NATIONAL OPEN UNIVERSITY OF NIGERIA

FACULTY Of SOCIAL SCIENCES

COURSE CODE: ECO 452

COURSE TITLE: APPLIED STATISTICS

NOUN
NATIONAL OPEN UNIVERSITY OF NIGERIA
APPLIED STATISTICS
ECO 452

FACULTY OF SOCIAL SCIENCES

COURSE GUIDE

Course Developers:
Dr. Adesina- Uthman Ganiyat Adejoke
Department of Economics, Faculty of Social Sciences,
National Open University of Nigeria.
&
Ogunjirin Olakunle
Yaba College of Technology
School of Liberal Studies, Department of Social Sciences.

Course Editor:
Dr. Ogunsakin Sanya
Department of Economics
Senior Lecturer, Ekiti State University, Ado-Ekiti.

Course Reviewer:
Obumneke Ezie, *Ph.D.*
Department of Economics, Bingham University, Karu, Nasarawa State.
eobumneke@yahoo.com

NATIONAL OPEN UNIVERSITY OF NIGERIA
National Open University of NigeriaHeadquarters
14/16 Ahmadu Bello WayVictoria Island
Lagos Abuja Annex
245 Samuel Adesujo Ademulegun Street
Central Business District
Opposite Arewa Suites
Abuja
e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

First Printed
ISBN:

For
National Open University of Nigeria *Multimedia Technology in Teaching and Learning*

CONTENT

Introduction

The course advanced statistics (ECO 452) is a first semester course which carries two credit units for fourth year level economics students in the School of Art and Social Sciences at the National Open University, Nigeria. The course is a very useful course to you in your academic pursuit, because it helps gain in-depth insight of the underlining statistical tools usually used by economists.

This course guide tells you what advanced statistics entails, what course materials you will be using and how you can work your way through these materials. It suggests some general guidelines for the amount of time required of you on each unit in order to achieve the course aims and objectives successfully. It also provides you some guidance on your tutor marked assignments (TMAs) as contained herein.

What you will learn in this Course

The course is made up of 14 units, covering areas such as:

- Sampling distribution defined
- Sampling distribution of proportion
- Sampling distribution of difference and sum of two means
- Probability distribution
- One-way factor analysis of variance
- Two-way factor analysis of variance
- Analysis of covariance
- Estimation of multiple regressions
- Partial correlation coefficient
- Multiple correlation coefficient and coefficient of determination
- Overall test of significance
- Time series and its components
- Quantitative estimation of time series
- Price index

Course Aims

The overall aims of this course include:

i. To introduce you to sampling methods
ii. Expose you to Probability distribution
iii. Teach you one-way factor analysis
iv. Expose you to two-way factor analysis
v. Introduce you to multiple regression
vi. Introduce you to time series analysis
vii. Expose you to index numbers

Course Objectives

There are 14 study units in the course and each unit has its own objectives. You should read the objectives of each unit and assimilate them. In addition to the objectives of each unit, the main objective of the course is to equip you with adequate information on the applied statistics and to enable you acquire enough professional competence to apply such knowledge to current theories and ways of conducting analysis in economics.

The objectives of the course will be achieved by:

- Analysing the Sampling distribution
- Discussing Sampling distribution of proportion
- Sampling distribution of difference and sum of two means
- Discussing Probability distribution
- Identifying One-way factor analysis of variance
- Discussing Two-way factor analysis of variance
- Analysing the Analysis of covariance
- Estimation of multiple regressions
- Discussing Partial correlation coefficient
- Examining Multiple correlation coefficient and coefficient of determination
- Discussing Overall test of significance
- Time series and its components
- Discussing the Quantitative estimation of time series
- Analyzing Price index

Working through the Course

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

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

Course Material

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

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

Study Units

There are four modules of 14 units in this course, which should be studied carefully.

MODULE ONE: Statistical Inference
Unit 1: Sampling distribution defined
Unit 2: Sampling distribution of proportion
Unit 3: Sampling distribution of difference and sum of two means
Unit 4: Probability distribution

MODULE TWO: Analysis of variance and analysis of covariance

Unit 1: One-way factor analysis of variance
Unit 2: Two-way factor analysis of variance
Unit 3: Analysis of covariance

MODULE 3: Multiple Regression Analysis
Unit 1: Estimation of multiple regressions
Unit 2: Partial correlation coefficient
Unit 3: Multiple correlation coefficient and coefficient of determination
Unit 4: Overall test of significance

MODULE 4: Time series analysis
Unit 1: Time series and its components
Unit 2: Quantitative estimation of time series
Unit 3: Price index

References and Other Resources

Every unit contains a list of references and further reading. Try to get as many as possible of those textbooks and materials listed. The textbooks and materials are meant to deepen your knowledge of the course.

Assignment File

In this file, you will find all the details of the work you must submit to your tutor for marking. The marks you obtain from these assignments will count towards the final mark you obtain for this course. Further information on assignments will be found in the Assignment File itself and later in this *Course Guide* in the section on assessment

Presentation Schedule

The Presentation Schedule included in your course materials gives you the important dates for the completion of tutor-marked assignments and attending tutorials. Remember, you are required to submit all your assignments by the due date. You should guard 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 Assignments (TMAs)

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

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

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

Final Examination and Grading

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

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

Course Marking Scheme

The table presented below indicate 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 indicate the units, number of weeks and assignments to be taken by you to successfully complete the course, applied statistics (ECO 452).

Unit	Title of Work	Weekly Activity	Assessment End of Unit
	Course Guide	1	
1	Analysing the Sampling distribution		
2	Discussing Sampling distribution of proportion		
3	Sampling distribution of difference and sum of two means		
4	Discussing Probability distribution		1st Assignment
5	Identifying One-way factor analysis of variance		
6	Discussing Two-way factor analysis of variance		
7	Analysing the Analysis of covariance		
8	Estimation of multiple regressions		2nd Assignment
9	Discussing Partial correlation coefficient		
10	Examining Multiple correlation coefficient and coefficient of determination		
11	Discussing Overall test of significance		3rd Assignment
12	Time series and its components		
13	Discussing the Quantitative estimation of time series		
14	Analyzing Price index		4th Assignment

How to Get the Most from This Course

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

Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your books or other material, and when to embark on discussion with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provides exercises for you to do at appropriate points.

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

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

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

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

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

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

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

   Whatever method you choose to use, you should decide on and write in your own dates for working breach unit.
3. Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they get behind with their course work. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.
4. Turn to Unit 1 and read the introduction and the objectives for the unit.
5. Assemble the study materials. Information about what you need for a unit is given in the 'Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your set books on your desk at the same time.
6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.

7. Up-to-date course information will be continuously delivered to you at the study centre.
8. Work before the relevant due date (about 4 weeks before due dates), get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments no later than the due date.
9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.
10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.
11. When you have submitted an assignment to your tutor for marking do not wait for it return 'before starting on the next units. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, both on the tutor-marked assignment form and also written on the assignment. Consult your tutor as soon as possible if you have any questions or problems.
12. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in this Course Guide).

Tutors and Tutorials

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

Your tutor will mark and comment on your assignments, keep a close watch on your progress and on any difficulties you might encounter, and provide assistance to you

during the course. You must mail your tutor-marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

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

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

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

Conclusion

On successful completion of the course, you would have developed critical thinking skills with the material necessary for efficient and effective use of statistical tools economics. However, to gain a lot from the course please try to apply anything you must have learnt in the course to practice by doing the calculation on paper yourself. We wish you success with the course and hope that you will find it both interesting and useful.

MODULE ONE: Statistical Inference
Unit 1: Sampling distribution defined
Unit 2: Sampling distribution of proportion
Unit 3: Sampling distribution of difference and sum of two means
Unit 4: Probability distribution

MODULE TWO: Analysis of variance and analysis of covariance
Unit 1: One-way factor analysis of variance
Unit 2: Two-way factor analysis of variance
Unit 3: Analysis of covariance

MODULE 3: Multiple Regression Analysis
Unit 1: Estimation of multiple regressions
Unit 2: Partial correlation coefficient
Unit 3: Multiple correlation coefficient and coefficient of determination
Unit 4: Overall test of significance

MODULE 4: Time series analysis
Unit 1: Time series and its components
Unit 2: Quantitative estimation of time series
Unit 3: Price index

MODULE ONE: Statistical Inference

Unit 1: Sampling distribution

Unit 2: Sampling distribution of proportion

Unit 3: Sampling distribution of difference and sum of two means

Unit 4: Probability distribution

UNIT ONE: SAMPLING DISTRIBUTION

1.1 Introduction

1.2

Generally statistical data are studied in order to learn something about the broader field which the data represents. In order to make statistical work meaningful, statistician generalize from what we find in the figure at hand to the wider phenomenon which they represent. In technical language we regard a set of data as a sample drawn from a larger "universe". We analyze the data of the sample in order to draw conclusion about the corresponding universe or population.

In a sense universe actually exists and it is theoretically possible to study the universe completely. But in another sense the universe is broader and in a sense less tangible.

This unit happens to be one of the four units in this module, for proper understanding of the topics in this unit a thorough knowledge of elementary statistics is required.

1.2 Learning Outcomes

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

- Discuss the meaning of Sample
- Evaluate the meaning of Population
- The meaning of Sampling theory
- Discuss the assumptions of normal distribution
- Evaluate parameter estimation
- Estimate sample mean, population mean etc.

1.3 Sampling Theory, Population and Sample Defined

Statistical inference is defined as the process by which on the basis of sample we draw conclusion about the universe from which sample is drawn. It can as well be defined as a process by which conclusion are drawn about some measure or attribute of a population based upon analysis of sample. Samples are taken and analyzed in order to draw conclusion about the whole population.

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.

In general, a study of the inference made concerning a population by using sample drawn

from it together with indication of accuracy of such inferences by using probability theory is called statistical inference. Population of a variable X is usually defined to consist of all the conceptually possible values that the variable may assume. Some of these values may have already been observed, others may not have occurred, but their occurrence is conceivably possible. The number of conceptually possible values of a variable is called size of the population. This size varies according to the phenomenon being investigated.

A population may 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.

Sample refers to a collection of observation on a certain variable. The number of observations included in the sample is called the size of the sample.

The main object of the theory of statistics 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 is various characteristics values that describes 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 populations statistician use the term parameter for the basic descriptive

measure of population while statistics is usually used for the basic descriptive measure of a sample.

Table M1.1.1: Basic Descriptive Measure of Population and Sample

	Population parameters	Symbol	Sample statistics	Symbol
I	Population mean	μ	Sample mean	$\bar{X}$
Ii	Population variance	σ_X^2	Sample variance	S_{X^2}
iii	Population standard deviation	σ_X	Sample standard deviation	S_X

Note: $E(X) = \mu = \frac{X_1 + X_2 + ...X_n}{n}$

Self-Assessment Exercise 1

What are descriptive measures that can be used in describing a sample or population?

1.4 Sampling Distribution of Parameter & Sample Estimates

The population mean is usually referred to as the expected value of the population and it is conventionally denoted as E(x) or μ. But for a discrete random variable the expected value is computed by the sum of the product of value of X_1 multiplied by their various probabilities.

$$E(X) = \mu = \sum_{i=1}^{n} Xf(X_1)$$

Where X_i is the probability of variable x.

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.

$$Var(x) = \sigma_X^2 = \frac{\sum (X - E(X))^2}{n} = \frac{\sum (X - \mu)^2}{n}$$

Where E (x) = population 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. This is denoted as:

$$\sigma_X = \frac{\sqrt{\sum (X - E(X))^2}}{n} = \frac{\sqrt{\sum (X - \mu)^2}}{n}$$

The standard deviation is a measure that describes how dispersed the values of x is around the population mean.

$$COV(XY) = \Sigma(XY) - \Sigma X \Sigma Y$$

Worked Example

Given the population 11, 12, 13, 14, 15 calculate the mean, standard deviation, and the variance of the given population.

Table M1.1.2: Table of Analysis for Sample Mean, Standard Deviation and Variance

X	$X - \mu$ X-E(X)	$(X - \mu)^2$ $(X-E(X))^2$
11	1 1 - 13 = 2	4
12	12 - 13 = 1	1
13	13 - 13 = 0	0
14	14 - 13 = 1	1
15	15 - 13 = 2	4
n = 5		10

$$\bar{X} = \mu \frac{11+12+13+14+15}{5} = \frac{65}{5} = 13$$

$$Var(x) = \sigma_X^2 = \frac{\sum (X - E(X))^2}{n} = \frac{\sum (X-\mu)^2}{n} = \frac{10}{5} = 2$$

$$\sigma_X = \sqrt{2} = 1.4142$$

Self-Assessment Exercise 2

Define standard deviation of a population

1.5 Estimation of Sample Statistics

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 Sx2 and sample standard deviation denoted by Sx.

Sample mean is defined as the average value in the sample it is denoted by $\bar{x}$. The sample arithmetic mean is calculated by adding up the observation of the sample and then dividing by the total number of observations.

$$\bar{X} = \frac{\sum_{i=0}^{n} X}{n}$$

Sample variance as it has been said before now, it is a measure of dispersion of the value of x in the sample around their average value. This is denoted as

$$S_{X^2} = \frac{\sum_{i=0}^{n} (X - \bar{X})^2}{n} = \frac{\sum X^2 - n\bar{X}^2}{n} = \frac{\sum X^2 - \bar{X}^2}{n}$$

The sample standard deviation is denoted by S_x this is taken to be the square root of the

variance.

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

Covariance; this statistics usually involves two variable. The covariance is defined as the sum of the product of the deviation of variable x and y from the various means.

$$COV(XY) = \frac{\sum_{i=1}^{n} (X - \bar{X})(Y - \bar{Y})}{n}$$

Question

From the information of population supplied in the preceding subsection i.e. 11, 12, 13, 14, 15

Table M1.1.3: Sample Statistics Table of Analysis

Possible samples	$\bar{X}$ = mean of each sample
(11,12)	$\frac{11+12}{2} = 11.5$
(11,13)	$\frac{11+13}{2} = 12$
(11,14)	$\frac{11+14}{2} = 12.5$
(11,15)	$\frac{11+15}{2} = 13$
(12,13)	$\frac{11+13}{2} = 12.5$
(12,14)	$\frac{11+14}{2} = 13$
(12, 15)	$\frac{11+15}{2} = 13.5$
(13,14)	$\frac{11+14}{2} = 13.5$
(13,15)	$\frac{11+15}{2} = 14$
(14,15)	$\frac{14+15}{2} = 14.5$

n = 10	

Sample mean = $\frac{11.5+12+12.5+13+12.5+13+13.5+13.5+14.+14.5}{10}=\frac{130}{10}=13$

All the information about the population and possible samples can be summarize in a frequency distribution as depicted in table below.

Table M1.1.14: Table of Possible Samples

X	F
11	1
11.5	1
12	1
12.5	2
13	2
13.5	2
14	1
14.5	1
15	1

Variance of sample mean $=\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{N}$

$$S_{X^2}=\frac{(11-13)^2+(11.5-13)^2+(12-13)^2+2(12.5-13)^2+2(13-13)^2+2(13.5-13)^2+(14-13)^2+(14.5-13)^2+(15-13)^2}{9}$$

$$S_{X^2}=\frac{(-2)^2+(1.5)^2+(-1)^2+2(-0.5)^2+2(0)^2+2(0.5)^2+(1)^2+(2)^2+(1.5)^2}{9}$$

$$S_{X^2}=\frac{4+2.25+1+0.5+0+0.5+1+4+2.25}{9}$$

$$S_{X^2}=\frac{15.5}{9}$$

$$S_{X^2}=1.722$$

$$S_X=\sqrt{1.722}$$

$$S_X=1.31233$$

From the foregoing analysis it would be observed that given X_1, X_2 ……X_n of any random sample of size n from any infinite population with population mean u and σ^2 then with sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$ we have

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

(ii) $Var(x) = \frac{\sigma^2}{n}$

Self-Assessment Exercise 3

Define the sample variance of any given population

1.5.1 Estimators for Mean and Variance

Given that X_1, X_2, X_3 …. X_n is a random sample of size n from normal population with mean μ and variance σ^2 i.e. $(X \sim N(\mu, \sigma^2))$, then the statistics $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X$

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

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

Worked Example

Given a random sample of 20 taken from a normal distribution with mean 90 and variance 25 find the probability that the mean is greater than 101.

Solution

$\bar{X} \sim 20(90, 25)$

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

$$Z = \frac{101-90}{\frac{25}{\sqrt{20}}} = \frac{11}{\frac{25}{4.47213}} = \frac{11}{5.590169} \cong 1.968$$

1.5.2 Assumptions Or Properties Of A Normal Distribution

A normal distribution, also known as a Gaussian distribution, is a type of continuous probability distribution for a real-valued random variable. It's one of the most important probability distributions in statistics because it fits many natural phenomena, such as heights, blood pressure, measurement errors, and IQ scores.

The normal distribution is defined by two parameters: the mean (μ) and the standard deviation (σ). The mean determines the center of the distribution, and the standard deviation the width of the distribution.

Here are the main assumptions or properties of a normal distribution:

1. **Symmetry**: A normal distribution is symmetric about the mean. This means that the left and right halves of the distribution are mirror images of each other.
2. **Mean, Median, Mode Co-incidence**: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.
3. **Bell-Shaped Curve**: The distribution has a bell shape, with the highest frequency occurring at the mean, and frequencies tapering off equally on either side.

4. **Asymptotic**: The curve is asymptotic to the x-axis, which means it approaches but never quite touches the x-axis.
5. **Empirical Rule (68-95-99.7 rule)**: This rule states that in a normal distribution, about 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and about 99.7% falls within three standard deviations.
6. **Dependence on Mean and Standard Deviation**: The entire form of the normal distribution depends on the mean and the standard deviation. The mean locates the center of the distribution, and the standard deviation determines the height and width of the distribution.

These assumptions make the normal distribution extremely tractable in mathematical and statistical analyses. Many statistical tests and methods are based on the assumption of normality, so understanding these properties is essential in statistics.

Self-Assessment Exercise 4

What are the assumptions of a normal distribution?

1.5.3 The Role and Significant of Statistics in Social Sciences

It is interesting to know that accuracy, validity, reliability, objectivity, analysis, efficiency are all characteristics of the roles expected of statistical research in decision making and policy formulation for societal development. Do you know that social statistics are

necessary in information gathering about socio-economics variables that are indices of economic growth and development? It started with what is known as the "statists" social research" and later grow to be known as "statistics", a new term for quantitative evidence. Social sciences' statistics is very significant because it assist in quantifying scientific developments and data on them therefore, making information on scientific studies more concise and precise. Social statistics is usually conducted to prove something for instance, how many women are affected by malaria compare to men in the society? How many people in the society are able to afford living in a duplex, flat, one-room apartment, face-to-face room or under the bridge? Consequently, it is significant to note that adequate cautions are usually put into stepwise data gathering, accuracy, and analysis for efficiency. The role of statistics in social sciences and its significant cannot be overemphasized.

Self-Assessment Question 5

Do you think that statistics in social sciences has role to play in societal problem solving?

1.6 Summary

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

It has been established that given a random sample of X_1, X_2, …. X_n with population

mean μ and standard variance r^2

(i) $\Sigma\bar{X} = \mu$

(ii) $Var(x) = \frac{\sigma^2}{n}$

Tutor Marked Assignment

Explain the descriptive measure of a sample statistics.

1.7 References/ Further Readings

Adedayo, O.A. (2006): Understanding statistics. JAS publishers Akoka, Yaba.

Dominick, S. and Derrick P. (2011): (Schaum outline series) Statistics and Econometrics (second edition) MCGRAW HILL, New York.

Edward, E.L. (1983): Methods of statistical analysis in Economics and Business. HOUGHTON MIFFLIN COMPANY BOSTON.

Esan F.O. and Okafor, R.O. (2010): Basis statistical method (revised edition) Toniichristo Concept Lagos.

Koutsoyianis, A. (2003): Theory of Econometrics (second edition). Palgrave publishers Ltd (formerly Macmillan press Ltd), London and Basic stoke.

Murray R. S. and Larry J. S. (1998): (Schaum outlines series). Statistics (Third edition) MCGRAW HILLS.

Olufolabo, O.O. and Talabi, C.O (2002): Principles and practice of statistics. HASFEM Nig Enterprises, Shomolu, Lagos.

Oyesiku, O.K. and Omitogun, O. (1999). Statistics for social and management sciences. Higher Education Books Publisher Lagos.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

The descriptive measures that can be used in describing a sample or population are mean,

variance and standard deviations.

Answer to Self- Assessment 2

The standard deviation is a measure of the amount of variation or dispersion of a set of values. When it comes to a population, the standard deviation is a measure that quantifies the extent to which individual data points in the population vary from the population mean.

Answer to Self- Assessment 3

The sample variance is a statistic that describes the dispersion or spread of a set of data points within a sample from a larger population. It's an estimate of the population variance, based on the data available in the sample.

Answer to Self- Assessment 4

The normal distribution is a continuous, symmetric, bell-shaped distribution for a real-valued random variable, characterized by its mean (center of the distribution) and standard deviation (spread or width of the distribution). Key assumptions include the equal location of the mean, median, and mode; its asymptotic nature to the x-axis; and the empirical rule stating that about 68%, 95%, and 99.7% of data falls within one, two, and three standard deviations from the mean, respectively.

