COURSE GUIDE

NATIONAL OPEN UNIVERSITY OF NIGERIA

Course Code:	**SMS 202**
Course Title:	**Business Statistics**
Course Developer/Writer:	**KADIRI KAYODE I.** **School of Management Sciences (SMS)** **National Open University of Nigeria.**
Programme Leader:	**Dr. I.D. Idrisu (NOUN)**
Course Coordinator:	**ANTHONY EHIAGWINA**
Course Editor:	

March, 2014

BUSINESS STATISTICS CONTENTS

Introduction

What You Will Learn In This Course

Course Aims

Course Objectives

Working Through This Course

Course

Materials Study

Units Set

Textbooks

Assignment File

Presentation Schedule

Assessment

Tutor-Marked Assignment

(TMAs) Final Examination And

Grading Course Marking

Scheme

Course Overview

How To Get The Most From This Course

Tutors And Tutorials

Summary.

INTRODUCTION:

Business Statistics is a one semester, 3 credit units second year level course. It will be available to all second degree of the school of Management Sciences at the National Open University, Nigeria. It will also be useful for those seeking introductory knowledge in business statistics.

The course consists of eighteen units that involved basic concepts and principles of statistics and decision making process, forms of data, methods of data estimation, summarizing data, graphical presentation of data, measures of both index number and dispersion, co-efficient of correlation and regression analysis, some elements of hypothesis tests and time series analysis, distributions of both discrete and continuous random variables.

The course requires you to study the course materials carefully, supplement the materials with other resources from Statistics Textbooks both to be prescribed and those not prescribed that may treat the contents

of the course.

This Course Guide tells you what the course is about, 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 you are likely to spend on each unit of the course in order to complete it successfully. It also gives you some guidance on your tutor--marked assignments. Detailed information on tutor-marked assignment is found in the separate file.

There are likely going to be regular tutorial classes that are linked to the course. It is advised that you should attend these sessions. Details of the time and locations of tutorials will be communicated to you by National Open University of Nigeria (NOUN).

What You Will Learn In The Course

The overall aim of BHM202 Business Statistics is to introduce you to the basic concepts and principles of statistics and decision making process, forms of data, methods of data estimation, summarizing data, graphical presentation of data, measures of both index number and dispersion, co-efficient of correlation and regression analysis, some elements of hypothesis tests and time series analysis, distributions of both discrete and continuous random variables.

Course Aims

The course aims to give you an understanding of statistical information and presentation for decision-making. It exposes you to measures that are computed and used for processing materials for decision-making. It also gives the basic knowledge of some concepts used for making decisions and carefully summarizes some Probability Distributions.

This will be achieved by:

1. Introducing you to nature and form of statistical data
2. Showing how the statistical data can be collected and presented
3. Showing you how to compute measurement of dispersion in a sample or population
4. Showing you how to compute value of chi-square contingency table
5. Introducing you to the basic concepts of hypothesis tests
6. Give the basic principles for the application of some important

forecasting and time series analysis

Course Objectives

To achieve the aims set above the course sets overall objectives; in addition, each unit also has specific objectives. The unit objectives are included at the beginning of a unit, you should read them before you start working through the unit. You may want to refer to them during your study of the unit to check on your progress. You should always look at the unit objectives after completing a unit. In this way you can be sure you have done what was required of you by the unit.

We set out wider objectives of the course as a whole below. By meeting these objectives, you should have achieved the aims of the course.

On successful completion of the course, you should be able to:

1: Role of Statistics (Application of Statistics)

2 Measurement of Variables

3: Measurement of Dispersion, Skewness and Kurtosis

4 Decision Analysis and Administration

5: Index Number

6: Statistical Data

7: Sample and Sampling Theory

8: Estimation Theory

9: Correlation Theory and Goodness of Fit

10: Pearson's Correlation Co-efficient

11: Spearman's Regression Analysis

12: Ordinary Lease Square Estimation (Regression)

13: Multiple Regression Analysis

14: Hypothesis AND T-tests

15 F- Tests

16: Chi-Square Distribution

17: ANOVA

18: Forecasting and Time Series Analysis

Working through This Course

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

Each unit contains self-assessment exercises called Student Assessment Exercises, (SAE). At some points in the course, you are required towrite TMA on computer basic and submit on NOUN TMA PORTAR for assessment purposes. At the end of the course there is a final Examination. This course should take about 15 weeks to complete. Some listed components 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 given below

Below you will find listed components 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.

