

Months	**Sales**	**3months Moving total**	**3months moving average trend**
Jan	35		
Feb	34		
Mar	36	105	35
April	31	101	33.7
May	28	95	31.7
June	30	89	29.7
July	27	85	28.3
Aug	26	83	27.7
Sept	31	84	28
Oct	35	92	30.7
Nov	37	103	34.3
Dec	39	111	37

Column 1 on the table represents the months

Column 2 represents the sale's figure

Column 3 is arrived at by adding the sales figure in 3s i.e

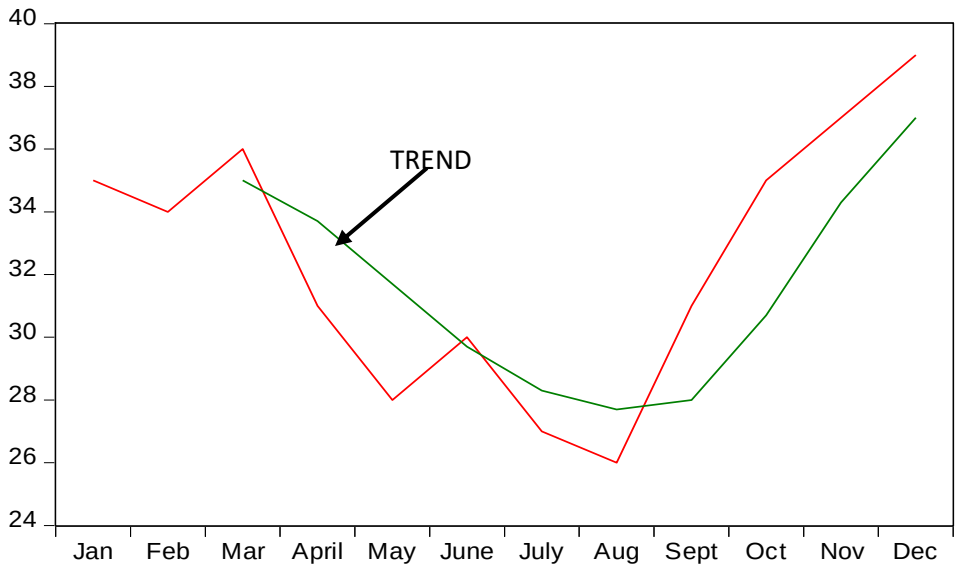
Jan + Feb + Mar = 1050

Feb + Mar + April = 1010

Mar + April + May = 950

Column 4 is arrived at by dividing the column 3 by the n which happen to be the moving average. This Rs called the trend.

GRAPHICAL REPRESENTATION OF MOVING AVERAGE TREND



3.3.4 Semi Moving Average Method

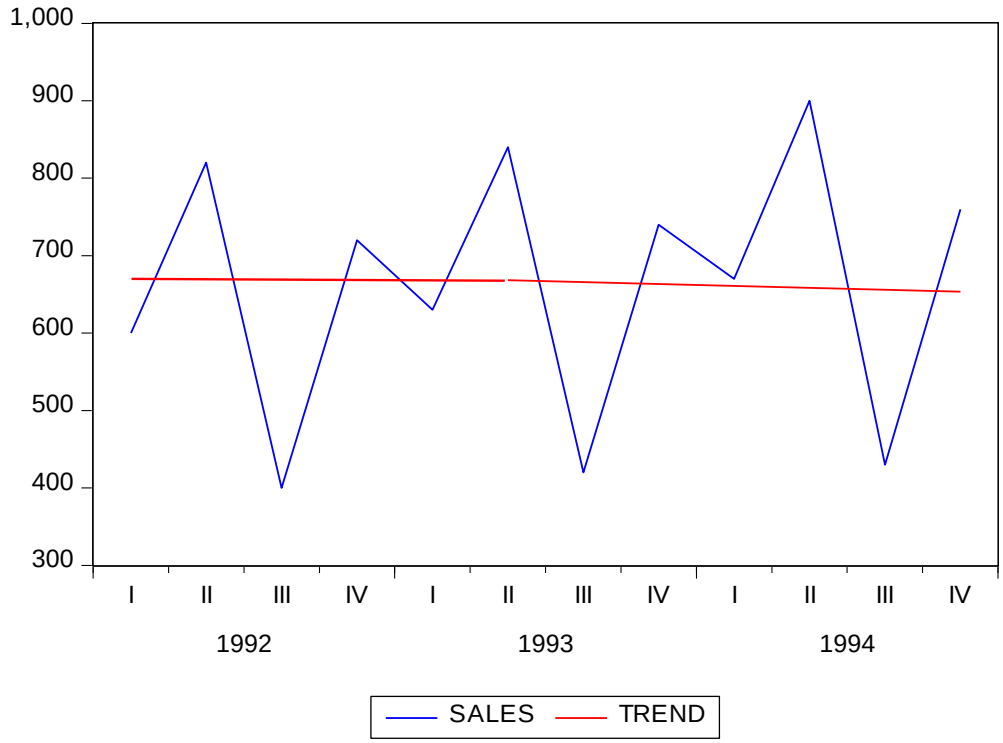
This method is usually used to estimate trends by separating or dividing that data into two equal parts and averaging the data each part, thus, obtaining two points on the graph of time series. A trend is then drawn between these two points and trend value can be determined. If the number of years is odd, the middle year is deleted and the group can then be divided into two equal parts.

Illustration

Semi- Moving Average Method Table of Analysis

Years	**Quarter**	**Y sales**	**X**	**Semi Average Total**	**Semi average method trend**
1992	1	600	- 6		
	2	820	- 5		
	3	400	- 4		
	4	720	- 3	4010	668.33
1993	1	630	- 2		
	2	840	- 1		
	3	420	1		
	4	740	2		
1994	1	670	3		

	2 3 4	900 430 760	4 5 6	 3920	653.3



4.0 CONCLUSION

In the course of our discussion on estimation of time series, you have learnt about least square method, moving average method and semi average method.

5.0 SUMMARY

The least square trend equation is written as\

6.0 TUTOR MARKED ASSIGNMENT

Table Showing the Number of Prescriptions Dispensed by a Chemist

Year	**Quarters**			
	1	**2**	**3**	**4**
2000	-	-	60	71
2001	69	67	62	69

2002	73	66	62	68
2003	72	66	65	67
2004	75	-	-	-

Prepare a 4-point moving average of the above information?

7.0 REFERENCES/FURTHER READINGS

Adedayo, O. A. (2006): Understanding Statistics. JAS Publishers, Akoka Lagos.

Dawodu, A.F. (2008): Modern business Statistics 1. NICHO Printing Works. Agbor, Delta State.

Esan, E.O. and Okafor, R.O. (2010): Basic Statistical Method. Tony Christo Concept, Lagos.

Olufolabo, O.O. &Talabi, C.O. (2002): Principles and Practice of Statistics HAS-FEM (NIG) Enterprises Somolu Lagos.

Owen F. and Jones, R. (1978): Statistics. Polytech Publishers Ltd, Stockport.

Oyesiku, O.K. and Omitogun, O. (1999): Statistics for social and Management Sciences. Higher Education Books Publisher, Lagos