Answer to Self- Assessment 5

Yes, statistics in the social sciences are not just about crunching numbers. They're a crucial tool for understanding and addressing the complex problems that societies face.

1. **Understanding Trends**: Statistics help us understand trends and patterns in society, such as changes in crime rates, employment, education outcomes, and health behaviors.
2. **Informing Policy**: By analyzing data, policymakers can make informed decisions about how to address societal problems. For example, data on poverty rates can guide the development of policies to assist low-income individuals and families.
3. **Evaluating Programs and Interventions**: Statistics can be used to evaluate the effectiveness of social programs and interventions. For example, a randomized controlled trial can determine whether a new educational program improves student outcomes.
4. **Identifying Inequalities**: Statistical analysis can reveal disparities and inequalities in society, such as racial or gender disparities in income or health outcomes. These insights can then inform efforts to promote equality.
5. **Forecasting and Planning**: Statistics can help forecast future societal trends, enabling better planning and resource allocation. For example, population projections can inform planning for healthcare, infrastructure, and other services.
6. **Testing Hypotheses**: Statistics provides tools to test hypotheses and theories about society, contributing to our understanding of human behavior and societal structures.

UNIT TWO: SAMPLING DISTRIBUTION OF PROPORTION

1.1 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.

1.2 Learning Outcomes

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
- Standard error

1.3 Sampling Distribution of Proportion Defined

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 characteristic. This sampling distribution is like 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

$$\text{var}(p) = \text{var}(p) = \frac{p(1-p)}{n} = \frac{pq}{n}$$

It is to be noted at this juncture that the sample proportion is also an unbiased estimator of the population proportion i.e. Σ (p) = P

Example

A coin is tossed 120 times, find the probability that head will appear between 45% and 55%.

Solution

From the above the prob(head) = 1/2 = p

Prob(not obtaining ahead) = 1/2= q = 1 - p

45% of tosses = $\frac{45}{100} \times 120 = 54$

While 55% of tosses gives = $\frac{55}{100} \times 120 = 66$

Mean $\mu p = np = 120 \times 1/2 = 60$

$$SD = \sqrt{npq} = \sqrt{\frac{0.25}{120}} = 0.04564$$

$$SD = \sqrt{npq} = \sqrt{120\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = \sqrt{30} = 5.477225575$$

$$prob(54 < p < 78) = p\left(\frac{54-60}{5.5} < z < \frac{66-60}{5.5}\right) = p\left(\frac{6}{5.5} < z < \frac{6}{5.5}\right)$$

$$= p\left(-1.0909 < z < 1.091\right) = (0.3621) \times 2 = 0.7242$$

Self-Assessment Exercise 1

What is the symbolic definition of sampling distribution of proportion?

1.3.1 Standard Error

Standard error usually represented by S.E. is defined as the square root of the population variance written as $\sqrt{\text{var}(p)}$

note $\text{var}(p) = \frac{pq}{n} = \frac{p(1-p)}{n}$

$\therefore \sqrt{\frac{p(1-p)}{n}}$

From the example in subsection 3.2 above

p = ½ = q
n = 120

$$\therefore SE = \sqrt{\frac{0.5(0.5)}{120}} = \sqrt{\frac{0.25}{120}} = \sqrt{0.0020833} = 0.0456$$

Self-Assessment Exercise 2

What does S.E stands for?

1.5 Sampling Distribution of Parameter Estimate

The sampling distribution of a parameter estimate is a theoretical distribution that would result from the infinite number of samples drawn from the same population. It gives us

valuable insights into the variability, precision, and accuracy of our parameter estimates, which are fundamental to hypothesis testing, constructing confidence intervals, and performing various other statistical inference procedures. In statistics, parameter estimation involves estimating the parameters of the population distribution from a sample. The parameter estimate from one sample is just a single point estimate, but if we were to take multiple samples, we would end up with multiple estimates, creating a distribution of parameter estimates, known as the sampling distribution.

To understand this concept, let's begin by defining a few terms:

- **Parameter**: A numerical characteristic of the population. For example, the population mean (mu) or population standard deviation (sigma).
- **Estimate**: An approximation of the parameter calculated from the sample. For example, the sample mean (x-bar) is an estimate of the population mean.
- **Sampling Distribution**: The probability distribution of a given statistic based on a random sample.
- **Standard Error**: The standard deviation of the sampling distribution.

Now, let's delve deeper into the concept of the sampling distribution of parameter estimate:

1. **Estimation of Parameters**: The parameters like population mean, population variance, etc., are usually unknown and need to be estimated. We use the corresponding statistics from sample data to estimate these parameters. The

sample mean is an unbiased estimator of the population mean, and the sample variance (dividing by n-1 instead of n) is an unbiased estimator of the population variance.

2. **Multiple Samples**: Suppose we collect multiple samples from the population and calculate the desired parameter (say mean) from each sample. Each of these means will likely be different because of the inherent randomness in the sampling process.
3. **Formation of Sampling Distribution**: If we plot the frequency of these sample means, we get what's known as a sampling distribution of the sample mean. The Central Limit Theorem (CLT) tells us that this distribution will approach a normal distribution as the sample size gets larger, regardless of the shape of the population distribution, provided the population has a finite standard deviation.
4. **Properties of Sampling Distribution**: The mean of the sampling distribution (also known as the expected value of the estimator) will be equal to the population mean. This property is called unbiasedness. The standard deviation of this sampling distribution is known as the standard error. It quantifies the variability in the parameter estimate across samples. As the sample size increases, the standard error decreases, indicating that larger samples provide more accurate and consistent estimates.
5. **Confidence Intervals**: The sampling distribution of a parameter estimate also allows us to compute confidence intervals for the estimate. A 95% confidence

interval gives a range of values that includes the true population parameter 95% of the time.

1.6 Summary

During the course of our discussion of this unit we have talked about; Sampling distribution of proportion and Standard error. In the course of our discussion, we defined the mean of a sampling distribution of proportion as np.

i.e. mean = np

$$\text{var}(p) = \frac{pq}{n} = \frac{p(1-p)}{n}$$

$$SD = \sqrt{npq}$$

Tutor Marked Assignment

A coin is tossed 90 times, find the probability that tail will appear between 35% and 55%.

1.7 Reference/Further Reading

Adedayo, O.A. (2006): Understanding statistics. JAS publishers Akoka Yaba.

Esan, F.O. and Okafor, R.O. (2010): Basis statistical method (revised edition) Toniichristo Concept, Lagos.

Murray, R.S. and Larry, J. S. (1998): (Schaum outlines series). Statistics (Third edition) MCGRAW HILLS.

Olufolabo, O.O. and Talabi, C.O. (2002): Principles and practice of statistics. HASFEM Nig Enterprises Shomolu Lagos.

Oyesiku, O.K. and Omitogun, O. (1999): Statistics for social and management sciences. Higher Education Books Publisher Lagos.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

The sampling distribution of proportion is said to have a mean of np, and variance

$$\text{var}(p) = \text{var}(p) = \frac{p(1-p)}{n} = \frac{pq}{n}$$

Answer to Self- Assessment 2

Standard error usually represented by S.E. is defined as the square root of the population variance written as $\sqrt{\text{var}(p)}$

UNIT THREE: SAMPLING DISTRIBUTION OF SUM AND DIFFERENCE OF TWO MEANS

1.1 Introduction

This unit is an extension of unit one and unit two of this module. In this unit we are going to look at sampling distribution of sum and difference of two means. 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.

1.2 Learning Outcomes

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

- Sampling distribution of sum of two means
- Sampling distribution of difference and

1.3 Sampling Distribution of Difference of Two Means and Sum

If two independent random sample of sizes n_1 and n_2 are selected from 2 different population of size N_1 and N_2 with population means p_1 and p_2 respectively and population

variance σ_1^2 and σ_2^2 respectively, then the sampling distribution of the difference of two means $(\bar{X}_1 - \bar{X}_2) = \mu_{p1} - \mu_{p2}$

$$\sigma_{X1-X2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Also the sampling distribution of sum of means is as defined below:

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

And the standard deviation

$$\sigma_{p1+p2}^2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Self-Assessment Exercise 1

Sampling distribution of the difference of two means is defined as-------

1.4 Worked Example of Sampling Distribution of Sum of Two Means

Given that p_1 = (30,50) and p_2 = (40,70) show that $\mu_{p1} + \mu_{p2} = \mu_{p1} + \mu_{p2}$;

(ii) $\mu_{p1} - \mu_{p2} = \mu_{p1} - \mu_{p2}$; and (iii) $\sigma_{p1+p2}^2 = \sigma_{p1+p2}^2$ for a sample drawn from each other.

Solution

Sampling sum

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

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

$$\therefore \mu_{p1+p2} = \frac{70+100+90+120}{4} = \frac{380}{4} = 95$$

Considering the 1st population p_2 (30,50)

$$\mu_{p1} = \frac{30+50}{2} = \frac{80}{2} = 40$$

Considering the 2nd population (40, 70)

$$\mu_{p2} = \frac{40+70}{2} = \frac{110}{2} = 55$$

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

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

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

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

Self-Assessment Exercise 2

What is the population and sample mean of P_1 = (70,90), P_2 = (60,80)

1.5 Worked Example of Sample Differences of Two Means

$\mu_{p1} - \mu_{p2} = 40 - 55$ from our calculation of means above

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

Taking the differences of possible =sample μ_{p1-p2}

$$\mu_{p1-p2} = \frac{(30-40)+(30-70)+(50-40)+(50-70)}{4}$$

$$\mu_{p1-p2} = \frac{-10-40+10-20}{4} = \frac{-60}{4} = -15$$

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

(iii) σ^2_{p1+p2} = variance of 70,10, 90 & 120
Note population mean = 95

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

$$\therefore \sigma^2_{p1+p2} = \frac{(70-95)^2+(100-95)^2+(90-95)^2+(120-95)^2}{4} = \frac{-25^2+5^2-5^2+25^2}{4}$$

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

Where $40 = \mu p_1 = mean$ of population 1

$$\sigma^2_{p1} = \frac{(-10)^2+(10)^2}{2} = \frac{100+100}{2} = 100$$

Where 55 = mean of population = μp_2

$$\sigma^2_{p2} = \frac{(15)^2+(15)^2}{2} = \frac{(225)+(225)}{2} = \frac{450}{2} = 225$$

$$\sigma^2_{p2} + \sigma_{p2} = 225+100 = 325$$

Self-Assessment Exercise 3

Sampling distribution of the difference of 2 mean $\bar{X}_1$ & $\bar{X}_2$ is usually written as?

1.6 Summary

In the course of our discussion of this unit you have learnt about Sampling distribution of difference of two means and Sampling distribution of sum of two means

In the course of our discussion on this unit we defined sampling distribution of the difference of two mean as $\mu p_1 - \mu p_2$ and standard deviation of the difference as

$$\sigma_{X1-X2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Tutor Marked Assignment

Given the following population $p_1 = (10,20)$ $p_2 = (30,40)$ show that

i. $\mu p_1 + p_2 = \mu p_1 + \mu p_2$
ii. $\mu p_1 - p_2 = \mu p_1 - \mu p_2$

1.7 Reference/Further Reading

Adedayo, O.A. (2006): Understanding statistics. JAS publishers, Akoka, Yaba.
Esan F.O. and Okafor, R.O. (2010): Basic statistical method (revised edition) Toniichristo Concept, Lagos.
Murray, R. S. and Larry J. S. (1998): (Schaum outlines series). Statistics (Third edition) MCGRAW HILLS.
Olufolabo, O.O. and Talabi, C.O. (2002): Principles and practice of statistics. HASFEM Nig Enterprises Shomolu, Lagos.
Oyesiku, O.K. and Omitogun, O. (1999): Statistics for social and management sciences. Higher Education Books Publishers, Lagos.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1
The sampling distribution of the difference of two means defined as: $(\bar{X}_1 - \bar{X}_2) = \mu_{p1} - \mu_{p2}$

Answer to Self- Assessment 2

To find the population and sample means of the given data sets P1 = (70, 90) and P2 = (60, 80), we'll calculate the mean for each set.

Population mean (μ): The population mean is the average of all the values in the population. Since the given data sets are not specified as samples, we'll treat them as populations.

For P1 = (70, 90): μ1 = (70 + 90) / 2 = 80

For P2 = (60, 80): $\mu 2 = (60 + 80) / 2 = 70$

Sample mean ($\bar{x}$): If the given data sets were samples instead of populations, we would calculate the sample mean. However, since they are not specified as samples, we won't calculate the sample mean in this case.

To summarize:

- The population mean for P1 is $\mu 1 = 80$.
- The population mean for P2 is $\mu 2 = 70$.

Note: The sample mean would be calculated as the sum of the sample values divided by the sample size. However, since the given data sets are not specified as samples, we don't calculate the sample mean.

Answer to Self- Assessment 3

The sampling distribution of the difference in means, $\bar{X}1 - \bar{X}2$, is normally distributed with mean $\mu 1 - \mu 2$ and standard deviation $\text{sqrt}[\sigma 1^2/n1 + \sigma 2^2/n2]$.

UNIT FOUR: PROBABILITY DISTRIBUTION

1.1 Introduction

For thorough understanding of this unit, it is assumed that you must have familiarized yourself with introductory statistics and unit one of this module. The main thrust of this unit is to introduce to you the concept of probability distribution, its discussion, calculation and interpretation of result. This unit is fundamental to the understanding of subsequent modules. This is because other unit and module will be discussed on the basis of the fundamentals concept explained here.

1.2 Learning Outcomes

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

i. Describe the concept of probability

ii. Explain the different probability distribution

iii. Calculate the different probability distribution

1.3 Probability Defined

Statisticians spends quality time measuring data and drawing conclusions based on his measurement sometimes, all the data is available to the statisticians and the measurement are bound to be accurate in such circumstances, it can be said that the statistician has perfect knowledge of the population.

There are a-times whereby this will not be the usual situation. In most cases, the statistician will not have the details he wants about the population and will be unable to collect the information he wants because of cost and labour involved.

However, because the entire population has not been examined, the statistician can never be completely sure of the result, so when quoting conclusion based on sample evidence, it is usual to state how confident the statistician is about his result. So you will often see estimates quoted with 85% confidence. This is simply talking about the probability that the estimate is right is 85%.

Probability is a fundamental concept in statistics and is used to quantify the likelihood or chance of an event happening. It's a mathematical framework for representing uncertain statements. Probability can range from 0 to 1, where 0 indicates that an event will not happen, 1 indicates that an event will definitely happen, and values in between represent varying degrees of likelihood.

Here are a few key concepts related to probability:

1. **Random Experiment**: An experiment or a process for which the outcome cannot be predicted with certainty.

2. **Sample Space**: The set of all possible outcomes of a random experiment. It's often denoted by the symbol S.
3. **Event**: An event is any subset of a sample space. It represents the outcomes of the random experiment that we are interested in.
4. **Probability of an Event**: Given a sample space S and an event A, the probability of A (often denoted as P(A)) is the measure of the likelihood that A will occur.

The probability of an event A is calculated as follows:

```
P(A) = (Number of outcomes where A occurs) / (Total number of outcomes in the sample space)
```

This is the classical or frequentist interpretation of probability. However, there are other interpretations as well.

5. **Probability Rules**:
    - **The Complementary Rule**: The probability that an event A does not occur is 1 minus the probability that it does occur. This is denoted as $P(A') = 1 - P(A)$.
    - **The Addition Rule**: The probability of the occurrence of at least one of two events A or B is equal to the sum of their individual probabilities minus the probability of their intersection (if they are not mutually exclusive). This is denoted as $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

- **The Multiplication Rule**: The probability of two events A and B occurring together (or one after the other) is equal to the probability of A times the probability of B given that A has occurred (if they are not independent). This is denoted as $P(A \cap B) = P(A) * P(B|A)$.

6. **Conditional Probability**: This is the probability of an event given that another event has occurred. If we are interested in the probability of event A given that event B has occurred, this is denoted as P(A|B).
7. **Independence**: Two events are independent if the occurrence of one event does not affect the occurrence of the other event. If A and B are independent, then $P(A \cap B) = P(A) * P(B)$.

These are some of the foundational concepts in probability that serve as the basis for much of statistics. Probability theory is essential in areas like hypothesis testing, confidence intervals, Bayesian statistics, and many more advanced statistical techniques.

The probability of a value X of a random variable is usually referred to as the limiting value of the relative frequency of that value as the total number of observation on the variable approaches infinity, the value which the relative frequency assumes at the limit as the number of observations tends to infinity. This can be written as

$$P(x) = \lim_{n\to} \frac{f}{\sum fx}$$

Self-Assessment Exercise 1

What is another name that probability can be called?

1.4 Probability Distribution of a Random Variable

If a variable is discrete, if its value are distinct i.e. they are separated by finite distance. To each we may assign a given probability. If x is a discrete random variable which may assume the values X_1, X_2.....X_n with respective probabilities $f(x_1)$, $f(x_2)$....., $f(x_n)$. Then the entire set of pairs of permissible value together with their respective probabilities is called probability distribution of a random variable x.

A random variable is a variable whose values are associated with the probability of being observed. A discrete random variable is one that can assume only finite and distinct value.

One of the discrete probability is the binomial distribution. This distribution is used to find the probability of X number of occurrences or success of an event, P(x) in n-trials of same experiment.

Binomial distribution is usually use to predict occurrence of events that are mutually exclusive in other words Binomial distribution is useful for problem that are concerned with determining the number of times an event is likely to occur or not occur during a given number of trials and consequently the probability of it occurring or not occurring. Symbolically it is written as;

$$P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$$

where,

$p =$ probability of success

$q = 1 - p =$ probability of failure

$X =$ Number of success desired

$n =$ Number of trials.

The mean of the binomial distribution is $\mu = np$, and the standard deviation is $\sigma = \sqrt{npq}$

Example 1

Using the binomial distribution, find the probability of getting 4 heads in 6 flips of a balanced coin.

$$P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X} = \frac{6!}{4!(6-4)!}(1/2)^4 (1/2)^{6-4}$$

$$= \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1}(1/16)(1/4) = 15(1/64) = \frac{15}{64} \cong 0.23$$

The expected number of heads in 6 flips $= \mu = np = (6)(1/2) = 3$ heads. The standard deviation of the probability distribution of 6 flips is

$$\sigma = \sqrt{np(1-p)} = \sqrt{(6)(1/2)(1/2)} = \sqrt{6/4} = \sqrt{1.5} \cong 1.22 \text{ heads}$$

Because $p = 0.5$ or $1/2,$ this probability distribution is symmetrical.

Example 2

Calculate the probability of getting exactly 2 heads of 3 tosses of a fair coin.

$$P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X} = \frac{3!}{2!(3-2)!}(0.5)^2(0.5)^{3-2}$$
$$= \frac{3\cdot 2\cdot 1}{2\cdot 1\cdot 1}(0.25)(0.5) = \frac{6}{2}(0.125) = 3(0.125) \cong 0.375$$

Self-Assessment Exercise 2

What do you understand by the word a random variable?

1.5 Poisson Distribution

Poisson distribution is another discrete probability distribution useful in describing the number of events that will occur in a specific period of time. It is usually used in determining the probability of a designated number of successes per unit of time. When the event or successes are independent and the average number of successes per unit of time remains constant. Symbolically it is written as;

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

Where;

$e =$ base of the natural logarithmic system, or 2.71828,

$x =$ designated number of successes, or the number of times an event occurs,

$\lambda =$ described as the expected rate of occurrence per unit time or mean rate of occurrence of events

Example

Past experience indicates that an average number of 6 customers per hour stop for gasoline at a gasoline pump.

a) What is the probability of 3 customers stopping in any hour?

b) What is the probability of 3 customers or less in any hour?

c) What is the expected value, or mean, and standard deviation for this distribution?

Solution:

a).

$$P(3) = \frac{6^3 e^{-6}}{3!} = \frac{(216)(0.00248)}{3 \cdot 2 \cdot 1} = \frac{0.53568}{6} = 0.08928$$

b).

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

$$P(0) = \frac{6^0 e^{-6}}{0!} = \frac{(1)(0.00248)}{1} = 0.00248$$

$$P(1) = \frac{6^1 e^{-6}}{1!} = \frac{(6)(0.00248)}{1} = 0.01488$$

$$P(2) = \frac{6^2 e^{-6}}{2!} = \frac{(36)(0.0248)}{2 \cdot 1} = 0.04464$$

$P(3) = 0.08928$ (from part a)

Thus, $P(X \le 3) = P(0) + P(1) + P(2) + P(3)$

$$= 0.00248 + 0.01488 + 0.04464 + 0.08928 = 0.15128$$

c).

The expected value, or mean, of this Poisson distribution is $\lambda = 6$ customers, and the standard deviation is $\sqrt{\lambda} = \sqrt{6} \cong 2.45$ customers.

Self-Assessment Exercise 3

Define standard deviation of Poisson distribution?

1.5.1 Probability Distribution of a Continuous Variable (Normal Distribution)

If a variable is continuous, it can assume an infinite number of values within a given interval. An important feature of probability distribution is that the areas under these curve represents probabilities. The total area under the curve of a probability distribution, being the sum of individual probabilities is equal to unity (1).