Course Materials

Major components of the course are:

(1) Course Guide (2) Study Units

(3) Textbooks

(4) Presentation Schedule.

Study Units

The course is in four modules and eighteen Study Units as follows:

Module 1: Role and Concepts of Statistics

Unit 1: Role of Statistics (Application of Statistics)

Unit 2 Measurement of Variables

Unit 3: Measurement of Dispersion, Skewness and Kurtosis

Unit 4 Decision Analysis and Administration

Module 2: INDEX NUMBER AND SAMPLING THEORIES

Unit 1: Index Number

nit 2: Statistical Data

Unit 3: Sample and Sampling Theory

Unit 4: Estimation Theory

Module 3: CORRELATION AND REGRESSION ANALYSIS

Unit 1: Correlation Theory and Goodness of Fit

Unit 2: Pearson's Correlation Co-efficient

Unit 3: Spearman's Regression Analysis

Unit 4: Ordinary Lease Square Estimation (Regression)

Unit 5: Multiple Regression Analysis

Module 4: STATISTICAL TEST

Unit 1: Hypothesis AND T-tests

Unit 2 F- Tests

Unit 3: Chi-Square Distribution

Unit 4: ANOVA

Unit 5: Forecasting and Time Series Analysis

The first four units concentrate on the roles and concepts of statistics. This constitutes Module 1. The next four units, module 2, concentrate on index number and research in management. Module3, deal with the correlation and regression analysis, The last five units Module 4, teach the principles underlying the applications of some important probability distributions., module 5, teach the principles underlying the applications of some important test of hypothesis and theory.

Each unit consists of one week direction for study, reading material, other resources and summaries of key issues and ideas. The units direct you to work on exercises related to the required readings

Each unit contains a number of self-tests. In general, these self-tests question you on the material you have just covered or required you to apply it in

some way and thereby help you to assess your progress and to reinforce your understating of the material. Together with tutor-marked assignments, these exercises will assist you in achieving the stated learning objectives of the individual units and of the course.

Set Textbooks

It is advisable you have some of the following books

ONWE J.O. NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OKOJIE, Daniel E. NOUN Statistics for Economist. Eco203

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

JUDE I.E, MICAN & EDIITH Statistics& Quantitative Methods for Construction & Business Managers.

Assessment

There are two types of the assessment of the course. First are the tutor-marked assignments (TMA); second, there is a computer base examination.

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

At the end of the course, you will need to sit for a final computer base examination of two hours' duration at designated centre. This examination will also count for 70% of your total course mark.

Tutor-Marked Assignments TMAs

There are four tutor-marked assignments in this course. You will submit all the assignments. You are encouraged to work all the questions thoroughly. Each assignment counts 12.5% toward your total course mark.

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 in all degree level education to demonstrate that you have read and researched more widely than the

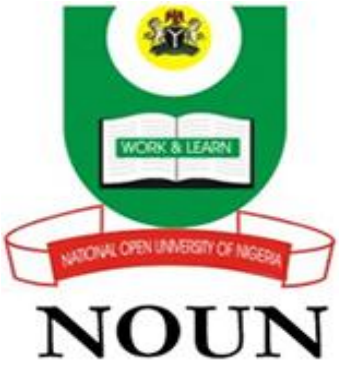
required minimum. You should use other references to have a broad viewpoint of the subject and also to give you a deeper understanding of the subject.

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

Final Examination and Grading

The final examination will be of three hours' duration and have a value of 70% of the total course grade. The examination will consist of questions which reflect the types of self testing, 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 the examination to revise the entire course. You might find it useful to review your self-tests, tutor-marked assignments and comments on them before the examination. The final examination covers information from all parts of the course.



NATIONAL OPEN UNIVERSITY OF NIGERIA

Course Code: **SMS 202**

Course Title: **Business Statistics**

Course Developer/Writer: **KADIRI KAYODE I.**
School of Management Sciences (SMS)
National Open University of Nigeria.

Programme Leader: **Dr. I.D. Idrisu (NOUN)**

Course Coordinator: **ANTHONY EHIAGWINA**

Course Editor:

March, 2014

BUSINESS STATISTICS

SMS STATISTICS

CONTENTS **PAGES**

Module 1: Role and Concepts of Statistics

Unit 1: Role of Statistics (Application of Statistics) ……..... 4

Unit 2 Measurement of Variables……..9

Unit 3: Measurement of Dispersion, Skewness and Kurtosis.....13

Unit 4 Decision Analysis and Administration.....29

Module 2: INDEX NUMBER AND SAMPLING THEORIES

Unit 1: Index Number 41

Unit 2: Statistical Data 51

Unit 3: Sample and Sampling Theory 55

Unit 4: Estimation Theory 65

Module 3: CORRELATION AND REGRESSION ANALYSIS

Unit 1: Correlation Theory and Goodness of Fit.....71

Unit 2: Pearson's Correlation Co-efficient.....75

Unit 3: Spearman's Regression Analysis.....83

Unit 4: Ordinary Lease Square Estimation (Regression).....95

Unit 5: Multiple Regression Analysis……...104

Module 4: STATISTICAL TEST

Unit 1: Hypothesis AND T-tests …….. 109

Unit 2 F- Tests…….....115

Unit 3: Chi-Square Distribution …….. 120

Unit 4: ANOVA…….....135

Unit 5: Forecasting and Time Series Analysis 152

MODULE 1 Roles and Concepts of Statistics

The general aim of this module is to provide you with a thorough understanding of Roles and Concepts of Statistics. Main focus here is to present you with the common roles and concepts of statistics as a general background to the course. The role of statistics and measurement of variables are brought to you.

The four units that constitute this module are statistically linked. By the end of this module you would have been able to list, differentiate and link these common statistics functions as well as identify and use them to solve related statistical problems. These units to be studied are;

Unit 1: Role of Statistics (Application of Statistics)
Unit 2: Basic Concepts in Statistics
Unit 3: Measurement of Variables
Unit 4 Measurement of Moments

UNIT 1: ROLE OF STATISTICS (APPLICATION OF STATISTICS)

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Definition of Statistics
    3.2 Role of Statistics
    3.3 Basic Concept in Statistics
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 Introduction

You will realize that the activities of man and those of the various organizations, that will often be referred to as firms, continue to increase. This brings an increase in the need for man and the firms to make decisions on all these activities. The need for the quality and the quantity of the information required to make the decisions increases also. The management of any firm requires scientific methods to collect and analyze the mass of information it collects to make decisions on a number of issues. Such issues include the sales over a period of time, the production cost and the expected net profit. In this regard, statistics plays an important role as a management tool for making decisions.

2.0 Objectives

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

- Understand the various definitions of statistics
- Describe the uses of statistics
- Define the basic concepts in statistics.

3.1 Definitions of Statistics

Statistics can be defined as a management tool for making decision. It is also a scientific approach to presentation of numerical information in such a way that one will have a maximum understanding of the reality represented by such information. Statistics is also defined as the presentation of facts in numerical forms. A more comprehensive definition of statistics shows statistics as a scientific method which is used for collecting, summarizing, classifying, analyzing and presenting information in such a way that we can have thorough understanding of the reality the information represents.

From all these definitions, you will realize that statistics are concerned with numerical data.. Examples of such numerical data are the heights and weights of pupils in a primary school when evaluating the nutritional well being of the pupils and the accident fatalities on a particular road for a period of time.

You should also know that when there are numerical data, there must be non-numerical data such as the taste of brands of biscuits, the greenness of some vegetables and the texture of some joints of a wholesale cut of meat. Non-numerical data cannot be subjected to statistical analysis except they are transformed to numerical data. To transform greenness of vegetables to numerical data, a five point scale for measuring the colour can be developed with 1 indicating very dull and 5 indicating very green.