The normal distribution as a continuous probability distribution and the most commonly used distribution in statistical analysis. The normal curve is bell-shaped and symmetrical about its mean. Usually, it extends indefinitely in both directions, but most of the area (probability) is clustered around the mean.

To find the probabilities for problems involving the normal distribution, first convert the x value into corresponding z value using:

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

Example

The mean weight of 500 male students at a certain college is 151 pounds, and the standard deviation is 15 pounds. Assuming that the weights are normally distributed, determine how many students weigh a). Less than 128 pounds; b). Between 120 and 155 pounds; c). More than 185 pounds.

Solution:

a).

$$z = \frac{128-151}{15} = -1.53$$

Required proportion of students = (area to the left of $z = -1.53$)

$= 0.5 - 0.4370 = 0.063$ (That is, looking up $z = 1.53,$ we get 0.4370)

Thus, the number of students weighing less than 128 pounds is $500(0.063) \cong 32.$

b).

$$z_{120} = \frac{120-151}{15} = -2.07$$

$$z_{155} = \frac{155-151}{15} = 0.27$$

Required proportion of students = (area between $z = -2.07$ and $z = 0.27$)

$= 0.4808 + 0.1064 = 0.5872$

Thus, the number of students weighing between 120 and 155 pounds is $500(0.5872) \cong 294.$

c).

$$z = \frac{185-151}{15} = 2.27$$

Required proportion of students = (area to the right of $z = 2.27$)

$= 0.5 - 0.4884 = 0.0116$ (That is, looking up $z = 2.27,$ in normal distribution Table, we get 0.4884). Thus, the number of students weighing more than 185 pounds is $500(0.0116) \cong 6.$

If W denotes the weight of a student at random, we can summarize the above results in terms of probability as:

$P(W \leq 128) = 0.063,\ P(120 \leq W \leq 155) = 0.5872, \text{and}\ P(W \leq 185) = 0.0116$

1.5.2 Attributes of A Normal Distribution Curve

The normal distribution, also known as the Gaussian distribution, is a type of continuous probability distribution for a real-valued random variable. Here are some key attributes of a normal distribution curve:

1. **Symmetry**: A normal distribution is symmetric about its mean. The left half and the right half of the distribution are mirror images of each other.
2. **Mean, Median, and Mode**: For a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.
3. **Shape**: The shape of a normal distribution is often referred to as a "bell curve." It starts at the mean and then decreases on both sides, creating a bell-like shape.
4. **Parameters**: The normal distribution is determined by two parameters, the mean (mu) and the standard deviation (sigma). The mean determines the center of the

distribution, and the standard deviation determines the spread or the width of the distribution.

5. **Standard Normal Distribution**: A standard normal distribution is a special case of the normal distribution where the mean is 0 and the standard deviation is 1.
6. **Empirical Rule**: Also known as the 68-95-99.7 rule, it states that in a normal distribution, about 68% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations.
7. **Area Under the Curve**: The total area under the normal distribution curve is equal to 1, which indicates that the probability of an event happening falls somewhere under the curve.
8. **Asymptotic Ends**: The ends of a normal distribution curve extend to negative and positive infinity, asymptotically touching but never crossing the horizontal axis. This means that the probability of an event happening can never be zero.
9. **Density Function**: The probability density function of a normal distribution is given by: `f(x) = (1/(sigma*sqrt(2*pi))) * e^(-(x-mu)^2 / (2*sigma^2))` where e is the base of the natural logarithm, pi is a mathematical constant, sigma is the standard deviation and mu is the mean of the distribution.
10. **No Skewness and Kurtosis**: Normal distribution does not exhibit skewness (lack of symmetry) or kurtosis (tailedness). However, it's worth noting that the measure

of kurtosis of a standard normal distribution is 3, and often excess kurtosis (kurtosis-3) is used, which is zero for a normal distribution.

These attributes make the normal distribution a fundamental element in the field of statistics and the theory of probability. It's used in various statistical methods, including hypothesis testing, regression analysis, and quality control processes.

Self-Assessment Exercise 4

Explain the attributes of a normal distribution curve

1.6 Summary

From our discussion so far, you have learnt about: Probability, Probability distribution. Different probability distribution, the binomial, Poisson, and normal distribution. In the course of our discussion of this unit, we have defined the different probability distributions binomial distribution is defined as:

$$P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$$

where,

$p =$ probability of success

$q = 1 - p =$ probability of failure

$X =$ Number of success desired

$n =$ Number of trials.

The mean of the binomial distribution is $\mu = np$, and the standard deviation is $\sigma = \sqrt{npq}$

Normal distribution

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

Tutor Marked Assignment

A study shows that 40% of the people entering a supermarket make a purchase. Using

(a) binomial distribution, (b) Poisson distribution find the probability that out of 30

people entering the supermarket 10 or more will make a purchase.

1.7 Reference/Further Reading

Adedayo, A.O. (2006): Understanding Statistics. JAS Publishers, Lagos.
Dominick, S. and Derrick, R. (2011): Statistics and Econometrics. (Schaum's outlines) McGraw Hill, New York.
Esan, E.O. and Okafor, R.O. (2010): Basic Statistical Methods (Revised Edition) Tonichristo Concept.
Ezie, O. and Ezie, K. P. (2023). Applied Statistics and Research Techniques: A Practical Guide for Data Analysis. Kabod Publisher, Kaduna.
Koutsoyianis, A. (2003): Econometric Methods (second edition). Palgrave publishers Ltd (formerly Macmillan press ltd), London and basin stoke.
Murray, R. S. and Larry, J. S. (1998): Statistics (Schaum outlines). McGraw Hill.
Olufolabo, O.O. and Talabi, C.O. (2002): Principles and practice of statistics, HASFEM (NIG) ENTERPRISES, Somolu, Lagos.
Oyesiku, O.K. and Omitogun, O. (1999): Statistics for social and management sciences. Higher Education Book Publishers Lagos.
Owen, F. and Jones R. (1983): Statistics. Polytech Publishers Ltd. Stockport.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Probability can also be referred to as "chance" or "likelihood". It quantifies the extent to which an event is likely to occur.

Answer to Self- Assessment 2

A random variable is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

1. **Discrete Random Variable**: A random variable is discrete if its possible values are countable. For example, the number of heads obtained in a fixed number of flips of a coin is a discrete random variable.
2. **Continuous Random Variable**: A random variable is continuous if its possible values include an entire interval on the number line. For example, the time it takes for a student to complete a test is a continuous random variable because it could be any non-negative real number.

Answer to Self- Assessment 3

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space. These events must occur with a known constant mean rate and independently of the time since the last event.

The standard deviation (σ) for a Poisson distribution is simply the square root of the mean (μ).

So, if λ is the mean number of successes (the mean rate of the Poisson distribution), then:

$\sigma = \text{sqrt}(\lambda)$

This characteristic is unique to the Poisson distribution and is one of its key properties. This means that as the average number of successes increases, the standard deviation (i.e., the variability) increases as well.

Answer to Self- Assessment 4

A normal distribution curve, also known as a Gaussian distribution or bell curve, has the following key attributes:

1. **Symmetry**: The curve is symmetric around the mean (average). This means that the left half of the distribution is a mirror image of the right half.
2. **Unimodal**: The curve has a single peak, known as the mode, which corresponds to the most frequently occurring score. For a normal distribution, the mode is equal to the mean and median.
3. **Mean, Median, Mode Coincidence**: The mean (average), median (middle value), and mode (most frequent value) of the distribution are all equal, and they all occur at the peak of the curve.
4. **Asymptotic Nature**: The curve approaches, but never touches, the x-axis. This means there are technically no minimum or maximum values.
5. **Empirical Rule**: Also known as the 68-95-99.7 rule. About 68% of values fall within 1 standard deviation from the mean, about 95% within 2 standard deviations, and about 99.7% within 3 standard deviations.

6. **Specific Shape**: The shape of the curve is completely described by the mean and standard deviation. The mean determines the center of the distribution. The standard deviation determines the spread of the distribution; a small standard deviation results in a narrow, peaked curve, while a large standard deviation results in a wide, flat curve.

These attributes make the normal distribution extremely useful in a variety of applications, particularly in statistics and data analysis, where it forms the basis for various methodologies and analyses.

MODULE TWO: ANALYSIS OF VARIANCE AND ANALYSIS OF COVARIANCE

Unit 1: One-way factor analysis of variance

Unit 2: Two-way factor analysis of variance

Unit 3: Analysis of covariance

UNIT ONE: ONE-WAY ANALYSIS OF VARIANCE

1.1 Introduction

A detailed knowledge and understanding of introductory statistics is assumed, it is also expected that students would have familiarized themselves with hypothesis testing. This unit is one of the three units in module 2 of the course.

1.2 Learning Outcomes

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

- Total sum of square
- Sum of square between groups

- Sum of square within the group
- Mean square

1.3 Logic of Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique used to analyze the difference among group means in a sample. The basic logic behind ANOVA is to compare the ratio of between-group variance (variation between different groups' means) to the within-group variance (variation within each group). If the between-group variance is significantly larger than the within-group variance, it can be concluded that there's a statistically significant difference among the group means.

Here are the steps to understand the logic behind ANOVA:

1. **Setting Up Hypotheses**:
    - Null Hypothesis (H0): All group means are equal (no difference in means)
    - Alternative Hypothesis (H1): At least one group mean is different from the others
2. **Partitioning of Variance**:

    ANOVA breaks down the total variance in the data into two types:

    - Between-group Variance: This is the variability of the group means around the grand mean (mean of all observations).

- Within-group Variance: This is the average variability within each group around their respective group mean.

3. **Calculating the F-statistic**:

   The F-statistic is calculated by dividing the between-group variance by the within-group variance. Mathematically,

   F = (Between-group variance) / (Within-group variance)

   If the group means are all equal, the expected value of the F-statistic is 1.

4. **Testing the Hypotheses**:

   If the observed F-statistic is significantly larger than 1 (determined using the F-distribution and degrees of freedom), we reject the null hypothesis and conclude that there is a statistically significant difference among the group means.

5. **Post Hoc Analysis**:

   If the null hypothesis is rejected, we know at least one group mean is different, but we don't know which ones. To identify the specific groups that differ from each other, we perform post hoc tests (like Tukey's HSD, Bonferroni, etc.).

Analysis of variance (ANOVA) is usually used to test null hypothesis that the means of two or more populations are equal versus the alternative that at least one of the means is different. The null hypothesis (Ho) tested in the case of ANOVA is that the means of the

population from which the sample is drawn are all equal i.e.

$H_0 : \mu_1 = \mu_2 = ...\mu_K$ while the alternative hypothesis says that H_o taken as a whole is not true i.e., $H_1 : \mu_1 \neq \mu_2 \neq ...\mu_K$

It is to be noted that each time ANOVA is used, all we are trying to do is to analyze or test the variances in order to test the null hypothesis about the means (i.e. $H_0 : \mu_1 = \mu_2 = ...\mu_K$). The ANOVA procedure is based on mathematical theory that the independent sample data can be made to yield two independent estimates 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.

In essence, the logic of ANOVA involves decomposing the total variability in the data into variability between groups and variability within groups, comparing them to see if the difference among group means is larger than what would be expected by chance.

Self-Assessment Exercise 1

State the null hypothesis of analysis of variance?

1.4 Assumptions of ANOVA

When performing an Analysis of Variance (ANOVA), there are certain assumptions that need to be satisfied for the results to be reliable. If these assumptions are not met, the results could be misleading. Here are the key assumptions:

1. **Independence of Observations**: The observations within each group and across different groups should be independent of each other. This means that the occurrence of one event does not influence the occurrence of another event. This assumption is often satisfied through proper study design.
2. **Normality**: The responses for each group are normally distributed in the population. In practice, if the sample size is large enough, thanks to the Central Limit Theorem, the violation of this assumption may not be severe because the sampling distribution of the mean tends to be normally distributed even if the population distribution is not.
3. **Homogeneity of Variance (Homoscedasticity)**: The variance within each of the groups should be approximately equal. This is also known as the assumption of homoscedasticity. One common technique to check this assumption is Levene's test. If this assumption is violated, it could inflate the Type I error rate (the likelihood of falsely claiming a significant effect). However, there are variants of ANOVA (like Welch's ANOVA) that are more robust against this assumption violation.
4. **Additivity and Linearity**: The expected value of the dependent variable should be a sum of the effect of the independent variables. ANOVA assumes a linear

relationship between the mean of the response variable and the categorical variables.

5. **No Multicollinearity**: Although this is not a strict assumption of ANOVA, if you are using ANOVA in a regression-like setting (such as Analysis of Covariance, or ANCOVA), it is assumed that there is no perfect linear relationship between explanatory variables.

Violation of these assumptions does not always mean that you cannot perform or trust the results of an ANOVA, but it does mean you should be cautious in your interpretation. Depending on the severity of the violation, there are different strategies to deal with them, such as data transformations, using non-parametric tests, or using a more robust variant of ANOVA.

Self-Assessment Exercise 2

State and discuss three assumptions of analysis of variance

1.5 Steps involved in ANOVA Analysis

(i) Estimate the population variance from the variance between sample means (MSB)

(ii) Estimate the population variance from the variance within the samples (MSW)

(iii) Compute the fisher ratio.

This is given as: $F = \frac{SSB / K - 1}{SSW / N - K} = \frac{MSB}{MSW}$

i.e. F = Variance of between the sample mean Variance of within the sample

(iv) Compute the various degree of freedom i.e. the degree of freedom for between, within and total groups. Degree of freedom for the sum between group is given as K-1 Degree of freedom within group is writer as N-K. Total degree of freedom as N - 1 Where K = no of samples, N = 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 (Ho) and concluded that the difference is not significant. If the calculated value is greater the critical value reject Ho 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.

Table M2.1.1: ANOVA Summary Table

Model	Sum of Squares	Degree of freedom (df)	Mean Square or Variance Estimate	F-value
Between Samples	*SSB*	$v_1 = K - 1$	*MSB*	F^*

Within Samples	*SSW*	$v_2 = N - K$	*MSW*	F-tables with $v_1 = K - 1$, $v_2 = N - K$ Degrees of freedom
Total	*SST*	$N - 1$		

Where:

MSB = Mean square between samples

MSW = Mean square within samples

SSB = sum of squares between samples

SSW = sum of squares within samples

N = Total number of observations

K = Number of groups

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(X_{ij} - \overline{\overline{X}}\right)^2$$

$$SSB = \frac{1}{n_j}\sum_{j=1}^{k}\left(X_j - \overline{\overline{X}}\right)^2$$

$$SSW = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}\right)^2$$

Alternatively, $SSW = SST - SSB$

$Total\,df = SSB\,df + SSW\,df$

$N - 1 = K - 1 + N - K$

When these 'sum of squares' are divided by their associated degrees of freedom, we get the following variances or mean square terms:

$$MSB = \frac{SSB}{K-1}; MSW = \frac{SSW}{N-K}$$

1.5.1 Worked Example

In a study to ascertain whether significant differences exist in the main sources of information to farmers in Agulu, Anambra, the following where obtained:

Table M2.1.2: Sources of Information

Village	News Paper (X1)	Radio (X2)	Television (X3)	Extension (X4)	Person to Person (X5)
A	5	20	10	15	8
B	10	10	8	15	6
C	8	15	5	5	7
D	6	10	5	10	12
E	4	80	3	10	4
F	4	2	10	4	6

To start with, one needs to state the null and alternative hypothesis, as well as the alpha level. Thus, the hypothesis is stated as:

H_0 : There is no significant difference in the main sources of information to farmers in Agulu, Anambra

H_1 : There is a significant difference in the main sources of information to farmers in Agulu, Anambra

The alpha level is fixed at 0.05 or 5percent.

Step 1: Obtain the totals and means of each of the five samples

Table M2.1.3: Total for Each Sample

Village	X1	X2	X3	X4	X5
A	5	20	10	15	8
B	10	10	8	15	6
C	8	15	5	5	7
D	6	10	5	10	12
E	4	8	3	10	4
F	4	2	10	4	6
Totals	37	65	41	59	43
Means	6.17	10.83	6.83	9.83	7.17

Step 2: Obtain the deviation of the sample values from the grand mean, $\bar{\bar{X}}$.
The grand mean itself is given by:
Total number of items or observations, N=30
Grand total of items = 37+65+41+59+43 = 245
Grand mean = $\bar{\bar{X}}$ = 245/30 = 8.17

Table M2.1.4: Items minus Grand Mean

Village	$X1-\bar{\bar{X}}$	$X2-\bar{\bar{X}}$	$X3-\bar{\bar{X}}$	$X4-\bar{\bar{X}}$	$X5-\bar{\bar{X}}$
A	-3.17	11.83	1.83	6.83	-0.17
B	1.83	1.83	-0.17	6.83	-2.17
C	-0.17	6.83	-3.17	-3.17	-1.17
D	-2.17	1.83	-3.17	1.83	3.83
E	-4.17	-0.17	-5.17	1.83	-4.17
F	-4.17	-6.17	1.83	-4.17	-2.17
Totals	-12.02	15.98	-8.02	9.98	-6.02

Table M2.1.5: Squares of Items minus Grand Mean

Village	$\left(X1-\bar{\bar{X}}\right)^2$	$\left(X2-\bar{\bar{X}}\right)^2$	$\left(X3-\bar{\bar{X}}\right)^2$	$\left(X4-\bar{\bar{X}}\right)^2$	$\left(X5-\bar{\bar{X}}\right)^2$
A	10.05	139.95	3.35	46.65	0.03
B	3.35	3.35	0.03	46.65	4.71
C	0.03	46.65	10.05	10.05	1.37
D	4.71	3.35	10.05	3.35	14.67
E	17.39	0.03	26.73	3.35	17.39
F	17.39	38.07	3.35	17.39	4.71
Totals	**52.91**	**231.39**	**53.55**	**127.43**	**42.87**

Step 1:

Total sum of squares of all 30 items:

$SST = 52.91+231.39+53.55+127.43+42.87 = 508.2$

Step 2:

Between sample sum of squares:

$$SSB = \frac{1}{6}\left((-12.02)^2 + (15.98)^2 + (-8.02)^2 + (9.98)^2 + (-6.02)^2\right)$$

Note: the division by 6 is because there are six villages in the sample

$$SSB = \frac{1}{6}\left(144.48 + 255.36 + 64.32 + 99.60 + 36.24\right)$$

$$SSB = \frac{600}{6} = 100$$

Step 3:

Within sample sum of squares = (total sum of squares) - (between sample sum of squares)

$SSW = 508.2 - 100 = 408.2$

Step 4:

Determine the degree of freedom

$K = 5$ (*Number of categories or groups*)

$N = 30$

For the total sum of squares, degree of freedom as already noted is N-l (30-l) =29; since there is a total of 30 items in the five samples or groups.

For the between samples sum of squares, the degree of freedom is 5-1 = 4 (since there are only five groups).

For the within sample sum of squares, the degree of freedom is 30 - 5 = 25 (the degree of freedom for the total sum of squares minus the between samples sum of squares degree of freedom).

Mean square between

$$MSB = \frac{100}{5-1} = \frac{100}{4} = 25$$

Mean square within

$$MSW = \frac{408.2}{30-5} = \frac{408.2}{25} = 16.33$$

Step 5:

Compute the f-statistic

To be able to make conclusions about whether the five samples are significantly different, we need to compare the variance estimate for the "between" group (which is 25) with that for the "within' group (which is 16.33).

$F = \frac{25}{16.33} = 1.53.$ That is:

$$F = \frac{SSB/K-1}{SSW/N-K} = \frac{MSB}{MSW}$$

$$F = \frac{100/(5-1)}{408.2/(30-5)} = \frac{100/4}{408.2/25} = \frac{25}{16.33} = 1.53$$

In our example, with a calculated F of 1.53, the table value of $F\frac{4}{25}$ (the 4 and 25 refer to the degree of freedom) under the 5 percent significance level is 2.76; while under 1percent, it is 4.18 (look at the F-statistical Table for these results).

Decision

Our calculated F is therefore not significant at the 5 percent level. We, therefore accept H_0 and conclude that there is no significant difference in the main sources of information to farmers in Agulu, Anambra.

Table M2.1.6: Results in ANOVA Table

Model	Sum of Squares	Degree of freedom (df)	Mean Square or Variance Estimate	F-value
Between Samples	100.0	4	25.00	1.53
Within Samples	408.2	25	16.33	
Total	508.2	29		

Self-assessment exercise 3

State the formulae for sum of square?

1.6 Summary

In the course of our study of one-way analysis of variance you must have learnt about;

- The meaning of ANOVA
- The assumptions of ANOVA
- Steps involved in ANOVA Analysis

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

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(X_{ij} - \overline{\overline{X}}\right)^2$$

$$SSB = \frac{1}{n_j}\sum_{j=1}^{k}\left(X_j - \overline{\overline{X}}\right)^2$$

$$SSW = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}\right)^2$$

1.7 References/Further Readings

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

Dominick, S. and Derrick, R. (2011): Statistics and Econometrics, (Schaum Outlines) McGraw-Hill Company, New York.

Edward, E.L. (1983): Statistical analysis in Economics and Business. Houghton Mufflin Company, Boston.