3.2 The Roles of Statistics

You will realize that statistics is useful in all spheres of human life. A woman with a given amount of money, going to the market to purchase foodstuff for the family, takes decision on the types of food items to purchase, the quantity and the quality of the items to maximize the satisfaction she will derive from the

purchase. For all these decisions, the woman makes use of statistics

Government uses statistics as a tool for collecting data on economic aggregates such as national income, savings, consumption and gross national product. Government also uses statistics to measure the effects of external factors on its policies and to assess the trends in the economy so that it can plan future policies.

Government uses statistics during census. The various forms sent by the government to individuals and firms on annual income, tax returns, prices, costs, output and wage rates generate a lot of statistical data for the use of the government

Business uses statistics to monitor the various changes in the national economy for the various budget decisions. Business makes use of statistics in production, marketing, administration and in personnel management.

Statistics is also used extensively to control and analyze stock level such as minimum, maximum and reorder levels. It is used by business in market research to determine the acceptability of a product that will be demanded at various prices by a given population in a geographical area. Management also uses statistics to make forecast about the sales and labour cost of a firm. Management uses statistics to establish mathematical relationship between two or more variables for the purpose of predicting a variable in terms of others. For the conduct and analyses of biological, physical, medical and social researches, we use statistics extensively.

3.3 Basic Concepts In Statistics

Let us quickly define some of the basic concepts you will continue to come across in this course.

- **Entity**: This may be person, place, and thing on which we make observations. In studying the nutritional well being of pupils in a primary school, the entity is a pupil in the school.

- **Variable**: This is a characteristic that assumes different values for different entities. The weights of pupils in the primary school constitute a variable.

- **Random Variable:** If we can specify, for a given variable, a mathematical expression called a function, which gives the relative frequency of occurrence of the values that the variable can assume, the function is called a probability function and the variable a random variable.

- **Quantitative Variable:** This is a variable whose values are given as numerical quantities. Examples of this is the hourly patronage of a restaurant

- **Qualitative Variable:** This is a variable that is not measurable in

numerical form or that cannot be counted. Examples of this are colours of fruits, taste of some brands of a biscuit.

- **Discrete Variable:** This is the variable that can only assume whole numbers. Examples of these are the number of Local Government Council Areas of the States in Nigeria, number of female students in the various programmes in the National Open University. A discrete variable has "interruptions" between the values it can assume. For instance between 1 and 2, there are infinite number of values such as 1.1, 1.11, 1.111, 1.IV land so on. These are called interruptions.

- **Continuous Variable:** This is a variable that can assume both decimal and non decimal values. There is always a continuum of values that the continuous variable can assume. The interruptions that characterize the discrete variable are absent in the continuous variable. The weight can be both whole values or decimal values such as 20 kilograms and 220.1752 kilograms.

- **Population:** This is the largest number of entities in a study. In the study of how workers in Nigeria spend their leisure hours, the number of workers in Nigeria constitutes the population of the study.

- **Sample:** This is the part of the population that is selected for a study. In studying the income distribution of students in the National Open University, the incomes of 1000 students selected for the study, from the population of all the students in the Open University will constitute the sample of the study.

- **Random Sample:** This is a sample drawn from a population in such a way that the results of its analysis may be used to generalize about the population from which it was drawn.

Exercise 1.1

What is the importance of Statistics to human activities? Your answer can be obtained in section 3.2 of this unit.

4.0 Conclusion

In this unit you have learned a number of important issues that relate to the meaning and roles of statistics. The various definitions and examples of concepts given in this unit will assist tremendously in the studying of the units to follow.

5.0 Summary

What you have learned in this unit concerns the meaning and roles of statistics, and the various concepts that are important to the study of statistics.

6.0 Tutor Marked Assignment

What is Statistics? Of what importance is statistics?

7.0 REFERENCES /FURTHER READING

AJAYI J. NOUN TEXT BOOK, BHM 106: Business Statistics

JUDE, MICAN & EDITH N. Statistical & Quantitative Methods for Construction & Business Managers

UNIT 2: MEASUREMENT OF VARIABLES

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Definition of Variable
    3.2 Measurement of Variables
    3.3 Variance of Binomial Distribution
4.0 Conclusion
5.0 Summary
6.0 Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

Variable can be used under the following conditions:

Information, which are not numeric in nature are cafled qualitative verb instance, information on colour of the skin, colour of the eye or hair, level, of education, si status, and other qualitative categories as building types are qualitative variables. (Variables can be assigned numerical values. This assignment of numerical values to information is called coding. Also these qualitative data can be arranged in order of impoi values assigned to them in that order. This is called ranking.

.2.0 OBJECTIVES

The aim of this unit is to enable student understand the meaning of variable and instances when it is applicable.

3.0 MAIN CONTENT

3. 3.1 What is a Variable?

A variable is any characteristic of an object or concept that is capable different values or falling into more than one distinct category. For instance, a bull object, but the characteristics of a building such as size, type, cost and age are varil

Also rain is an object, but the amount of rain is a variable. Other variables Inc height, sex, weight, colour of the skin, hair colour, genotype, blood group, maci religious affiliation, level of education attained, place of residence of a person strength of Dangote cement, tensile strength, number of bags of cement in the star of bags of cement used in the site per day, expenditure, income of household per r degree of satisfaction, level of intelligence etc. Therefore, any characteristic of an vary in time and space is called a variable.

Statistical raw data are generated or provided by these variables. That is, attached to the variables constitutes statistical data. A single value of a variable it observation, an item, a score or a case.

Quantitative variable can be classified into two major types, viz. dis continuous variables.

3.1.2. Discrete variables are variable those values are whole numbers or integers. Th fractional part, they are countahs or finite. Examples of discrete variables include housing unit, number of students in a class, number of goals scored in a fooft number of cars sold etc.

3.1.3. Continuous variables are variables that assume any value ii'ithin an interval or r have the property of infinite divisibility. They can assume fractional values. Example weight, height, cost, scores, income, breaking strength etc.

3.2. Measurement of variables

There are four measurement scales available as insrrumenrs for measurinl These scales easily identiQi variables. The scales are nominal, ordinal, interval and Nominal scale -

This scale groups the objects into distinct categories to facilitate referencing. Alt is attached to each distinct category. Examples of nominal scale variables include sex, in religious or party affIliation, genotype, blood group, place of residence, etc. Also, we the various categories of the nominal variables with numbers (or codes): When this number or code is mere label or mere identification mark, which do not permit an operation. For instance, marital status may be categorized as married, separated, divorce married, If we assign 1 to married, 2 to separated, 3 to divorced and 4 to never marry, the married. If we assign 1 to married, 2 to separted, 3 to divorced and 4 to never marry, these numbers ine coc The numbers do not indicate order of importance of the various categories and the sum of land 3 can not produce categoiy 4. This is the lowest scale of measurement. O, sCar.

This scale ranks or orders the mutually exdusive categories of the variables according to the importance attached to each category. This scale has all the properties of the nominal scale plus the additional property of ordering or ranking the categories. Examples are, a teacher rating his students according to their performance — A, B, C, D, E, and F or 1", 3 income groups of individuals dassifled as high, medium, and low, dassiflcanon of a city according to high, medium and low density of population concentration. The numbers assigned to each variable category only help toortleror rank the observations in ascending or descending order. Many statist!ca! npcrntions that are based on ranking or rank ordering are permissible under this scale. Examples of such statistical techniques are Spearman's rank correlation coefficient, Wilcoxon rank- sum test, signed rank test etc. This scale is higher than the normal scale.

This has the combined properties of the nominal and ordinal scale plus the additional property of measuring the distance or interval between two measurements. This scale gives information on how much one category is more or less than the other. Examples are age in years, income, pressure, and temperature. This scale has no absolute zero. That is, the selected zero point in this scale is arbitrary. That a student scored zero percent in examination does not mean that he does not know anything in that course. Interval variables are quantitative and may be discrete or continuous. As such arithmetic operations of addiction and subtraction are permitted. Many statistical procedures are permissible in this scale, the mean, standard deviation, product moment correlation coefficient and other statistical inferences are possible on this scale.

Ratio scale

This scale has all the properties of the nominal, ordinal and interval scales including the additional property of having an absolute zero point. This is the highest level of measurement. Examples are measurement of height, weight, volume, price of an item, votes scored in an election, etc. many statistical procedures are available for ratio scale data.