Ezie, O. and Ezie, K. P. (2023). Applied Statistics and Research Techniques: A Practical Guide for Data Analysis. Kabod Publisher, Kaduna.

Olufolabo, O.O. and Talabi, C.O. (2002): Principles and Practice of Statistics; HAS-FEM ENTERPRISES Somolu, Lagos.

Owen, F. and Jones, R. (1978): Statistics, Polytech Publishers Ltd, Stockport.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

In Analysis of Variance (ANOVA), the null hypothesis (H0) typically asserts that all population means from the different groups being compared are equal.

Specifically, if you're comparing k different groups, the null hypothesis would be:

H0: $\mu 1 = \mu 2 = \mu 3 = ... = \mu k$

Here, $\mu 1, \mu 2, ..., \mu k$ represent the population means for each of the k groups.

So in ANOVA, the null hypothesis states that there's no significant difference between the means of the groups being compared. If the data provide sufficient evidence to reject the null hypothesis, it suggests that at least one group mean is significantly different from the others.

Answer to Self- Assessment 2

Analysis of Variance (ANOVA) is based on several key assumptions. Here are three of the most important ones:

1. **Normality**: This assumption states that the residuals (the differences between the observed and predicted values) are normally distributed. It doesn't assume that the variables themselves are normally distributed, but the errors are. This assumption can be checked using visual methods such as Q-Q plots, or statistical tests like the Shapiro-Wilk test.
2. **Homogeneity of Variance**: This assumption, also known as homoscedasticity, states that all groups have the same variance. In other words, the variability in scores for each group being compared should be roughly the same. The Levene's test or Bartlett's test can be used to check this assumption. If this assumption is violated, you may need to use a different statistical test, such as the Welch's ANOVA, which does not assume equal variances.
3. **Independence of Observations**: This assumption requires that the observations within each group, and across groups, are independent of each other. This means that the occurrence of one event does not affect the probability of the other. This assumption is typically validated through study design rather than statistical tests. For example, if participants are randomly assigned to groups, this helps ensure independence.

Violations of these assumptions can lead to incorrect conclusions, so it's important to check these assumptions when using ANOVA. There are ways to address violations, such as data transformations or using different statistical tests.

Answer to Self- Assessment 3

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(X_{ij} - \overline{\overline{X}}\right)^2$$

$$SSB = \frac{1}{n_j}\sum_{j=1}^{k}\left(X_j - \overline{\overline{X}}\right)^2$$

$$SSW = \sum_{i=1}^{r}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}\right)^2$$

UNIT TWO: TWO-WAY ANALYSIS OF VARIANCE CONTENTS

1.1 Introduction

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

1.2 Learning Outcomes

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

- Explain the meaning of two-way ANOVA
- test for two null hypotheses of two-way ANOVA
- Explain the assumptions of Two-way ANOVA
- Understand how to evaluate two-way ANOVA.

1.3 Two- Way Analysis of Variance Defined

Two-way Analysis of Variance (ANOVA) is a statistical method used to examine the influences and interactions of two different categorical independent variables on one

continuous dependent variable. In other words, it's used when we want to compare the means of one variable (dependent) categorized by two different factors (independent variables).

The two-way ANOVA, also known as factorial ANOVA, extends the concept of the one-way ANOVA where only one factor is considered by adding an additional factor.

Let's break down the two-way ANOVA:

1. **Two Factors (Independent Variables)**: Two-way ANOVA is used when we want to investigate the effect of two different categorical variables. These variables are often referred to as "factors". Each factor will have its own levels or groups. For example, a two-way ANOVA could examine the influence of Diet (factor A: high carb, high protein, low calorie) and Exercise (factor B: none, light, heavy) on weight loss (dependent variable).
2. **Main Effects**: The main effect is the effect of one of your factors on the dependent variable, ignoring the effect of the other factor. In the example above, you would have a main effect of Diet, and a main effect of Exercise.
3. **Interaction Effects**: An interaction effect occurs when the effect of one factor depends on the level of the other factor. In other words, it tests whether the effect of one independent variable on the dependent variable is the same at all levels of the other independent variable. If an interaction effect is present, it means that the effect of Diet on weight loss depends on the level of Exercise, or vice versa.

4. **Hypotheses**: The null hypotheses for a two-way ANOVA are as follows:
    - H0 for main effect of Factor A: The population means of the dependent variable are the same at all levels of Factor A.
    - H0 for main effect of Factor B: The population means of the dependent variable are the same at all levels of Factor B.
    - H0 for interaction effect: There is no interaction between Factor A and Factor B.
5. **ANOVA Table and F-tests**: The output of a two-way ANOVA is usually an ANOVA table, which includes the sum of squares, degrees of freedom, mean squares (variance), F statistic, and p-value for each of the main effects and the interaction effect. An F-test is performed for each effect to test the corresponding null hypothesis.

The assumptions of two-way ANOVA are similar to those of one-way ANOVA: independence of observations, normality of the data, and homogeneity of variances. It is also assumed that there is no multicollinearity between the factors.

As with one-way ANOVA, if significant effects are found, post hoc tests may be conducted to determine which specific groups differ from each other. The interpretation of the two-way ANOVA results depends on whether the interaction term is significant. If the interaction term is significant, it indicates that the effect of one factor changes depending on the level of the other factor. If the interaction term is not significant, then you can consider the main effects of each factor independently.

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

Self-Assessment Exercise 1

Define two-way ANOVA?

1.4 Assumptions of Two-Way ANOVA

Two-way ANOVA, like all statistical tests, has assumptions that need to be met for the test results to be valid. Violating these assumptions can lead to incorrect conclusions. The assumptions of a two-way ANOVA include:

1. **Independence**: The observations are independent of each other. This means that the sampling of observations is done without replacement, the sample is random, and one subject does not influence another subject.
2. **Normality**: The dependent variable is approximately normally distributed for each combination of the groups of the two independent variables. This can be tested

using methods such as Shapiro-Wilk test, Kolmogorov-Smirnov test, or visual methods such as Q-Q plots. If sample sizes are large enough, violation of this assumption might be less problematic due to the Central Limit Theorem.

3. **Homogeneity of variances**: The variances of the dependent variable are equal across all groups, a condition also known as homoscedasticity. This can be tested using Levene's test or Bartlett's test. If the homogeneity of variances assumption is violated, you might need to apply a transformation to your data or use a more robust variant of ANOVA, like Welch's ANOVA.
4. **No perfect multicollinearity**: In the context of factorial ANOVA (which two-way ANOVA falls under), there should not be perfect multicollinearity between the independent variables. This means that one independent variable should not be a perfect linear function of another independent variable.
5. **Additivity**: The effect of the different factors on the dependent variable is additive. In other words, there should be no interaction between the factors (unless explicitly included in the model).
6. **Random Errors**: The errors are random, and there are no patterns in the residuals (the difference between the observed and predicted values). The residuals should be normally distributed with a mean of 0, which can be checked using a residuals plot.

If these assumptions are not met, there are several strategies that you can use, such as applying a transformation to your dependent variable (e.g., log transformation), using

non-parametric statistical tests, or using statistical techniques that are robust to violations of these assumptions.

1.5 Two-way Classification and The Formulars

Table M2.2.1: Two-way classification table

	Treatment (Factor A)		
		1 2 3 ..... t	Total
Block factor B	1	Y11 Y12 Y13..... Yij	Bi
	2 3 4 5 . . .	Y21 Y22 Y23..... Y2t	B2
	B	Ybi Yb2 Yb3..... Ybt	Bb

The Formulars

Steps for constructing the two-way ANOVA

Step 1: Calculate the total variation:

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij} - \overline{\overline{X}}\right)^2$$

Step 2: Calculate the variation between treatment (column) means

$$SSC = r\sum_{j=1}^{c}\left(\bar{X}_j - \overline{\overline{X}}\right)^2$$

Step 3: Calculate the variation between blocks (rows) means

$$SSR = c\sum_{i=1}^{r}\left(\bar{X}_i - \bar{\bar{X}}\right)^2$$

Step 4: Calculate the variation due to random error

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij} - \bar{X}_i - \bar{X}_j\bar{\bar{X}}\right)^2$$

Alternatively, $SSE = SST - SSC - SSR$

Step 5: Calculate the average variations

$$Total\,df = SSC\,df + SSR\,df + SSE\,df$$
$$cr - 1 = c - 1 + r - 1 + (c-1)(r-1)$$

When these sums of squares are divided by their associated degrees of freedom, we get the following variances or *mean square* terms:

$$MSC = \frac{SSC}{c-1}; MSR = \frac{SSR}{r-1}; MSE = \frac{SSE}{(c-1)(r-1)}$$

Step 6: Compute the F-statistic

The test statistic F for analysis of variance is given by:

$$F_1 = \frac{MSC}{MSE}$$

$$F_2 = \frac{MSR}{MSE}$$

The following table shows the general arrangement of the ANOVA Table for two-way analysis of variance.

Table M2.2.2: ANOVA Summary Table for Two-Way Classification

Source of variation	Sum of squares	Degrees of freedom	Mean squares	Test-statistic or f-value
Between columns	SSC	$c-1$	MSC	$F_1 = MSC/MSE$
Between rows	SSR	$r-1$	MSR	$F_2 = MSR/MSE$
Residual error	SSE	$(c-1)(r-1)$	MSE	
Total	SST	$cr-1$		

Generally, for c samples or treatments (columns), r blocks (rows), and number of observations $n = r - c,$ the partitioning of total variation in the sample data is shown below:

$SST = SSB or SSC + SSR + SSE$

Where:

$SST =$ Total variation

$SSB or SSC =$ variation between samples

$SSR =$ variation between blocks

$SSE =$ variation due to random error

It is to be noted that; two (2) separate null hypothesis is considered.

(i) Ho; There is no difference between mean of treatment

(ii) Ho; There is no difference between mean of block.

Decision Criteria

Testing for the equality of population means

- $H_0 : \mu_1 = \mu_2 = ...\mu_k$ *against* $H_1 : \mu_1 \neq \mu_2 \neq ...\mu_k \ (j = 1, 2, ..., c)$ Decision Rule

- Reject H_0 if the calculated value of $F_1 >$ its critical value $F_{\alpha,(c-1),(c-1)(r-1)}$
- Otherwise accept H_0.

Testing for the equality of block effects

$H_0 : b_1 = b_2 = ... = b_r$ *against* $H_1 : b_1 \neq b_2 \neq ...b_r \ (i = 1, 2, ..., r)$

- Reject H_0 if the calculated value of $F_2 >$ its critical value $F_{\alpha,(r-1),(c-1)(r-1)}$ Otherwise accept H_0.

Self-Assessment Exercise 2

State the formulae for Two-way Anova?

1.5.1 Worked Example

Table M2.2.3 gives the first-year earnings (in thousands of dollars) of students with master's degrees from 5 schools and for 3 class rankings at graduation. Test at the 5% level of significance that the means are identical (a) for school populations and (b) for class-ranking populations.

(a) The hypotheses to be tested are:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ *against* $H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$

where μ refers to the various means for factor A (school) populations.

Table M2.2.3: First-Year Earnings of MA Graduates of 5 Schools and 3 Class Ranks (in Thousands of Dollars)

Class Ranks	School 1	School 2	School 3	School 4	School 5	Sample Mean
Top 1/3	20	18	16	14	12	$\bar{X}_1 = 16$
Middle 1/3	19	16	13	12	10	$\bar{X}_2 = 14$
Bottom 1/3	18	14	10	10	8	$\bar{X}_3 = 12$
Sample mean	$\bar{X}_1 = 19$	$\bar{X}_2 = 16$	$\bar{X}_3 = 13$	$\bar{X}_4 = 12$	$\bar{X}_5 = 10$	$\bar{\bar{X}} = 14$

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij} - \bar{\bar{X}}\right)^2 =$$

$$(20-14)^2 + (18-14)^2 + (16-14)^2 + (14-14)^2 + (12-14)^2 +$$
$$(19-14)^2 + (16-14)^2 + (13-14)^2 + (12-14)^2 + (10-14)^2 +$$
$$(18-14)^2 + (14-14)^2 + (10-14)^2 + (10-14)^2 + (8-14)^2 = 194$$

$$SSC = r\sum_{j=1}^{c}\left(\bar{X}_j - \bar{\bar{X}}\right)^2 \text{ (Between column variations)}$$
$$= 3\left[(19-14)^2 + (16-14)^2 + (13-14)^2 + (12-14)^2 + (10-14)^2\right]$$
$$= 3(25+4+1+4+16) = 150$$
$$SSR = c\sum_{j=1}^{r}\left(\bar{X}_i - \bar{\bar{X}}\right)^2 = 5\left[(16-14)^2 + (14-14)^2 + (12-14)^2\right]$$
$$= 5(4+0+4) = 40$$
$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij} - \bar{X}_i - \bar{X}_j \bar{\bar{X}}\right)^2 = SST - SSC - SSR = 194 - 150 - 40 = 4$$

These results are summarized in Table below. From Tabulated F-value, $F_1 = 3.84$ for degrees of freedom 4 and 8 and $\alpha = 0.05$. Since the calculated $F_1 = 70,$ we reject H_0 and accept $H_1,$ that the population means of first-year earnings for the 5 schools are different.

Table M2.2.3: ANOVA Summary Table for Two-Way Classification

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-value

Explained by schools (A) (Between columns)	$SSC = 150$	$c-1=4$	$MSC = \frac{150}{4} = 37.5$	$F_1 = \frac{MSC}{MSE} = \frac{37.5}{0.5} = 70$
Explained by ranking (B) (Between rows)	$SSR = 40$	$r-1=2$	$MSR = \frac{40}{2} = 20$	$F_2 = \frac{MSR}{MSE} = \frac{20}{0.5} = 40$
Residual error	$SSE = 4$	$(c-1)(r-1)=8$	$MSE = \frac{4}{8} = 0.5$	
Total	$SST = 194$	$cr-1=14$		

(b) The hypotheses to be tested are:

$H_0 : \mu_1 = \mu_2 = \mu_3$ *against* $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

Where; μ refers to the various means for factor B (class-ranking) populations. From Table M2.2.3, we get that the calculated value of $F_2 = MSC / MSR = 40.$ Since this is larger than the tabular value of $F_2 = 4.46$ for *df* 2 and 8 and $\alpha = 0.05$, we reject H_0 and accept H_1, that the population means of first-year earnings for the 3 class rankings are different. Thus, the type of school and class ranking are both statistically significant at the 5% level in explaining differences in first-year earnings.

Self-Assessment Exercise 3

Discuss the assumptions of Two-way ANOVA?

1.6 Summary

In the course of our discussion on two-way analysis of variance, we have learnt about:

(i) Sum of square of Column (or Factor A)

(ii) Sum of square of Row (or Factor B)

(iii) Sum of square of the error term

(iv) Mean square of Column (or Factor A)

(v) Mean square of Row (or Factor B)

(vi) F-ratio of both Column and Row

(vii) Sum of Square of total variation.

In our discussion the following definition were inferred to:

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij} - \overline{\overline{X}}\right)^2$$

$$SSC = r\sum_{j=1}^{c}\left(\bar{X}_j - \overline{\overline{X}}\right)^2$$

$$SSR = c\sum_{i=1}^{r}\left(\bar{X}_i - \overline{\overline{X}}\right)^2$$

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(X_{ij} - \bar{X}_i - \bar{X}_j\overline{\overline{X}}\right)^2$$

Alternatively, $SSE = SST - SSC - SSR$

$$MSC = \frac{SSC}{c-1}; MSR = \frac{SSR}{r-1}; MSE = \frac{SSE}{(c-1)(r-1)}$$

$$F_1 = \frac{MSC}{MSE}$$

$$F_2 = \frac{MSR}{MSE}$$

Tutor Marked Assignment

Submit a one-page essay on the definition and derivation of MSE, SST and F-ratio.

1.7 References/Further Reading

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

Dominick, S. and Derrick, R. (2011): Statistics and Econometrics, (Schaum's Outlines) McGraw-Hill Company, New York.
Edward, E. L. (1983): Statistical analysis in Economics and Business. Houghton Mifflin Company. Boston.
Ezie, O. and Ezie, K. P. (2023). Applied Statistics and Research Techniques: A Practical Guide for Data Analysis. Kabod Publisher, Kaduna.
Olufolabo, O.O. and Talabi, C.O. (2002): Principles and Practice of Statistics; HAS-FEM ENTERPRISES Somolu, Lagos.
Owen, F. and Jones, R. (1978): Statistics, Polytech Publisher Ltd, Stockport.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Two-way Analysis of Variance, often called two-way ANOVA, is a statistical procedure that can be used to compare the means of three or more groups that have been split on two independent variables. In other words, it is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable.

The two-way ANOVA not only aims to assess the main effect of each independent variable but also whether there is any interaction between them. This means the effect of one independent variable may depend on the level of the other independent variable.

Answer to Self- Assessment 2

The test statistic F for analysis of variance is given by:

$$F_1 = \frac{MSC}{MSE}$$

$$F_2 = \frac{MSR}{MSE}$$

Answer to Self- Assessment 3

Two-way ANOVA, like other forms of ANOVA, has several key assumptions that must be met for the analysis to be valid:

1. **Normality**: The residuals (i.e., the error terms) should be normally distributed. This does not mean that the dependent variable itself has to be normally distributed, but the errors are expected to follow a normal distribution. This

assumption can be checked using a variety of methods, including visual plots like histograms or Q-Q plots, or formal statistical tests such as the Shapiro-Wilk test.

2. **Independence of Observations**: The observations should be independent of each other. This means the data should not be repeated measures (from the same subject) and there should be no relationship between the observations. This is more a condition of the experimental design and data collection, and it is usually not possible to test for this statistically.
3. **Homogeneity of Variance (Homoscedasticity)**: The variances of the different groups should be equal. This assumption can be checked using tests like Levene's Test or Bartlett's Test. Violations of this assumption can sometimes be remedied with data transformations.
4. **No interaction effect**: This is a specific assumption of the two-way ANOVA. The effects of the two independent variables should be additive, meaning there should be no interaction effect between the variables. If an interaction effect is present, it means the effect of one variable depends on the level of the other variable. If this assumption is violated, you should not interpret the main effects without considering the interaction effect.

Like any statistical test, violations of these assumptions can lead to results that are not accurate. However, there are various strategies to deal with assumption violations, including data transformations and non-parametric statistical tests.

UNIT THREE: ANALYSIS OF COVARIANCE CONTENTS

1.1 Introduction

In general, research is conducted for the purpose of explaining the effect of the independent variable on the dependent variable, and the purpose of research design is to provide a structure for the research. In the research design, the researcher identifies and controls independent variable that can help to explain the observed variation in the dependent variable which in turn reduces error variables (unexplained variation).

In addition to controlling and explaining variation through research design, it is also possible to use statistical control to explain the variation in the dependent variable, statistical control is usually used when experimental control is difficult, if not impossible, can be achieved by measuring one or more variable in addition to the independent variable of primary interest and by controlling the variation attributed to these variables through statistical analysis rather than through research design. The analysis procedure

employed in this statistical control is analysis of covariance (ANCOVA).

1.2 Learning Outcomes

At the end of our discussion on analysis of covariance, you should be able to;

- Define analysis of variance
- Define covariate
- Define adjusted Y
- Develop table of analysis of covariance
- Calculate the various terms that may be needed on the computation of ANCOVA Table

1.3 Analysis of Variance Defined

Analysis of covariance is an extension of the one-way analysis of variance that added quantitative variable (covariate) when used, it is assumed that their inclusion will reduce the size of the error variance and thus increase the power of the design. Analysis of covariance (ANCOVA) is a statistical test related to analysis of variance (ANOVA). It tests whether there is a significant difference between groups after controlling for variance explained by a covariate.

A covariate is a continuous variable that correlates with the dependent variable. This means that you can, in effect, “partial out” a continuous variable and run an ANOVA on the result.

This is one way that you can run a statistical test with both categorical and continuous independent variables.

The purpose of analysis of covariance is to remove one or more unwanted factor or variables in the analysis. A variable whose effect one wishes to eliminate by means of a covariance analysis called a covariate sometimes called concomitant variable.

ANCOVA works by adjusting the total sum of square, group sum of squares and error sum of square of the independent variable to remove the influence of the covariate.

Assumptions Of Analysis Of Covariance

- Variance is normally distributed
- Variance is equal between group
- All measures are independent
- Relationship between dependent variable and the covariate as linear
- The relationship between the dependent variable and the covariate is the same for all groups.

Self-assessment exercise 1

What is analysis of covariance?

1.4 Estimation of ANCOVA

Hypothesis for ANCOVA

- H0 and H1; need to be stated slightly different for an ANCOVA than a regular ANOVA.