Note that the scale of measurement of variables determines the type of statistical tool to be

employed.

4.0 CONCLUSION

In probability theory and statistics, the Binomial distribution is the discrete probability distribution of the number of successes in a sequence of *n*independent yes/no experiments,

each of which yields success with probability*p*. Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial; when $n = 1$, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance. The Binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution is a good approximation, and widely used.

5.0 SUMMARY

You have been made to understand in this unit that the meaning of variables. And the measurement of various variables.. Therefore, in summary, the measurement of variable describes the behaviour of a scale, if the following conditions apply:

1. The Ratio Scale.
2. Nominal Scale.
3. Ordinal Scale.
4. Interval Scale.

If in your application of variables, these conditions are met, then statistical scale has a meaning.

6.0 TUTOR-MARKED ASSIGNMENT

1. What is a variable? Distinguish between quantitative and qualitative variables, discrete and continuous variables.

2. Write short notes on:

Nominal scale (ii) Ordinal scale

(iii) Interest scale (iv) Ratio scale

7.0 REFERENCES/FURTHER READINGS

ONWE J.O. NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OTOKOTI O.S. Contemporary Statistics

JUDE, MICAN & EDITH N. Statistical & Quantitative Methods for Construction & Business Managers

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 3: MEASURES OF DISPERSION, SKEWNESS AND KURTOSIS

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

- 3.1 Measurement of Dispersion
- 3.2 Measure of Skewness
- 3.3 Kurtosis

4.0 Conclusion

5.0 Summary

6.0 Assignment

7.0 References / Further Reading

1.0 INTRODUCTION

The second most important characteristics which describe a set of data is the amount of variation, scatter, or spread in the data. In this chapter, we discuss in detail the various measures of dispersion and skewness. The purpose of these measures is to amplify the imperfect summary of any statistical distribution usually provided by the three measures of averages commonly used: the mean, the median, and the mode. These averages are inherently unsatisfactory because no single measure of average can tell you everything about a distribution, and the wider the dispersion of a given data around the average, the less satisfactory the average becomes. In order to improve your understanding of population averages, you need to know how wide the dispersion is around the average, and whether it is symmetrical (un-skewed) or asymmetrical (skewed).

The first set of measures to be discussed here are measures of dispersion, and the second set measures of skewness.

Fig. 1: Normal Curve

2.0 OBJECTIVE

The main aim of this unit is to ensure students' proper understanding of the measurement of dispersion and skewness; appreciate its applicability in day-to-day business and scientific live and be able to use it as appropriate in practical statistical studies

3.0 MAIN CONTENT

3.1 MEASURES OF DISPERSION

The common measures of dispersion include:

(a) The Range

(b) The Quartile Deviation

(c) Mean Deviation

(d) Variance

(e) Standard Deviation

(f) Coefficient of Variation

The variation or dispersion can be said to measure the degree of uniformity of observations in a given set of data. The greater the variation, the more un-uniform the observations in a given set of data

The Range

The Range (R) of a given set of ungrouped data can be determined from an ordered array as the difference between the highest observation and the lowest observation in a distribution..

Let X_h = Highest observation

X_L = Lowest observation

Then, $R = X_h - X_L$

Given the arrayed data: X = 2,5,8,9,12,13,18,

the range will be:

$R = 18 – 2 = 16.$

The range can be an unsatisfactory measure of dispersion because it is affected by extreme values or items which renders it unrepresentative of majority of the set of data.

The Quartile Deviation

Unlike the range, quartile deviation does not take extreme values or items. Quartiles are the boundaries separating the items in a given distribution or set of data into quarters.

There are, therefore, three quartiles: the *lower quartile* (at the 25 percent mark); the *median* (at the 50 percent mark); and, the *upper quartile* (at the 75 percent mark). To compute the quartiles of *ungrouped data*, you simply use:

0.25 (n + 1), for the lower quartile

0.50 (n + 1), for the median quartile

0.75 (n + 1), for the upper quartile

For *grouped data*, you simply use:

0.25n for the lower quartile

0.5n for the median quartile

0.75n for the upper quartile

Example

Consider the following output distribution of the employees of a manufacturing company:

Table 3.1: Output of Employees

Units of Output	Number of Employees (f)
21 – 30	7
31 – 40	11
41 – 50	14
51 – 60	8
61 – 70	5

Table 3.1 indicates that there are 45 items or observations (ie. total number of employees or sum of the frequencies, Σf).