H0: the group means are equal after controlling for the covariate

H1: the group means are not equal after controlling for the covariate

Below are the lists of notations for the calculation of ANCOVA.

$$SS_{bg(y)} = \frac{\sum^{k}\left(\sum^{n} Y\right)^2}{n} - \frac{\left(\sum^{k}\sum^{n} Y\right)^2}{kn}$$

Note: $k =$ number of groups; $n =$ number of subjects per group.

$$SS_{bg(x)} = \frac{\sum^{k}\left(\sum^{n} X\right)^2}{n} - \frac{\left(\sum^{k}\sum^{n} X\right)^2}{kn}$$

$$SS_{wg(y)} = \sum^{k}\sum^{n} Y^2 - \frac{\sum^{k}\left(\sum^{n} Y\right)^2}{n}$$

$$SS_{wg(x)} = \sum^{k}\sum^{n} X^2 - \frac{\sum^{k}\left(\sum^{n} X\right)^2}{n}$$

$$SP_{bg} = \frac{\sum^{k}\left(\sum^{n} Y\right)\left(\sum^{n} X\right)}{n} - \frac{\left(\sum^{k}\sum^{n} Y\right)\left(\sum^{k}\sum^{n} X\right)}{kn}$$

$$SP_{wg} = \sum^{k}\sum^{n}(XY) - \frac{\sum^{k}\left(\sum^{n} Y\right)\left(\sum^{n} X\right)}{n}$$

$$SS_{total(y)} = SS_{bg(y)} + SS_{wg(y)}$$

$$SS_{total(x)} = SS_{bg(x)} + SS_{wg(x)}$$

$$SP_{total} = SS_{bg} + SS_{wg}$$

$$SS'_{bg} = SS_{bg(y)} - \left[\frac{\left(SP_{bg} + SP_{wg}\right)^2}{SS_{bg(x)} + SS_{wg(x)}} - \frac{\left(SP_{wg}\right)^2}{SS_{wg(x)}}\right]$$

$$SS'_{wg} = SS_{wg(y)} - \frac{\left(SP_{wg}\right)^2}{SS_{wg(x)}}$$

Thus, the summarised One-way ANCOVA formula is given as:

$$F = \frac{MS'_{bg}}{MS'_{wg}} = \frac{SS'_{bg} / k-1}{SS'_{wg} / k(n-1)-1}$$

Note: $k =$number of groups (or levels of the independent variable); $n =$ number of subjects per group.

Table M2.3.1: Analysis of Covariance for a Single Factor Experiment with One Covariate

Source of Variance	Adjusted SS	Degree of freedom (df)	Mean Square or Variance Estimate	F-value
Between Groups	SS'_{bg}	$v_1 = k-1$	MS'_{bg}	$F*$
Within Groups	SS'_{wg}	$v_2 = k(n-1)-1$	MS'_{wg}	F-tables with v_1, v_2 Degrees of freedom

1.5 Worked Example

Examine whether differential treatment of learning-disabled children affect reading scores, after adjusting for differences in the children's prior reading ability.

Table M2.3.2: Small-Sample Data for Illustration of One-Way Between Subject ANCOVA

	Groups					
	Treatment 1		Treatment 2		Control	
	Pre	Post	Pre	Post	Pre	Post
	85	100	86	92	90	95
	80	98	82	99	87	80
	92	105	95	108	78	82
Total	257	303	263	299	255	257

When ANCOVA formula is applied to the data in Table M2.3.2, the six sums of squares and products are as follows:

$$SS_{bg(y)} = \frac{(303)^2 + (299)^2 + (257)^2}{3} - \frac{(859)^2}{(3)(3)} = 432.889$$

$$SS_{wg(y)} = (100)^2 + (98)^2 + (105)^2 + (92)^2 + (99)^2 + (108)^2 + (95)^2 + (80)^2 + (82)^2 - \frac{(303)^2 + (299)^2 + (257)^2}{3} = 287.333$$

$$SS_{bg(x)} = \frac{(257)^2 + (263)^2 + (255)^2}{3} - \frac{(775)^2}{(3)(3)} = 11.556$$

$$SS_{wg(x)} = (85)^2 + (80)^2 + (92)^2 + (86)^2 + (82)^2 + (95)^2 + (90)^2 + (87)^2 + (78)^2 - \frac{(257)^2 + (263)^2 + (255)^2}{3} = 239.333$$

$$SP_{bg} = \frac{(257)(303) + (263)(299) + (255)(257)}{3} - \frac{(775)(859)}{(3)(3)} = 44.889$$

$$SP_{wg} = (85)(100) + (80)(98) + (92)(105) + (86)(92) + (82)(99) + (95)(108) + (90)(105) + (87)(80) + (78)(82) - \frac{(257)(303) + (263)(299) + (255)(257)}{3} = 181.667$$

$$SS'_{bg} = 432.889 - \left[\frac{(44.889 + 181.667)^2}{11.556 + 239.333} - \frac{(181.667)^2}{239.333} \right] = 366.202$$

$$SS'_{wg} = 287.333 - \frac{(181.667)^2}{239.333} = 149.438$$

$$F = \frac{MS'_{bg}}{MS'_{wg}} = \frac{SS'_{bg} / k - 1}{SS'_{wg} / k(n-1) - 1} = \frac{366.202 / (3-1)}{149.438 / 3(3-1) - 1} = \frac{183.101}{29.888} = 6.13$$

From a standard F table, we find that the obtained F of 6.13 exceeds the critical F of 5.79 at $\alpha = 0.05$ with 2 and 5 df. We, therefore, reject the null hypothesis of no change in Literacy Test reading scores associated with the three treatment levels, after adjustment for pre-test reading scores.

Table M2.3.3: One-way ANCOVA Result Summary

Source of Variance	Adjusted SS	Degree of freedom (df)	Mean Square or Variance Estimate	F-value

Between groups	366.202	2	183.101	6.13*
Within groups	149.438	5	29.888	

Self-Assessment Exercise 2

State the formular for one-way ANCOVA

1.6 Summary

In the course of our discussion on analysis of covariance you have learnt about the following:

- Definition of analysis of covariance
- Estimation of analysis of covariance
- Computation of analysis of covariance table
- Adjustment of the dependent

In the course of our discussion the following were inferred.

$$SS_{bg(y)} = \frac{\sum^{k}\left(\sum^{n} Y\right)^2}{n} - \frac{\left(\sum^{k}\sum^{n} Y\right)^2}{kn}$$

Note: $k =$ number of groups; $n =$ number of subjects per group.

$$SS_{bg(x)} = \frac{\sum^{k}\left(\sum^{n} X\right)^2}{n} - \frac{\left(\sum^{k}\sum^{n} X\right)^2}{kn}$$

$$SS_{wg(y)} = \sum^{k}\sum^{n} Y^2 - \frac{\sum^{k}\left(\sum^{n} Y\right)^2}{n}$$

$$SS_{wg(x)} = \sum^{k}\sum^{n} X^2 - \frac{\sum^{k}\left(\sum^{n} X\right)^2}{n}$$

$$SP_{bg} = \frac{\sum^{k}\left(\sum^{n} Y\right)\left(\sum^{n} X\right)}{n} - \frac{\left(\sum^{k}\sum^{n} Y\right)\left(\sum^{k}\sum^{n} X\right)}{kn}$$

$$SP_{wg} = \sum^{k}\sum^{n}(XY) - \frac{\sum^{k}\left(\sum^{n} Y\right)\left(\sum^{n} X\right)}{n}$$

Tutor Marked Assignment

Submit a one page discussion on the definition of analysis of covariance and its assumption.

1.7 References/ Further Readings

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

Dominick, S. and Derrick, R. (2011): Statistics and Econometrics, (Schaum's Outlines) McGraw-Hill Company, New York.

Ezie, O. and Ezie, K. P. (2023). Applied Statistics and Research Techniques: A Practical Guide for Data Analysis. Kabod Publisher, Kaduna.

Kuotsoyanis, A. (2003): Theory of Econometrics (second edition).Palgrave publishers Ltd (formerly Macmillan publishers Ltd), Houndmills, Basingstoke, New York.

www.youtube.com

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Analysis of Covariance (ANCOVA) is a general linear model which blends ANOVA and regression. ANCOVA evaluates whether the means of a dependent variable (DV) are

equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates (CV).

Answer to Self- Assessment 2

$$F = \frac{MS'_{bg}}{MS'_{wg}} = \frac{SS'_{bg} / k-1}{SS'_{wg} / k(n-1)-1}$$

MODULE 3: MULTIPLE REGRESSION ANALYSIS

Unit 1: Estimation of multiple regressions

Unit 2: Correlation coefficient

Unit 3: Multiple correlation coefficient and coefficient of determination

Unit 4: Overall test of significance.

UNIT ONE: ESTIMATION OF MULTIPLE REGRESSIONS

1.1 Introduction

In introductory statistic, simple linear regression is one of the topics discussed. 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).

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.

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.

1.2 Learning Outcomes

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, ... bn

(iv) Test of significance

Coefficient of multiple determinations Test

of overall significance of the regression

Partial correlation coefficient

1.3 Multiple Regression and Assumptions

Multiple regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple regression is to model

the linear relationship between your explanatory (independent) variables and response (dependent) variable.

In a simple linear regression, where you have one explanatory variable and one response variable, the regression model is described as:

$Y = \beta0 + \beta1*X + \varepsilon$

where Y is the dependent variable, X is the explanatory variable, β0 is the y-intercept, β1 is the slope (which gives the change in Y for a unit change in X), and ε is the random error term.

In multiple regression, the model is extended to include more than one explanatory variable. For example, in a multiple regression with two explanatory variables, the model would be:

$Y = \beta0 + \beta1X1 + \beta2X2 + \varepsilon$

This model states that the response Y is a function of two explanatory variables, X1 and X2. The coefficients β1 and β2 represent the change in the dependent variable Y for a one-unit change in X1 and X2, respectively, holding the other variable constant.

The basic assumptions of multiple regression are similar to those of simple linear regression:

1. **Linearity**: The relationship between each independent variable and the dependent variable is linear.
2. **Independence**: The observations are independent of each other.
3. **Homoscedasticity**: The variance of the errors is constant across all levels of the independent variables.
4. **Normality**: For any fixed values of the independent variables, the dependent variable is normally distributed.
5. **No Multicollinearity**: The independent variables are not perfectly correlated with each other. While some correlation between predictors is acceptable, perfect or near-perfect multicollinearity can pose problems, as it can make the model unstable and the estimates of the coefficients unreliable.

Multiple regression is widely used in many fields, including social sciences, economics, marketing, data science, and more, for forecasting, time series modeling, and hypothesis testing. The interpretation of multiple regression analysis can be complex as it requires considering multiple variables at once, as well as potential interactions or correlation among those variables.

Self-Assessment Exercise 1

Define multiple regression model of four variables?

1.4 Goals of Multiple Regression Analysis

Multiple regression is a powerful statistical tool that has several key goals and applications across various fields such as economics, business, social sciences, and health sciences, among others. Here are some of the main goals of multiple regression:

1. **Prediction**: One of the main goals of multiple regression is to use it for predicting the dependent variable based on the values of the independent variables. After a multiple regression model has been developed and validated, it can be used to forecast future outcomes. For example, a company might use multiple regression to predict future sales based on advertising spend, price changes, and economic indicators.
2. **Determining Key Factors**: Multiple regression can be used to identify the key factors or predictors that have the most influence on the outcome variable. By looking at the coefficients of the regression model, we can identify which predictors have a significant effect on the outcome, and the size and direction of their effects.
3. **Quantifying Relationships**: Multiple regression can be used to quantify the relationships between the dependent variable and independent variables. The estimated regression coefficients tell us the change in the mean of the dependent variable for a one-unit change in an independent variable, while holding all other variables constant.
4. **Controlling for Confounding Variables**: In observational studies, there may be confounding variables that are related to both the independent and dependent

variables. Multiple regression allows us to control for these confounding variables and estimate the effect of the independent variables on the dependent variable more accurately.

5. **Testing Hypotheses**: Multiple regression can be used to test hypotheses about the relationships between variables. For example, a researcher could use multiple regression to test whether there's a significant relationship between an outcome variable (like job satisfaction) and predictor variables (like salary, job level, and years of education).
6. **Modeling Interaction Effects**: Multiple regression can also handle interaction effects, where the effect of one independent variable on the dependent variable depends on the level of another independent variable. This can provide more nuanced and accurate models of complex relationships.

It's important to remember that while multiple regression is a powerful tool, it also has assumptions that must be met for the analysis to be valid. These include linearity, independence of errors, homoscedasticity, normality of errors, and lack of perfect multicollinearity among predictors. Always check these assumptions when performing multiple regression analysis.

1.5 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:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$$

$$\hat{\beta}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Where:

$x_1 = X_1 - \overline{X}_1$, $x_2 = X_2 - \overline{X}_2$, $y = Y - \overline{Y}$, $\overline{Y} = \frac{\sum Y}{n}$, $\overline{X}_1 = \frac{\sum X_1}{n}$, and $\overline{X}_2 = \frac{\sum X_2}{n}$

Self-assessment exercise 2

State the goals of multiple regression?

Worked Example

A study is conducted involving 10 students to investigate the relationship and effects of revision time and lecture attendance on exam performance.

Table M3.1.1: Student's exam performance, revision time and lecture attendance

Obs.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Y	40	44	46	48	52	58	60	68	74	80
X1	6	10	12	14	16	18	22	24	26	32
X2	4	4	A5	7	9	12	14	20	21	24

Stands for:
(Y) Exam performance
(X1) Revision time
(X2) Lecture attendance.

Table M3.1.2: Calculating the coefficient of regression

				y	x_1	x_2	$x_1 y$	$x_2 y$	$x_1 x_2$	x_1^2	x_2^2
Obs	Y	$X1$	$X2$	$Y - \overline{Y}$	$X1 - \overline{X1}$	$X2 - \overline{X2}$					
1	40	6	4	-17	-12	-8	204	136	96	144	64
2	44	10	4	-13	-8	-8	104	104	64	64	64
3	46	12	5	-11	-6	-7	66	77	42	36	49
4	48	14	7	-9	-4	-5	36	45	20	16	25

5	52	16	9	-5	-2	-3	10	15	6	4	9
6	58	18	12	1	0	0	0	0	0	0	0
7	60	22	14	3	4	2	12	6	8	16	4
8	68	24	20	11	6	8	66	88	48	36	64
9	74	26	21	17	8	9	136	153	72	64	81
10	80	32	24	23	14	12	322	276	168	196	144
Total	570	180	120				956	900	524	576	504

$$\overline{X1} = \frac{\Sigma X1}{n}; \overline{X2} = \frac{\Sigma X2}{n}; \overline{Y} = \frac{\Sigma Y}{n}$$

$$\overline{X1} = \frac{180}{10} = 18; \overline{X2} = \frac{120}{10} = 12; \overline{Y} = \frac{570}{10} = 57$$

$$\beta_1 = \frac{(\Sigma x_1 y)(\Sigma x_2^2) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{(986)(504) - (900)(524)}{(576)(504) - (524)^2} = 0.65$$

$$\beta_2 = \frac{(\Sigma x_2 y)(\Sigma x_1^2) - (\Sigma x_1 y)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{(900)(576) - (956)(524)}{(576)(504) - (524)^2} = 1.11$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X1} - \beta_2 \overline{X2}$$

$$= 57 - (0.65 * 18) - (1.11 * 12) = 31.98$$

Regression Model;

$$Y = 31.98 + 0.65X1 + 1.11X2$$

Self-Assessment Exercise 3

Discuss 5 assumptions of multiple regression

1.6 Summary

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

In this unit, multiple regression model is given as

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$$

$$\hat{\beta}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Where $\hat{\beta}_1$ measures the change in Y for a unit change in X_1, while holding X_2 constant

$\hat{\beta}_2$ measure change in Y per units change in X_2 holding X_1 constant.

Tutor Marked Assignment

i. Explain what multiple regression is all about.

1.7 References/ Further Readings

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

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

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.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Multiple regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple regression is to model the linear relationship between your independent (explanatory) and dependent (response) variables.

A multiple regression model with four independent variables might look something like this:

$Y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \beta4X4 + \varepsilon$

Here:

- Y is the dependent variable that we're trying to explain or predict.
- β0 is the y-intercept of the model, it represents the predicted value of Y when all independent variables X1, X2, X3, X4 are zero.
- β1, β2, β3, and β4 are the regression coefficients for the independent variables X1, X2, X3, and X4 respectively. They represent the change in Y associated with a one-unit change in the respective independent variables, assuming all other variables are held constant.
- X1, X2, X3, X4 are the independent variables (predictors) that we are using to predict Y.
- ε is the error term, it represents the difference between the actual and predicted values of Y, it's assumed to be normally distributed with a mean of zero.

The regression coefficients (β0, β1, β2, β3, β4) are estimated using a method called least squares, which minimizes the sum of squared residuals (the differences between the observed and predicted values of the dependent variable).

This model allows you to examine the relationship between each independent variable and the dependent variable while controlling for the effects of the other independent variables. This means you can isolate the contribution of each independent variable to the prediction of the dependent variable.

Answer to Self- Assessment 2

Multiple regression is a powerful statistical method used for several purposes:

1. **Prediction**: One of the primary uses of multiple regression is to predict the value of the dependent variable based on the values of multiple independent variables. This can be particularly useful in fields such as finance, economics, medicine, and

social sciences where future outcomes need to be predicted based on a set of predictors.

2. **Variable selection**: Multiple regression can be used to identify the most important predictors out of a larger set. By comparing the standardized regression coefficients, it's possible to see which variables have the most impact on the dependent variable.
3. **Trend identification**: Multiple regression can be used to identify trends and relationships between variables. The signs and sizes of the regression coefficients can indicate the direction and strength of the relationship between each independent variable and the dependent variable.
4. **Control for confounding variables**: Multiple regression allows researchers to control for confounding variables. This means examining the effect of a particular independent variable on the dependent variable, while holding all other independent variables constant.
5. **Testing Hypotheses**: Multiple regression can be used to test hypotheses about the relationships between the dependent variable and one or more independent variables. For example, a researcher might want to know whether a particular independent variable has a significant effect on the dependent variable after controlling for other variables.
6. **Estimate the impact of changes**: With multiple regression, we can estimate how much a change in one or more independent variables will affect the dependent variable. This helps in evaluating the effectiveness of certain decisions or interventions.

Answer to Self- Assessment 3

Multiple regression analysis requires several key assumptions to ensure accurate and reliable results. Here are five of these assumptions:

1. **Linearity**: The relationship between the independent and dependent variables is linear. This means that the line of best fit through the data points is a straight line rather than a curve.
2. **Independence**: The residuals (i.e., the differences between the observed and predicted values of the dependent variable) are independent. In other words, the residuals from one prediction have no effect on the residuals from another. This is a key assumption of the general linear model.
3. **Homoscedasticity**: This assumption means that the variances around the line of best fit remain constant as you move along the line. If the variances around the line are not constant, we refer to this as heteroscedasticity.
4. **Normality**: The residuals of the model are normally distributed. This assumption can be checked by creating a histogram of the residuals or a Q-Q plot.

5. **No Multicollinearity**: Multicollinearity occurs when the independent variables are highly correlated with each other. This can make it difficult to determine the effect of each independent variable on the dependent variable and can inflate the variance of the regression coefficients, making them unstable and difficult to interpret. It can be checked using variance inflation factor (VIF) or tolerance values.

Violation of these assumptions can lead to problems with the validity and reliability of the regression model. However, various diagnostics can be used to check these assumptions, and there are often ways to fix violations, such as by transforming variables or using different types of regression models.

UNIT TWO: CORRELATION COEFFICIENT

1.1 Introduction

It is assumed that you must have read unit 1 of this module that talks about multiple regression, a detailed understanding of this will be assumed. this unit is building on the unit 1 of this module. This unit will be dealing with thorough explanation of the parameters involved in the regression analysis.

1.2 Learning Outcomes

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

- Explain the meaning of Correlation co-efficient

- Assumptions of correlation
- Explain the goals of correlations
- evaluate how to analyse correlation co-efficient

1.3 Correlation Defined

Correlation analysis is used to describe the strength and direction of the linear relationship between two variables. Correlation is a statistical concept that measures the strength and direction of the linear relationship between two variables. It quantifies how changes in one variable are associated with changes in another variable. Correlation is often denoted by the symbol "r" and ranges from -1 to 1.

Correlation, like covariance, is a measure of the degree to which any two variables vary together. In other words, two variables are said to be correlated if they tend to simultaneously vary same direction. If both the variables tend to increase (or decrease) together, the correlation is said to be direct or positive, e.g. the length of an iron bar will increase as the temperature increases.

Here are some key points about correlation:

1. **Strength of Relationship**: The magnitude of the correlation coefficient indicates the strength of the relationship between the variables. A correlation coefficient of 1 or -1 indicates a perfect positive or negative linear relationship, respectively. A correlation coefficient close to 0 indicates a weak or no linear relationship.

2. **Direction of Relationship**: The sign of the correlation coefficient (+ or -) indicates the direction of the relationship. A positive correlation coefficient indicates a positive linear relationship, meaning that as one variable increases, the other variable tends to increase as well. In contrast, a negative correlation coefficient indicates a negative linear relationship, meaning that as one variable increases, the other variable tends to decrease.
3. **Linear Relationship**: Correlation measures the strength and direction of the linear relationship between variables. It does not capture non-linear relationships. If the relationship between variables is not linear, the correlation coefficient may not accurately represent the association.
4. **Range of Values**: The correlation coefficient ranges from -1 to 1. A value of -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.
5. **Correlation Does Not Imply Causation**: It's important to note that correlation does not imply causation. Just because two variables are correlated does not necessarily mean that changes in one variable cause changes in the other variable. Correlation simply indicates a statistical association between the variables.
6. **Pearson Correlation vs. Spearman Correlation**: The most commonly used correlation coefficient is the Pearson correlation coefficient, which measures the linear relationship between two continuous variables. If the variables are not normally distributed or the relationship is not strictly linear, the Spearman

correlation coefficient can be used, which measures the monotonic relationship (the direction of change but not the magnitude).

7. **Correlation Matrix**: In cases where you have more than two variables, a correlation matrix can be constructed to examine the relationships between all pairs of variables simultaneously. This matrix displays the correlation coefficients between each pair of variables.

Correlation is widely used in various fields such as social sciences, economics, finance, and data analysis. It helps in understanding the relationship between variables, identifying patterns, and guiding further analysis. However, it's important to interpret correlation carefully and consider other factors before making causal claims or drawing conclusions based solely on correlation.

1.4 Goals and Assumptions of Correlation

1.4.1 Goals of Correlation

Correlation is a fundamental statistical concept used extensively across numerous disciplines. Here are some key goals when using correlation in statistical analysis:

1. **Identifying Relationships**: The primary goal of calculating correlation is to quantify the degree and direction of association between two variables. The correlation coefficient tells us whether the variables are related and if so, whether the relationship is positive or negative.

2. **Variable Selection**: In building predictive models, correlation analysis can help in selecting the most meaningful variables that have strong relationships with the response variable. This can improve the accuracy of the model and also make it more interpretable.
3. **Understanding Data**: Correlation analysis can be used as part of exploratory data analysis to understand the relationships between variables in a dataset. Identifying these relationships can help provide insights into the data, guiding further analysis.
4. **Multicollinearity Detection**: In regression analysis, multicollinearity (when predictor variables are highly correlated with each other) can be a problem because it can make the model unstable and the estimates of the coefficients unreliable. By identifying highly correlated variables, we can reduce multicollinearity, thereby improving the model.
5. **Testing Hypotheses**: Correlation can also be used for hypothesis testing, to test whether the observed correlation differs significantly from zero. This can give us more confidence in the existence of a relationship between the variables.
6. **Predictive Analysis**: Although correlation doesn't imply causation, it can still be valuable in predictive analytics. Two variables that are strongly correlated can often be used to predict one another, even if the relationship between them is not causal.

Remember, correlation is a measure of linear association between two variables. It does not imply causation, meaning we cannot say that change in one variable causes change in

the other based solely on correlation. Furthermore, correlation does not capture non-linear relationships, and correlation values close to 0 don't necessarily mean that there's no relationship between the variables, just that there's no linear relationship.

1.4.2 Assumptions of Correlation

Here are these assumptions:

1. **Linearity**: The assumption of linearity specifies that there is a linear relationship between the two variables. The Pearson correlation coefficient (r) measures the strength and direction of the linear relationship between two variables. If the relationship is not linear, the correlation coefficient may not capture the strength and direction of the relationship accurately. Non-linearity can be checked by scatterplots.
2. **Bivariate Normality**: This assumption requires that the pair of variables follows a bivariate normal distribution. In other words, for any fixed value of one variable, the other variable is normally distributed, and vice versa. Violation of this assumption doesn't render the correlation coefficient invalid, but it can make significance tests and confidence intervals inaccurate.
3. **Homoscedasticity**: Homoscedasticity means that the variability in one variable is the same across all values of another variable. If the variance of one variable differs substantially for different values of the other variable (heteroscedasticity), then the correlation coefficient might not be a good summary of the relationship.

4. **Independence of Observations**: The observations are assumed to be independent of each other. This assumption is violated when there is autocorrelation or when data are collected over time (time series data) or space.

Remember, correlation doesn't imply causation, meaning we cannot say that change in one variable causes change in the other based solely on correlation. If the assumptions of correlation are violated, you might still be able to use other methods to quantify the relationship between variables, such as Spearman's rank correlation (which doesn't require linearity or homoscedasticity) or regression methods that can model non-linear relationships.

1.5 Estimation and Explanation of Correlation Coefficient

The sample linear correlation coefficient for n pairs of observations (Xi, Yi) usually denoted by the letter r, is defined by:

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \sum(Y - \overline{Y})^2}}$$

The population correlation co-efficient for a bivariate distribution, denoted by ρ, has already been defined as:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{\left[\sum X^2 - (\sum X)^2/n\right]\left[\sum Y^2 - (\sum Y)^2/n\right]}}$$

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

Or $r = \frac{\Sigma xy}{\sqrt{\left[\Sigma x^2\right]\left[\Sigma y^2\right]}}$; where $x = X - \overline{X}$; and $y = Y - \overline{Y}$

This is a more convenient and useful form, especially when $\overline{X}$ and $\overline{Y}$ are not integers.

The linear correlation co-efficient, is also the square root of the linear co-efficient of determination, r^2.

Self-Assessment Exercise 1

Define correlation

Worked Examples

Use this Table to answer the questions below:

Using Table M3.2.1, compute the correlation coefficient between the variables Revenue (*Y*) and Tax (*X*).

Table M3.2.1: Data for the Correlation Estimation

n	Y (Revenue)	X (Tax)
1	40	8
2	90	18
3	50	10
4	60	12
5	10	2
6	20	4
7	50	6
8	120	20
9	80	14
10	90	16

Table M3.2.2: Estimated Correlation Results

n	Y	X	XY	X^2	Y^2	$x = X - \bar{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
1	40	8	320	64	1600	-3	-21	9	441	63
2	90	18	1620	324	8100	7	29	49	841	203
3	50	10	500	100	2500	-1	-11	1	121	11
4	60	12	720	144	3600	1	-1	1	1	-1

5	10	2	20	4	100	-9	-51	81	2601	459
6	20	4	80	16	400	-7	-41	49	1681	287
7	50	6	300	36	2500	-5	-11	25	121	55
8	120	20	2400	400	14400	9	59	81	3481	531
9	80	14	1120	196	6400	3	19	9	361	57
10	90	16	1440	256	8100	5	29	25	841	145
Total	**610**	**110**	**8520**	**1540**	**47700**	**0**	**0**	**330**	**10490**	**1810**

$n = 10; \Sigma Y = 610; \Sigma X = 110; \Sigma x^2 = 330; \Sigma y^2 = 10,490; \Sigma xy = 1810;$
$\Sigma XY = 8,520; \Sigma X^2 = 1540; \ \Sigma Y^2 = 47,700.$

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}} \text{ or } r = \frac{\Sigma xy}{\sqrt{\left[\Sigma x^2\right]\left[\Sigma y^2\right]}}$$

$$r = \frac{(10)(8520) - (610)(110)}{\sqrt{(10)(1540) - 12100}\sqrt{(10)(47700) - 372100}} = 0.975 \qquad (11.6)$$

$$\text{or } r = \frac{1810}{\sqrt{330}\sqrt{10490}} = 0.975 \qquad (11.7)$$

The value of 0.975 shows that there is a strong (or large) and positive correlation between tax and revenues.

Self-Assessment Exercise 2

Discuss the assumptions of correlation

1.6 Summary

In the course of our discussion the following formulas were made use of the population correlation co-efficient for a bivariate distribution, denoted by ρ, defined as:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$r = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{\left[\sum X^2 - (\sum X)^2/n\right]\left[\sum Y^2 - (\sum Y)^2/n\right]}}$$

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

Or $r = \frac{\Sigma xy}{\sqrt{\left[\Sigma x^2\right]\left[\Sigma y^2\right]}}$; where $x = X - \overline{X}$; and $y = Y - \overline{Y}$

Tutor Marked Assignment

Given that Y and X, derive the expression for correlation coefficient.

1.7 References /Further Readings

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

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

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.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Correlation is a statistical technique that is used to measure and describe the strength and direction of the relationship between two variables.

A correlation coefficient, typically denoted as 'r', quantifies the degree to which two variables are related. Correlation coefficients range from -1 to 1.

- If r = 1, there's a perfect positive correlation. This means that as one variable increases, the other also increases.
- If r = -1, there's a perfect negative correlation. This means that as one variable increases, the other decreases.

- If r = 0, there's no linear correlation. This means that there's no linear relationship between the variables.

Answer to Self- Assessment 2

When conducting a correlation analysis, there are several assumptions that should ideally be met to ensure the accuracy and reliability of the results:

1. **Linearity**: The assumption of linearity specifies that the relationship between the two variables is linear. In other words, the pattern of the data points should roughly follow a straight line when plotted on a scatter plot. Non-linear relationships will not be accurately captured by a standard correlation coefficient.
2. **Independence of observations**: Each pair of observations (X, Y) should be independent from all other pairs of observations. This means that the data should not be repeated measures or have any form of inherent ordering.
3. **Homoscedasticity**: This means that the spread of residuals (variances) should be consistent for all values of your independent variables. If the variances are unequal across the variables (heteroscedastic), the correlation coefficient may not accurately reflect the relationship.
4. **Bivariate normal distribution**: The pair of variables should follow a bivariate normal distribution. This means that each variable follows a normal distribution and that the form of the relationship between the two variables is linear.

Violation of these assumptions can lead to biased or misleading results. There are different types of correlation coefficients available that make different assumptions and may be more suitable for different types of data, such as Spearman's rho for ordinal data or data with non-linear relationships.

UNIT THREE: MULTIPLE CORRELATION CO-EFFICIENT AND COEFFICIENT OF DETERMINATION

1.1 Introduction

This unit is an extension of unit one and two of this module. This unit requires thorough knowledge of unit 1 and unit two. In this unit we are going to look at multiple Correlation Coefficients (R) and multiple coefficient of determination (R^2).

1.2 Learning Outcomes

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

- Estimate multiple correlation coefficient (r)
- Estimate coefficient of determination
- Interprete your answer i.e. statistical interpretation

1.3 Multiple Correlation Coefficient (R) and Coefficient of Determination (R^2)

1.3.1 Multiple Correlation Coefficient (R)

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. Its value can be positive or negative; multiple correlation coefficients is always taken to be positive. In practice the multiple correlation coefficients is of little importance.

The multiple correlation coefficient, often denoted as R, is a measure used in multiple regression to gauge the strength and direction of the linear relationship between one dependent variable and several independent variables.

In the context of multiple regression, R is defined as the correlation between the observed outcome values and the values predicted by the model. Like the simple correlation coefficient, R ranges from -1 to 1, with -1 indicating a perfect negative linear

relationship, 1 indicating a perfect positive linear relationship, and 0 indicating no linear relationship.

However, in most contexts of multiple regression, R is non-negative and is the square root of R^2 (the coefficient of determination), making R fall between 0 and 1. This is because, in a multiple regression, the prediction of the dependent variable is based on the combination of the independent variables, which typically results in predicted values that are more closely associated with the observed outcome values than any individual independent variable.

Here are a few important aspects of the multiple correlation coefficient:

1. **Strength of Relationship**: A higher absolute value of R indicates a stronger relationship between the dependent variable and the set of independent variables.
2. **Direction of Relationship**: As mentioned earlier, in most multiple regression contexts, R is non-negative. However, in the general sense, if R is considered to have a sign, a positive value indicates that, on average, higher predicted values correspond with higher observed values. A negative value indicates that higher predicted values correspond with lower observed values.
3. **Perfect Correlation**: An R of 1 indicates that the model predicts the observed values perfectly with no error (although this rarely happens in practice).
4. **No Correlation**: An R of 0 suggests that the model explains none of the variability in the outcome variable.

5. **R and R^2**: R is the square root of R^2. So, if R^2 (the coefficient of determination) is the proportion of the variance in the dependent variable that can be predicted from the independent variable(s), R can be interpreted as the correlation between the observed and predicted values of the dependent variable.

1.3.2 Coefficient of Determination (R^2)

The Coefficient of Determination, often denoted as R^2, is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

The more meaningful coefficient is the coefficient of determination R or r^2. Coefficient of determination (R) is defined as the proportion of the total variation in Y explained by the multiple regression of Y on X1 and X2. 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 X1 and X2. The coefficient of determination is denoted by R or r^2. Because of the relative importance of coefficient of determination (R^2) we concentrate more on the coefficient of determination (R^2).

It is a key output of regression analysis and is commonly used to gauge the goodness-of-fit of a regression model.

The coefficient of determination ranges between 0 and 1, where:

- An R^2 of 1 indicates that the regression predictions perfectly fit the data. All changes in the dependent variable are completely explained by changes in the independent variable(s).
- An R^2 of 0 indicates that the regression model explains none of the variability of the dependent variable around its mean. Essentially, the independent variable(s) provide no predictive capability for the dependent variable in this model.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.20 means that 20% of the dependent variable's variability can be explained by the model.

Here are some important points to remember about R^2:

- **R^2 is not an error measure**: R^2 tells you the proportion of variance explained by the model, but it doesn't tell you if the predictions from the model are close to the actual values. The root-mean-square error (RMSE) or mean absolute error (MAE) can be used to quantify prediction error.
- **R^2 always increases as more predictors are added to the model**: Even if those predictors are not truly related to the response, adding them to the model will increase the R^2 value. This could lead to overfitting, where the model fits the sample data too closely and performs poorly on new data. Adjusted R^2 can be used to account for this by penalizing the addition of unnecessary predictors.
- **R^2 doesn't indicate whether a model is adequate**: You could have a low R^2 value for a very good model, or a high R^2 value for a model that doesn't fit the

data. It's also possible to have a high R^2 for a model that's not appropriate for the data, so it's important to also use other metrics and plots to assess model adequacy.

- **R^2 doesn't tell you if the coefficients and predictions are biased**: A high R^2 isn't a guarantee of unbiased estimates. Always check the residual plots to verify the assumptions of your regression model.

1.4 The Importance of Correlation and Coefficient of Determination

Correlation and the Coefficient of Determination (R^2) are both crucial statistical measures used in many fields, such as economics, psychology, biology, and business. They provide valuable insights about the relationships among variables and how well certain models can predict outcomes.

1.4.1 The Importance of Correlation

Correlation is important for the following reasons:

1. **Relationship Between Variables**: It provides an initial understanding of how strongly pairs of variables are related. It tells us about the direction (positive or negative) and degree (strong or weak) of relationship between two variables.
2. **Variable Selection**: In predictive modeling, correlation analysis can help identify which predictors have strong relationships with the outcome variable, and therefore should be included in the model.

3. **Multicollinearity Detection**: High correlation between independent variables (multicollinearity) can cause problems in regression analysis, as it may lead to unstable estimates of regression coefficients. Identifying and addressing multicollinearity is therefore an important step in regression modeling.
4. **Predictive Analytics**: Even if correlation doesn't imply causation, knowing that variables are correlated can still be useful in predictive analytics. For instance, if sales and advertising spend are highly correlated, we might be able to use advertising spend to predict sales.

1.4.2 Importance of Coefficient of Determination

The **Coefficient of Determination (R^2)** is important for the following reasons:

1. **Model Fit**: R^2 is a measure of how well the regression model fits the observed data. A higher R^2 means that the model explains a larger proportion of the variance in the outcome variable, which usually means the model has a better fit.
2. **Predictive Power**: R^2 indicates the percentage of the dependent variable's variation that the independent variables explain collectively. It provides a measure of the model's predictive power.
3. **Model Comparison**: R^2 can be used to compare different regression models. For example, if we fit different models to the same data, the model with the higher R^2 might be a better choice, as it explains more of the variance in the outcome.

4. **Interpretability**: R^2 provides an intuitive measure of the overall strength of the relationship between the dependent variable and a set of independent variables.

Self-Assessment Exercise 1

Explain the importance of correlation

1.5 Estimation of Coefficient of Determination

$R^2 = 1 - (SSR / SST)$

where:

- SSR (Sum of Squares of the Residuals) is the sum of the squares of the difference between the predicted Y value (based on the regression equation) and the mean Y value. It represents the error of the prediction from the model.
- SST (Total Sum of Squares) is the sum of the squares of the difference between the observed Y value and the mean Y value. It represents the total variability in the outcome variable.

Conceptually, it is often written as:

Coefficient of determination: $R^2 = r^2 = \frac{\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2}{\Sigma y^2}$

$$R = r = \sqrt{R^2} = \sqrt{r^2} = \sqrt{\frac{\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2}{\Sigma y^2}}$$

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.

Self-assessment exercise 2

The coefficient of determination usually lies between------ and ------

1.5.1 Worked Example

From our calculation in unit 1 of this module especially the table in unit 1, we are going to derive our values from the table in unit 1.

Table M3.1.2: Calculating the coefficient of regression

				y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2	y^2
Obs	Y	$X1$	$X2$	$Y-\overline{Y}$	$X1-\overline{X1}$	$X2-\overline{X2}$						
1	40	6	4	-17	-12	-8	204	136	96	144	64	289
2	44	10	4	-13	-8	-8	104	104	64	64	64	169
3	46	12	5	-11	-6	-7	66	77	42	36	49	121
4	48	14	7	-9	-4	-5	36	45	20	16	25	81
5	52	16	9	-5	-2	-3	10	15	6	4	9	25
6	58	18	12	1	0	0	0	0	0	0	0	1
7	60	22	14	3	4	2	12	6	8	16	4	9
8	68	24	20	11	6	8	66	88	48	36	64	121
9	74	26	21	17	8	9	136	153	72	64	81	289
10	80	32	24	23	14	12	322	276	168	196	144	529
Total	570	180	120				956	900	524	576	504	1634

$$\overline{X1}=\frac{\Sigma X1}{n};\overline{X2}=\frac{\Sigma X2}{n};\overline{Y}=\frac{\Sigma Y}{n}$$

$$\overline{X1}=\frac{180}{10}=18;\overline{X2}=\frac{120}{10}=12;\overline{Y}=\frac{570}{10}=57$$

$$R^2=r^2=\frac{\hat{\beta}_1\Sigma yx_1+\hat{\beta}_2\Sigma yx_2}{\Sigma y^2}=\frac{0.65(956)+1.11(900)}{1634}=\frac{621.4+999}{1634}=0.992=99.2\%$$

This implies that the explanatory variable (x_1 and x_2) can only account for 99.2% variation in variable Y i.e. both x_1 and x_2 contributes 99.2% to the explanation of the variation in Y.

$$R = r = \sqrt{R^2} = \sqrt{r^2} = \sqrt{\frac{\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2}{\Sigma y^2}}$$

$$R = r = \sqrt{0.992} = 0.996$$

Self-Assessment Exercise 2

When r^2=0.85, what is the economic interpretation of this?

1.6 Summary

In the course of our discussion of this unit, you have learnt about the following:

- Concept of multiple correlation

- Coefficient of determination

- Estimation of R^2 & r

- Interpretation of r & r^2

In our discussion of this unit we defined coefficient of determination R^2 as:

$$R^2 = r^2 = \frac{\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2}{\Sigma y^2}$$

$$R = r = \sqrt{R^2} = \sqrt{r^2} = \sqrt{\frac{\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2}{\Sigma y^2}}$$

The closer the r^2 is to 1, the better

Tutor Marked Assignment

i. The measure of proportion or percentage of variation in Y explained by the

explanatory variable $x_1 \ldots x_n$ jointly is given by ----------

ii. Multiple coefficient of determination measures----------

1.7 References/ Further Readings

Damordar, N. G., Dawn, C. P. and Sangeetha, G. (2012): Basic Econometrics (5th edition) Tata McGraw Hill Education Private Limited. New Delhi, India.

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

Oyesiku O.K. and Omitogun O. (1999): Statistics for social and management science (2nd edition). Higher Education Books Publisher, Lagos.

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

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Correlation is a crucial statistical concept that has a wide variety of applications. Here are a few reasons why it's important:

1. **Understanding Relationships**: Correlation provides an important tool for discovering and quantifying the relationships between variables. By understanding these relationships, researchers can identify patterns and connections in the data.
2. **Predictive Analytics**: Correlation is foundational to predictive analytics and machine learning. Understanding the relationship between various features and a target variable is a common task in developing predictive models. If two variables are strongly correlated, we can predict one variable from the other with a higher degree of accuracy.
3. **Risk Management**: In finance and investing, understanding the correlation between different assets or investments can help in creating a diversified portfolio that minimizes risk.
4. **Hypothesis Testing**: Correlation can be used to test hypotheses about the relationship between two variables. For example, if a researcher hypothesizes that studying more leads to higher test scores, they could use correlation to measure the strength of the relationship between study time and test scores.
5. **Multi-collinearity in Regression Analysis**: In multiple regression analysis, examining the correlation between each pair of variables can help identify

multicollinearity (when two or more variables are highly correlated). This is important because multicollinearity can affect the interpretability of the regression coefficients and the model's performance.

6. **Data Reduction**: If two variables are highly correlated, they are likely conveying similar information, and the dimensionality of the data can be reduced by removing one or more of the correlated variables.

It's important to remember that correlation does not imply causation. Just because two variables are correlated, it doesn't mean that one variable causes the other to occur. They could be related due to a third factor, or the correlation could be a coincidence.

Answer to Self- Assessment 2
The coefficient of determination usually lies between 0 and 1

Answer to Self- Assessment 3

The coefficient of determination, denoted as r^2, is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well observed outcomes are replicated by the model.

If $r^2 = 0.85$, it implies that 85% of the variation in the dependent variable can be explained by the independent variable(s) included in the model. This is generally considered a strong relationship, and the model is a good fit for the data. In all cases, it is also crucial to evaluate the economic significance and theoretical consistency of the independent variables, beyond the statistical significance indicated by a high r^2 value.

UNIT FOUR: TEST OF SIGNIFICANCE

1.1 Introduction

This unit completes this module, so it is required that thorough knowledge of unit one to unit three is very germane. It is important to test for the significance of the value of the regression, Coefficients, and the level of prediction or explanation given by the regression equation.

1.2 Learning Outcomes

At the end of this unit the student(s) should be able to calculate and understand:

- The calculation of F-statistics (Fcal)
- Check the corresponding tabulated value of F.statistics through its degree of freedom.
- Compare the F-statistics and F-tab
- Interpret your answer

1.3 Test of Significance Defined

Test of significance is a procedure by which sample results are used to verify truity of falsity of a null hypothesis. The key idea behind test of significance is that of a test statistics (estimator and the sampling distribution of such a statistics under the null hypothesis).

The decision to accept or reject H_o is made on the basis of the test statistics obtained from the data at hand.

Here are the key steps involved in a test of significance:

1. **State the Hypotheses**: The first step in a test of significance is to set up the null hypothesis (H0) and the alternative hypothesis (H1 or Ha). The null hypothesis usually represents a skeptical perspective, or status quo, while the alternative hypothesis represents what we are trying to prove.
2. **Formulate an Analysis Plan**: The analysis plan describes how to use the sample data to accept or reject the null hypothesis. It should specify the significance level (typically 0.05), the test statistic, and the method for calculating it.
3. **Analyze Sample Data**: Using sample data and the plan formulated in step 2, calculate the value of the test statistic.
4. **Interpret the Results**: If the test statistic falls in the critical region, reject the null hypothesis in favor of the alternative hypothesis. If not, fail to reject the null hypothesis. The critical region is determined by the significance level (α); if a test of hypothesis has a significance level of 0.05, the data must provide evidence against the null hypothesis so strong that it would happen less than 5% of the time due to random chance alone.

The result of a significance test is expressed in terms of a p-value, which is the probability of obtaining a result as extreme as, or more extreme than, the observed data, assuming that the null hypothesis is true. If the p-value is less than or equal to the significance level, we reject the null hypothesis.

It's important to note that failing to reject the null hypothesis doesn't mean that the null hypothesis is true, just that we don't have enough evidence to conclude otherwise.

Similarly, rejecting the null hypothesis doesn't prove the alternative hypothesis; it just suggests that it may be true and warrants further investigation.

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).

Self-Assessment Exercise 1

The key steps involved in a test of significance are----

1.4 Estimation of Test of Significance

The F-statistics or F-ratio for the test of significance can be written as:

$$H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$$

$$F_{k-1,n-k} = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

Table M3.4.1: Anova Tables for 3-Variables Regression

Source of Variation	Sum of Squares	Degree of freedom (df)	Mean Square or Variance Estimate	F-value
Due to Regression (ESS)	$\hat{\beta}_1 \Sigma yx_1 + \hat{\beta}_2 \Sigma yx$	$v_1 = K-1$	$\frac{\hat{\beta}_1 \Sigma yx_1 + \hat{\beta}_2 \Sigma yx_2}{K-1}$	$F^* = \frac{\hat{\beta}_1 \Sigma yx_1 + \hat{\beta}_2 \Sigma yx_2 / (K-1)}{\Sigma \mu^2 / N-K}$ $= \frac{R^2/(K-1)}{1-R^2/N-K}$
Due to Residual (RSS)	$\Sigma \mu^2$	$v_2 = N-K$	$\frac{\Sigma \mu^2}{N-K}$	F-tables with $v_1 = K-1,$ $v_2 = N-K$ Degrees of freedom
Total	Σy^2	$N-1$		

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.

Decision Criteria: If the F-ratio calculated exceeds the critical F-value from the table at the alpha percent level of significance we reject Ho; otherwise, do not reject it. Alternatively, if the F-cal of the observed F is sufficiently low, accept Ho.

Self-Assessment Exercise 2

State the decision criteria for test of significance?

1.5 Worked Example

Example 1:

Given the regression model Y = B0+ BX1 + B2 X2+U, how would you state the null hypothesis to test for test for significance of x1 and x2 on Y

Solution:

In a multiple regression model, a hypothesis test typically aims to determine whether the coefficients of the predictor variables (in this case, X1 and X2) are significantly different from zero. The null and alternative hypotheses would be set up as follows:

For X1:

- Null Hypothesis (H0): The coefficient B1 equals 0. This means that X1 has no effect on Y, assuming that other predictors are in the model. In other words, B1 = 0.
- Alternative Hypothesis (H1 or Ha): The coefficient B1 does not equal 0. This means that X1 does have an effect on Y, again assuming other predictors are in the model. In other words, B1 ≠ 0.

For X2:

- Null Hypothesis (H0): The coefficient B2 equals 0. This means that X2 has no effect on Y, assuming that other predictors are in the model. In other words, B2 = 0.
- Alternative Hypothesis (H1 or Ha): The coefficient B2 does not equal 0. This means that X2 does have an effect on Y, again assuming other predictors are in the model. In other words, B2 ≠ 0.

The hypotheses are then tested using statistical software, which will provide p-values for each coefficient. If the p-value for a given coefficient is less than a certain significance level (typically 0.05), you would reject the null hypothesis and conclude that the predictor has a significant effect on the outcome. If the p-value is larger than the significance level, you would fail to reject the null hypothesis and conclude that there is not enough evidence to suggest the predictor has an effect on the outcome.

Remember that these tests are conditional on the other variables being in the model. For example, the test for B1 is really testing whether X1 has an effect after X2 (and any other predictors) have been taken into account.

Example 2:

Calculate the F-statistic using the Table below:

Table M3.4.2: Calculating F-statistic

				y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2	y^2
Obs	Y	$X1$	$X2$	$Y-\overline{Y}$	$X1-\overline{X1}$	$X2-\overline{X2}$						
1	40	6	4	-17	-12	-8	204	136	96	144	64	289
2	44	10	4	-13	-8	-8	104	104	64	64	64	169
3	46	12	5	-11	-6	-7	66	77	42	36	49	121
4	48	14	7	-9	-4	-5	36	45	20	16	25	81
5	52	16	9	-5	-2	-3	10	15	6	4	9	25
6	58	18	12	1	0	0	0	0	0	0	0	1
7	60	22	14	3	4	2	12	6	8	16	4	9
8	68	24	20	11	6	8	66	88	48	36	64	121
9	74	26	21	17	8	9	136	153	72	64	81	289
10	80	32	24	23	14	12	322	276	168	196	144	529
Total	570	180	120				956	900	524	576	504	1634

$$\overline{X1}=\frac{\Sigma X1}{n};\overline{X2}=\frac{\Sigma X2}{n};\overline{Y}=\frac{\Sigma Y}{n}$$

$$\overline{X1}=\frac{180}{10}=18;\overline{X2}=\frac{120}{10}=12;\overline{Y}=\frac{570}{10}=57$$

$$R^2=r^2=\frac{\hat{\beta}_1\Sigma yx_1+\hat{\beta}_2\Sigma yx_2}{\Sigma y^2}=\frac{0.65(956)+1.11(900)}{1634}=\frac{621.4+999}{1634}=0.992=99.2\%$$

Adopting this F-statistic method:

$$F=\frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

We have,

$$F = \frac{0.992/(3-1)}{(1-0.992)/(10-3)} = \frac{0.4960}{0.00114} = 435.08$$

$$F_{k-1,n-k} = F_{2,7} = 4.74$$

Since $F_{cal} > F_{tab}$, reject H_o and accept H_1.

Self-Assessment Exercise 3

What will be the decision criteria if $F_{cal} < F_{tab}$?

1.6 Summary

From our discussion on this unit, you have learnt about:

Definition test of significance; Estimation of test of significance; The interpretation of resulting. The meaning of ANOVA F-statistic and how it can be estimated.

Tutor Marked Assignment

Given the regression model $Y = B_0 + BX_1 + B_2 X_2 + U$, how would you state the null hypothesis to test for test for significance of x_1 and x_2 on Y.

1.7 References/ Further Readings

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

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

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.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Conducting a test of significance, also known as hypothesis testing, generally involves the following steps:

1. **Formulate the Hypotheses**: The first step in hypothesis testing is to set up the null hypothesis (H0) and the alternative hypothesis (Ha). The null hypothesis often represents a theory that has been put forward, either because it is believed to be true or because it is used as a basis for argument, but has not been proved. The alternative hypothesis is a statement that will be accepted in place of the null hypothesis if the null hypothesis is rejected.
2. **Choose the Significance Level (α)**: The significance level, also denoted as alpha or α, is a threshold for determining when to reject the null hypothesis. Common choices for α are 0.05 (5% chance) or 0.01 (1% chance).
3. **Select the Appropriate Test Statistic**: Depending upon the nature of the data and the reason for the analysis, choose an appropriate statistical test (t-test, chi-square test, ANOVA, etc.).
4. **Calculate the Test Statistic and Corresponding P-value**: Based on the selected statistical test, calculate the test statistic. Then, compute the P-value associated with the observed value of the statistic. The P-value is the probability of obtaining results as extreme as the observed results of a statistical hypothesis test, assuming the null hypothesis is true.
5. **Compare P-value with the Significance Level**: If the P-value is less than the chosen significance level (α), you reject the null hypothesis in favor of the alternative hypothesis. If the P-value is greater than or equal to α, you do not reject the null hypothesis.
6. **Draw Conclusions and Communicate the Results**: Based on the previous step, draw a conclusion about the hypotheses. The conclusion should be presented in the context of the problem, answering the original research question.

Answer to Self- Assessment 2

The decision criteria for a test of significance, also known as a hypothesis test, is usually based on the p-value and the chosen significance level, commonly denoted as alpha (α).

Here's the basic decision rule:

1. If the p-value ≤ α, then you reject the null hypothesis.
2. If the p-value > α, then you fail to reject (or "do not reject") the null hypothesis.

In other words, if your p-value is less than or equal to your significance level, you have evidence to suggest that your null hypothesis is unlikely to be true and should therefore be rejected.

Typically, α is set to 0.05, meaning there's a 5% chance you'll reject the null hypothesis when it is true. In other words, you're willing to accept up to a 5% chance of making a Type I error (rejecting the null hypothesis when it's true).

Remember, failing to reject the null hypothesis does not prove the null hypothesis is true. It just suggests that there's not enough evidence against it at your chosen significance level.

Answer to Self- Assessment 3

The decision criteria for a test of significance involving an F statistic (such as ANOVA) usually compares the calculated F-value (Fcal) to a critical F-value from the F-distribution table (Ftab), which is based on a chosen significance level and the degrees of freedom.

Here's the basic decision rule:

1. If Fcal > Ftab, then you reject the null hypothesis. This means the sample data provide enough evidence to conclude that there is a significant effect or difference.
2. If Fcal ≤ Ftab, then you fail to reject (or "do not reject") the null hypothesis. This means the sample data do not provide enough evidence to conclude that there is a significant effect or difference.

In your question, if Fcal < Ftab, then you would not reject the null hypothesis. This suggests that the variability among the group means can be attributed to random chance, rather than differences among the groups.

MODULE 4: TIME SERIES ANALYSIS

Unit 1: Time series and its components

Unit 2: Measurement and Quantitative estimation of time series

Unit 3: Index Numbers

UNIT ONE: TIME SERIES AND ITS COMPONENTS

1.1 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.

1.2 Learning Outcomes

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

- Understand or define time series
- Explain the component part of time series
- Understand methods of estimating time series
- Estimate and graphical representation of the trend

1.3 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 (Y_t) where Y_t 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.

Self-Assessment Exercise 1

The essence of time series is forecast. (true/false)

1.4 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

1.4.1 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.

GRAPHICAL REPRESENTATION OF SECULAR TREND

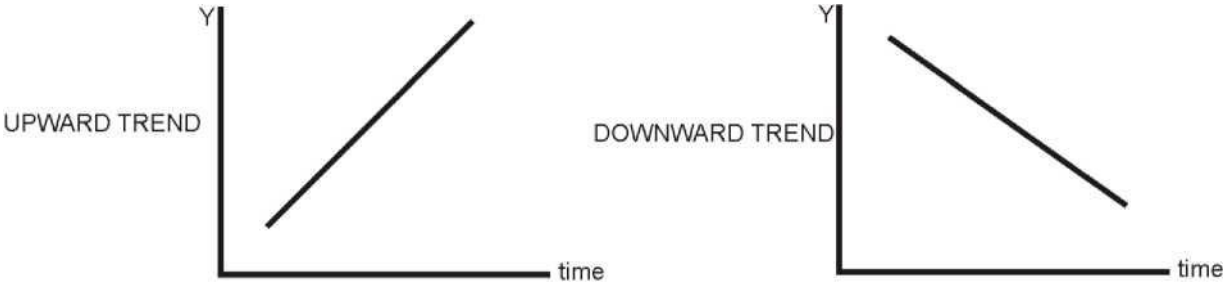

Fig. M4.1.1

1.4.2 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 xmas, new year or any festive period(s).

GRAPHICAL REPRESENTATION OF SEASONAL VARIATION

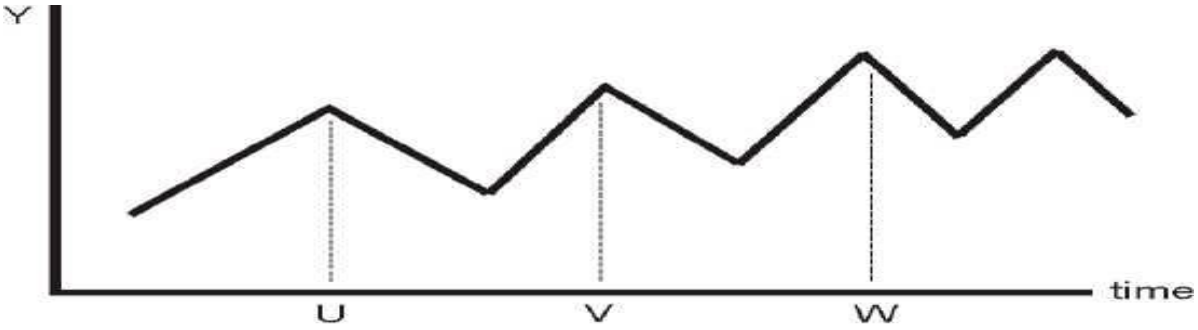

Fig. M4.1.2

1.4.3 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.

GRAPHICAL REPRESENTATION OF CYCLICAL VARIATION

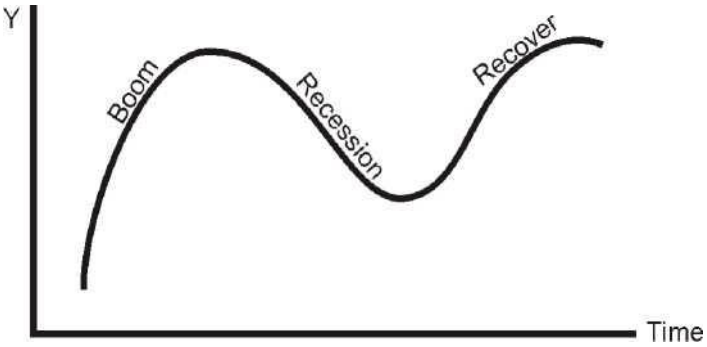

Fig. M4.1.3

1.4.4 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.

Graphical Representation Of Irregular Variation

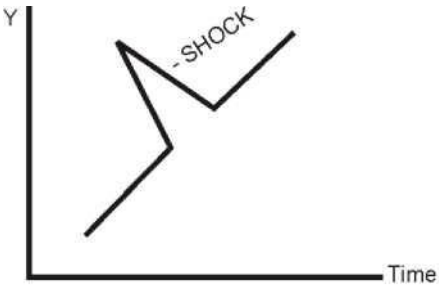

Fig. M4.1.4

Self-Assessment Exercise 2

The trend of secular trend can either be upward or downward. (true/false)

1.5 Goals of Time Series

Time series analysis is a branch of statistics that deals with data recorded over a period of time. It aims to understand the underlying structure and function that produce the observed data. Time series analysis has a wide range of applications, such as economic forecasting, sales forecasting, budgetary analysis, stock market analysis, yield projections, process control, and quality control, to name a few.

The main goals of time series analysis include:

1. **Descriptive Analysis**: To identify patterns, trends, and cycles in the data over time. This includes things like seasonal patterns, trends, and other regular cycles.
2. **Forecasting**: Perhaps the most important goal of time series analysis is forecasting. Once we understand the underlying patterns in the data, we can extrapolate them into the future to make predictions. This can be incredibly valuable for businesses trying to forecast sales, economists trying to predict economic indicators, or investors forecasting stock prices, etc.
3. **Intervention Analysis**: Sometimes we need to know the effect of a particular intervention on the time series data. For example, a company might want to know the effect of a marketing campaign on sales. This can be done using intervention analysis, which compares the time series before and after the intervention.
4. **Update Theory or System Understanding**: Time series analysis can lead to a better understanding of the system that generated the data. By developing a mathematical model that describes the growth pattern, we can increase our understanding of the system.

5. **Control**: For a system that can be controlled by changing input variables, time series analysis can be used to help optimize those inputs to achieve desired results.

1.6 Summary

In the course of our discussion on time series analysis you have learnt about

- Time series; Time series data; Component of time series and goals of time series.

Majorly time series decomposes itself into the following;

- Secular trend or long term movement

- Seasonal variation

- Cyclical variation

- Irregular or residual variation

Tutor Marked Assignment

Discuss the goals of time series

1.7 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): Basis Statistical Method. Tony Chriisto 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.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

True. One of the main reasons for analyzing time series data is to create models that can forecast future values based on historical data. Time series analysis can help to understand the underlying structure and patterns in the data, such as trend, seasonality, cyclical patterns, and irregular variations, which can be used to make more accurate predictions. However, it's worth noting that forecasting is not the only purpose of time series analysis. It's also used to understand the past behaviour of the series, detect anomalies, or test theoretical hypotheses about the causes of the observed behaviour.

Answer to Self- Assessment 2

True. The secular trend, often simply referred to as the "trend" in time series analysis, refers to the long-term movement in data over time. This can indeed be either upward (increasing over time) or downward (decreasing over time). A trend is a consistent, sustained direction in data. For example, if a company's sales are consistently increasing year over year, this would be an upward trend. Conversely, if a company's sales are consistently decreasing year over year, this would be a downward trend.

UNIT TWO: MEASUREMENT AND ESTIMATION OF TIME SERIES

1.1 Introduction

This unit is an extension of unit one of this module, here, you are going to learn more about estimation of time series data, also, a thorough understanding of unit one of this module is required for proper understanding of this module.

1.2 Learning Outcomes

At the of this unit, you should be able to

- estimate any time series data
- Understand methods of estimating time series.
- Estimate and do the graphical representation of the trends.

1.3 Measurement of Trend

There are several methods of estimating a trend in time series data. Here are a few:

1. **Graphical Method**: A simple plot of data over time can be a good starting point for identifying a trend. This is generally the most basic approach and provides a visual representation of possible trends in the data.
2. **Moving Averages**: If the time series data is influenced by seasonality, calculating moving averages can help identify the trend component. A moving average is calculated over a specific number of periods (for example, every five years) to smooth out short-term fluctuations and highlight long-term trends.
3. **Linear Regression**: Here, time is used as the independent variable to estimate the parameters of a linear trend line (i.e., $Y = a + bX$, where 'a' is the intercept, 'b' is the slope, and 'X' is time). The slope 'b' will indicate whether there is an increasing or decreasing linear trend over time.
4. **Exponential Smoothing**: This is a more advanced method that gives more weight to more recent observations. This method can be useful when recent data points are considered more useful for predicting future values.
5. **Decomposition Methods**: These techniques decompose a time series into trend, seasonal, and residual components. One common method is the STL (Seasonal and Trend decomposition using Loess) method, which uses a non-parametric method to estimate the trend and seasonal components.
6. **Polynomial Fitting**: If the trend is not linear, we might need to fit a polynomial to capture the trend accurately. Quadratic (second-degree) or cubic (third-degree) polynomials can sometimes provide a good fit to the data.

Each of these methods has its strengths and limitations, and the choice of method should be guided by the characteristics of the data and the purpose of the analysis. Some methods may be more suitable for data with seasonal patterns, while others may be more appropriate for data with non-linear trends. Furthermore, real-world data can often be noisy, and trend detection can be a challenging task that requires careful analysis and interpretation.

1.4 Least Square Method of Estimation

As defined in Preceding unit of this module, Time series refers to sequence of observations that gives information on how data has been behaving in the past. Estimation has to do with how time series are calculated, in this sub section we shall talk about three methods of estimation or measurement. These are method of least square, moving average and semi average method.

1.4.1 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

Formular 1

Trend equation of the least square method is given as

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = n\Sigma X + b\Sigma X^2$$

Where Σ = summation term derived from the data of the problem at hand

Σ x = sum of X values

Σ Y = sum of Y values

Σ xy = sum found by multiplying each Y by corresponding X value and adding the Products

n = no of items involved in the whole time series

The least square estimates of a and b are the solution to the normal equation above which can be solve simultaneously.

Formular 2

The general formular is as given below:

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

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

Where;

$$\bar{Y} = \frac{\Sigma Y}{n}$$
$$\bar{X} = \frac{\Sigma Y}{n}$$

Worked Example

Given the 7weeks information below about the sales of a company

Table M 4.1.1: Table Showing the Sales of a Company

Wk	Sales
1	15
2	25
3	38
4	32
5	40
6	37
7	50

Let X represents the weeks

Y represent the sales value

Table M4.1.2: 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
Σ X= 28	Σ XY=237	Σ XY=1079	ΣX^2=140

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

n=7

$$\bar{X} = \frac{28}{7} = 4$$

$$\bar{Y} = \frac{237}{7} = 33.857$$

$$b = \frac{7(1079) - 28(237)}{7(140) - 28^2} = \frac{7553 - 6636}{980 - 784} = \frac{917}{196} = 4.67857$$

$$a = \bar{Y} - \hat{b}\bar{X} = 33.857 - (4.67857)4 = 33.857 - 18.7142 = 15.1427$$

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

Y ≅ 62

For the 12th week

Y = 15.1427 + 4.6785(12)

Y = 15.1427 + 56.142

Y = 71.2847

Y = 71

1.5 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.

Example 2

The table below contained information about the actual sales of a company

Table M4.1.3: Table Showing the Sales of a Company

Month	Sale (units)
Jan	350
Feb	340
Mar	360
April	310
May	280
June	300
July	270
August	260
Sept	310
Oct	350
Nov	370
Dec	390

Prepare a 3 month moving average forecast

Solution

Table M4.1.4: 3- Month Moving Average Method Table of Analysis

Months	Sales	3months Moving total	3months moving average trend
Jan	350		
Feb	340	1050	350
Mar	360	1010	336.7
April	310	950	316.7

May	280	890	296.7
June	300	850	283.3
July	270	830	276.7
Aug	260	840	280
Sept	310	920	306.7
Oct	350	1030	343.3
Nov	370	1110	370
Dec	390		

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 3 s 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

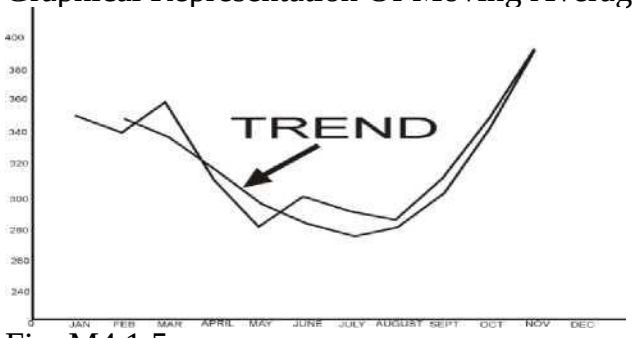

Fig. M4.1.5

Example 3

From the time series data below determine the trend on sales of a company

Table M4.1.5: Table showing sales of a company per quarter

Years	1	2	3	4
1982	600	820	400	720
1983	630	840	420	740
1984	670	900	430	760

Prepare a 4-quarter moving average

Solution

Table M 4.1.6: **4-point moving average table of analysis**

Year	Quarter	Sales	4 point moving total	4-point average moving or 4 quarterly average	2 point total or centre total	Moving average (trend)
1982	1	600	-	-	-	-
	2	820	-	-	-	-
			2540	635	-	
	3	400			1277.5	638.75
			2570	642.5	-	-
	4	720			1290	645
			2590	647.5	-	-
1983	1	630			1300	650
			2610	652.5	-	-
	2	840			1310	655
			2630	657.5	-	-
	3	420			1325	662.5
			2670	667.5	-	-
	4	740			1350	675
			2730	682.5		
1984	1	670			1367.5	683.75
			2740	685	-	-
	2	900			1375	687.5
			2760	690	-	-
	3	430	-	-	-	-
	4	760	-	-	-	-

Column 1 represent years

Column 2 represents quarter periods

Column 3 represents sales values

Column 4 is arrived at by adding the sales value in 4s

Column 5 is derived by dividing column 4 by the no's of quarters

Column 6 is the total of column 5 when taken in 2s

Column 7 is arrived at by dividing column 6 by 2

Graphical Representation Of Four Quarter Moving Average Trend

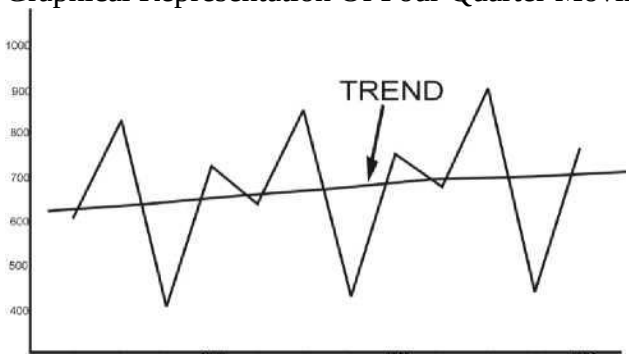

Fig. M4.1.6

1.5.1 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.

Example 4

Table M4.1.8: 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	4010	668.33

	4	720	-3		
1993	1	630	-2		
	2	840	-1		
	3	420	1		
	4	740	2		
1994	1	670	3		
	2	900	4		
	3	430	5		
	4	760	6	3,920	653.3

Column 4 represents the total of the 1st half and 2nd half.

Column 5 is arrived at by dividing the column 4 by 6 this represents the trend value, when plotted in a graph it gives the trend line.

Graphical Representation Of Semi Moving Average Trend

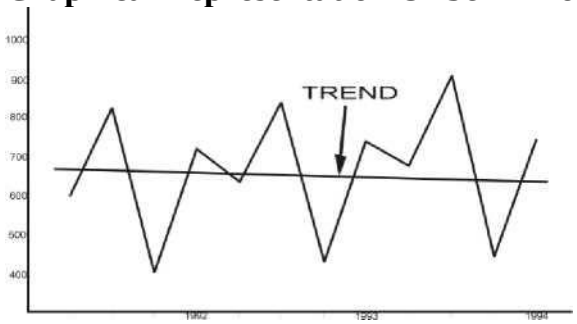

Fig. M4.1.7

Self-Assessment Exercise 1

State the least square equation of a time series data?

1.5.2 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.

Graphical Representation of Free Hand Method

Fig. M4.1.8

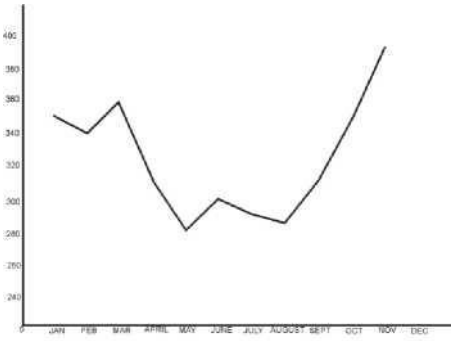

Fig. M4.1.8

1.6 Summary

In the course of our discussion on estimation of time series, you have learnt about

- least square method

- moving average method

- semi average method.

The least square trend equation is written as

Y = a + b x + e

Where a = intercept = $a = \bar{Y} - \hat{b}\bar{X}$

slope = $b = \dfrac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$

For moving average develop the following

- n – moving total
- determine the moving average
- plot the trend value to know the trend line

For semi average

- divide the data into 2 equal parts
- when you have an odd data given, eliminate or delete the data in the middle

- get the half way total of each division

- divide the half way total by n depending on data supplied.

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?

1.7 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.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

The general formular is as given below:

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

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

Where;

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

$$\bar{X} = \frac{\Sigma Y}{n}$$

UNIT THREE: INDEX NUMBERS

1.1 Introduction

In introductory statistics a lot of meaning has be given to the average, this (average) has been confirmed not to be necessarily representative of the data it describes. statisticians have constructed a device that attempt to measure the magnitude of economic change over time, a device called index number. This device is also used for international comparison of economic data. This device called index number is what this unit shall be looking at, we shall examine the basic principles by which index numbers are constructed.

1.2 Learning Outcomes

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

- Define index number

- Calculate the index number through different methods
- Explain the Use(s) of index number
- Explain the relevance of index number

1.3 Index Number Defined

In statistical analysis of one very large and important class of problems, we must combine different set of data into a single measure e.g. we may wish to study the behaviour of wholesale prices and to do this, we calculate an index number which describes the changes, not in the various individual prices in which we are interested but in the group of prices taken as a whole.

The relevance of this statistical device is shown by the fact that governmental and other agencies devotes very substantial amount of money every year to the work of collecting appropriate data performing the necessary calculations for the construction of index numbers. The most widely known of this measure is the consumer price index or cost of living index.

In general, index numbers are used in the study of prices (wholesale, retail, form, export etc), output (manufacturing mining). The purpose of such measures is to get a summary of a whole range of similar activities, thereby, one will be able to investigate problem on relatively broad basis.

Index numbers are used extensively in economics and finance to represent

trends and compare changes across different sectors, regions, or time periods. Some of the most well-known index numbers include the Consumer Price Index (CPI), the Wholesale Price Index (WPI), and various stock market indices.

1.3.1 Characteristics of Index Numbers

Here are some of the main characteristics of index numbers:

1. **Measures Relative Change**: Index numbers measure the relative change in a variable or a group of variables over a certain period. They are not concerned with the absolute levels of the variables, but with their proportionate (or percentage) changes.
2. **Base Period**: Index numbers require a point of reference, called the base period. The value of the variable in the base period is usually normalized to 100, and changes in subsequent periods are expressed relative to this base period.
3. **Expressed as a Number or Percentage**: Index numbers are usually expressed as a number or percentage. For example, if the base period is set to 100, and the index number for a subsequent period is 120, this indicates a 20% increase from the base period.
4. **Aggregation of Variables**: Index numbers often combine several different variables into a single index. For example, the Consumer Price Index (CPI) combines prices from a 'basket' of many different goods and services. The

variables included in an index should be homogeneous and should together represent the overall characteristic that the index is intended to measure.

5. **Weighting of Variables**: In many indexes, different variables are given different weights to reflect their relative importance. For example, in the CPI, items that make up a larger proportion of consumer spending are given more weight.
6. **Choice of Formula**: Several different formulas can be used to calculate index numbers, such as the Laspeyres, Paasche, or Fisher formulas. The choice of formula can affect the value of the index and the interpretation of changes over time.
7. **Periodicity**: Index numbers can be calculated for different periods depending on the objective of the study - daily, monthly, quarterly, or annually are common intervals.

It's important to remember that while index numbers provide a convenient summary of complex data, they also have limitations. They are subject to issues such as substitution bias, quality changes, and new goods bias, which can complicate their interpretation. Therefore, while they are useful tools, they should be used and interpreted with care.

The main characteristics of index numbers are:

1. **Measures Relative Changes**: Index numbers measure the relative changes in variables over time or between different geographical locations.

2. **Uses a Base Period**: An index number uses a base period as a reference point, which is typically set to 100. All subsequent changes are relative to this base period.
3. **Expressed as a Percentage**: Index numbers are usually expressed as a percentage. An index number of 110, for example, indicates a 10% increase from the base period.

Self-Assessment Exercise 1

What is the basis for an index number?

1.4 Computation of Index Number

Under this subsection we are going to look at the different index number available, how the index number can be calculated through different methods.

1.4.1 Price Relative Index Number

This method is usually in use where just one commodity is involved. It measures the rate of change in single commodity.

Price relative $= \frac{P_n}{P_0} \times 100$

Where P_n refer to price of current year and P_o represents the price of the base period or reference period.

1.4.2 Simple Price Index Number

Simple price index number is defined as the sum total of the price of related items divided by the sum total of its base or reference period.

$$SPI = \frac{\Sigma P_n}{\Sigma P_0} \times 100$$

1.4.3 Weighted Price Index Number

Here the concept weight is introduced to index number. These weights indicates the importance of the particular commodity depending on whether we use base year, given year or typical year quantities denoted by Q_o,Q_n. We are going to look at the works and Marshall edge-worth Laspeyres, Paasche and Fisher on index number.

Laspeyres gave its own index number as

$$LPI = \frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0} \times \frac{100}{1}$$

$$LQI = \frac{\Sigma P_n Q_n}{\Sigma P_0 Q_0} \times \frac{100}{1}$$

Paasche gave its own as

$$PPI = \frac{\Sigma P_n Q_n}{\Sigma P_0 Q_n} \times \frac{100}{1}$$

$$PQI = \frac{\Sigma P_n Q_n}{\Sigma P_n Q_0} \times \frac{100}{1}$$

1.4.4 Fisher's Ideal Price Index

Fisher defined its own index number as the square root of the works of both Paashe and Laspeyres.

$$F = \sqrt{(Laspeyres\ Index) \times (Paasche\ Index)}$$

$$F = \sqrt{\left(\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0}\right)\left(\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_n}\right) \times 100}$$

1.4.5 Marshall Edge Worth Price Index

Marshall edge-worth defined its own index as

$$MEPI = \frac{\Sigma P_n (Q_0 + Q_n)}{\Sigma P_0 (Q_0 + Q_n)} \times 100$$

$$MEQI = \frac{\Sigma Q_n (P_0 + P_n)}{\Sigma Q_0 (P_0 + P_n)} \times 100$$

Where;

MEPI = Marshall edge-worth price index and

MEQI = Marshall edge-worth quantity index

Worked Example

Given the following about Open University, you are to compute the various price index numbers for 1990 using 1986 as base year.

Table M4.2.1: Table Showing Information about Open University

Commodity	Quantities		Prices	
	1986	1990	1986	1990
A	30	70	75	360
B	40	100	160	300
C	50	150	250	960
D	15	33	180	291

Solution

From the above 1986 values is the base year which represents the Po values and Pn value is represented by 1990 values.

Table M4.2.2: Laspeyre, Paasche and Fisher's Table of Analysis

Price			Quantities					
Comm.	1986	1990	1986	1990				
	Po	Pn	Qo	Qn	PnQn	PoQo	PnQo	PoQn
A	75	360	30	70	25,200	2,250	10,800	5,250
B	160	300	40	100	30,000	6,400	12,000	16,000
C	250	960	50	150	144,000	12,500	48,000	37,500
D	180	291	15	33	9,603	2,700	4,365	5,940
	665	1,911			208,803	23,850	75,165	64,690

$$SPI = \frac{\Sigma P_n}{\Sigma P_0} \times 100 = \frac{1911}{665} \times 100 = 287.36$$

This is simply implying that cost of the commodity had risen by 287.4% between 1986 and 1990.

$$LPI = \frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0} \times 100 = \frac{75165}{23850} \times 100 = 315.157$$

$$LQI = \frac{\Sigma P_0 Q_n}{\Sigma P_0 Q_0} \times 100 = \frac{64690}{23850} \times 100 \cong 271\%$$

Using Laspeyres price index it is showing the rate of rise in price as by 315.16% between 1986 and 1990.

Where LPI = Laspeyres Price Index

LPI = Laspeyres Quantity Index

Paasche method

$$PPI = \frac{\Sigma P_n Q_n}{\Sigma P_0 Q_n} \times 100 = \frac{208803}{64690} \times 100 \cong 322.77\%$$

$$PQI = \frac{\Sigma P_n Q_n}{\Sigma P_n Q_0} \times 100 = \frac{208803}{75165} \times 100 \cong 277.79\%$$

Using Paasche price index the rate of increase in price is 322.8% between 1986 and 1990.

Fisher's Ideal Price Index

$$FPI = \sqrt{LPI \times PPI} = \sqrt{315.16 \times 322.77} = \sqrt{101724.1932} = 318.$$

$$FQI = \sqrt{LQI \times PQI} = \sqrt{271.2396 \times 277.79286} = \sqrt{75347.67419} = 274.495\%$$

Marshall Edge-Worth Index Number

Table M4.2.3: Marshal Edge-Worth Table of Analysis

Po	Pn	Qo	Qn	Po+ Pn	Qo+ Qn	Pn(Qo+ Qn)	Po(Qo+ Qn)	Qo(Po+ Pn)	Qn(Po+Pn)
75	360	30	70	435	100	36,000	7,500	13,050	30,450
160	300	40	100	460	140	42,000	22,400	18,400	46,000
250	960	50	150	1201	200	192,200	50,000	60,500	181,500
180	291	15	33	471	48	13,968	8,640	7,065	15,543
						283,968	**88,540**	**99,015**	**273,493**

Using Marshall edge-worth price index

$$MEPI = \frac{\Sigma P_n(Q_0 + Q_n)}{\Sigma P_0(Q_0 + Q_n)} \times 100 = \frac{283968}{88540} \times 100 = 320.722\%$$

$$MEQI = \frac{\Sigma Q_n(P_0 + P_n)}{\Sigma Q_0(P_0 + P_n)} \times 100 = \frac{273493}{99015} \times 100 = 276.213\%$$

1.5 Problems Involved in Index Number Construction

While index numbers are incredibly useful for comparing changes over time or between groups, there are several challenges and problems involved in their construction:

1. **Selection of Base Year**: The choice of the base year is crucial as it serves as the reference point for the index. A poorly chosen base year can lead to misleading results. Typically, the base year should be a normal year, not affected by extreme events or anomalies.
2. **Selection of Items**: It's critical to choose items that represent the sector or population you're interested in. If the items don't accurately reflect the typical consumption or production patterns, the index won't be useful or representative.
3. **Changes in Quality**: Over time, the quality of goods and services can change, making it hard to compare prices directly. A car today is not the same product as a car from 30 years ago, for instance, and this needs to be taken into account when constructing price indices.
4. **Changes in Consumption Patterns**: Over time, consumption patterns can change, and new goods and services may be introduced while others may disappear. For instance, the consumption of digital goods and services has significantly increased in recent years. This change needs to be reflected in the basket of goods and services used for the index.
5. **Weighting of Items**: Assigning weights to the items in the basket can be challenging. Weights should reflect the relative importance of each item in the overall basket, which may require detailed expenditure data.

6. **Substitution Bias**: Over time, as prices change, consumers substitute cheaper goods for more expensive ones. If this is not accounted for, the price index can be overstated. This is known as substitution bias.
7. **Choice of Formula**: There are various ways to calculate an index, such as the Laspeyres, Paasche, and Fisher indices, each with its own strengths and weaknesses. The choice of formula can affect the resulting index.
8. **Data Availability**: Accurate and timely data is crucial for constructing index numbers. However, collecting this data can be time-consuming and expensive, and in some cases, the necessary data may not be available at all.

Despite these challenges, index numbers provide a valuable tool for understanding trends and making comparisons over time or between different groups. However, it's important to understand the limitations of any index number and interpret its results accordingly.

Self-Assessment Exercise 2

Define Fisher's Ideal formular for price index?

1.6 Summary

In the course of our discussion on this unit you have learnt about

- Definition of price index

- Calculation about:

Simple price index

Weighted price index; where we talked about

Laspeyres index number

Paasche index number

Fisher's ideal index number

Marshall edge-worth index number

Below is the summary of all the price indices we talked about in this unit.

Price relative $= \frac{P_n}{P_0} \times 100$

Simple price index $= \frac{\Sigma P_n}{\Sigma P_0} \times 100$

Laspeyres index $= \frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0} \times \frac{100}{1}$

Paasche index $= \frac{\Sigma P_n Q_n}{\Sigma P_0 Q_0} \times \frac{100}{1}$

Fisher's ideal price index $= \sqrt{(Laspeyres\ Index) \times (Paasche\ Index)}$

Marshall Edge-worth Index

$$MEPI = \frac{\Sigma P_n (Q_0 + Q_n)}{\Sigma P_0 (Q_0 + Q_n)} \times 100$$

$$MEQI = \frac{\Sigma Q_n (P_0 + P_n)}{\Sigma Q_0 (P_0 + P_n)} \times 100$$

Tutor Marked Assignment

Explain the problems with Index number computation

1.7 References/ Further Readings

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

Dawodu, A.F. (2008): Modern Business Statistics NIICHO Printing Works, Agbor, Delta State.
Edward, E.L. (1983): Statistical Analysis in Economic and Business. (2nd edition) Houghton Mifflin Company, Boston.
Esan, E.O. and Okafor, R.O. (2010): Basic statistical methods (Revised Edition) Toniichristo Concept, Lagos.
Olufolabo, O.O. and Talabi, C.O. (2002): Principle and practice of statistics. HASFEM (NIG) Enterprises, 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. (2nd edition) HEBP, Lagos.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

The basis for an index number lies in the concept of relative comparison. Instead of presenting absolute values, an index number expresses the value of a variable as a percentage or ratio relative to a specified base period or reference point. The base period is usually assigned an index number of 100, representing the baseline or starting point.

Answer to Self- Assessment 2

Fisher defined its own index number as the square root of the works of both Paashe and Laspeyres.

$$F = \sqrt{(Laspeyres\ Index) \times (Paasche\ Index)}$$

$$F = \sqrt{\left(\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_0}\right)\left(\frac{\Sigma P_n Q_0}{\Sigma P_0 Q_n}\right) \times 100}$$