Using these information, the quartiles are as follows:

Lower quartile (Q1) = $0.25n = 0.25(45) = 11.25^{th}$ item

Median quartile (Q2) = $0.5n = 0.5(45) = 22.5^{th}$ item

Upper quartile (Q3) = $0.75n = 0.75(45) = 33.75^{th}$ item

The values of the quartile items are determined simply as follows:

Lower quartile: Since, according to table 3.1, there are 7 items in the first group (ie, group of 21 – 30), the quartile item is the (11.25 – 7) = 4.25th item of the second group. Thus,

$$\text{Value of the lower quartile (Q1)} = 30 + \frac{4.25}{11} \times 10 \text{ units}$$

$$= 30 + 3.66$$

$$= 34 \text{ approximately.}$$

Therefore, the value of the lower quartile is about 34 units.

In a similar process, the value of the median and upper quartiles can be determined, thus:

Value of Median quartile: The 22.5th item in the distribution is in the 41 – 50 group and is the (22.5 – 18) = 4.5th item out of 14 in the group (note that the figure 18 is the cumulative frequency of the first and second groups,and the figure 10 appearing in the calculations is the class interval of the distribution). The value of the median quartile (Q2) is therefore:

$$Q2 = 40 + \frac{4.5}{14} \times 10$$

$$= 40 + 3.21 = 43.21$$

$$= 43 \text{ units approximately.}$$

Value of the Upper quartile (Q3): The 33.75th item in the distribution in the third group, the group of (41 – 50), and since there are 32 items in the third group (the cumulative frequency), the median is the (33.75 – 32) = 1.75th item in the fourth group. The value of the upper quartile is therefore:

$$Q3 = 50 + \frac{1.75}{8} \times 10$$

$$= 50 + 2.19 = 52.19$$

$$= 52 \text{ units approximately.}$$

The quartile deviation referred to as the semi-interquartile range is defined as one-half the difference between the upper quartile and the lower quartile. Thus,

$$\text{Quartile Deviation} = \frac{Q3 - Q1}{2}$$

In this example, therefore, the quartile deviation is:

$$\frac{52 - 34}{2} = 9 \text{ units}$$

The distribution in table 3.1 can then be described as having a median value of 43 units and a quartile deviation around the median value of 9 units.

The Mean Deviation (MD)

The Mean Deviation can be defined simply by the following relationship:

$$MD = \frac{\Sigma /X-\bar{X}/}{n}$$

where $\Sigma /X-\bar{X}/$ = sum of the absolute values of deviation from arithmetic mean

n = number of observation

As an example, consider again the arrayed data, X = 2,5,8,9,12,13,18.

The mean deviation, MD, can be computed as follows:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{67}{7} = 9.57$$

By tabulation,

X	$(X - \bar{X})$ / X - X /
2	-7.57	7.57
5	-2.57	2.57
8	-1.57	1.57
9	-0.57	0.57
12	2.43	2.43
13	3.43	3.43
18	8.43	8.43

Σ /X-X/ = 26.57

Thus,

$MD = \frac{\sum /X-X/}{n} = \frac{26.57}{7} = 3.7957$

The Variance

The Variance for a given set of an ungrouped data can be defined by:

$$\text{Variance} = S^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

where X represents the numerical values of the given set of an ungrouped data.

Continuing with our earlier example, where

X = 2,5,8,9,12,13,18

and by tabulation:

X	X^2
2	4
5	25
8	64
9	81
12	144
13	169
18	324

$\sum X = 67$; $\sum X^2 = 811$;

$(\sum X)^2 = (67)^2 = \underline{4489} = 641.29$

n 7 7

Thus, $S^2 = \underline{\sum x^2-(\sum x)^2/n} = \underline{811-641.29}$

n-1 7-1

$= \underline{169.71} = 28.285$

6

Thus, the variance of the given set of ungrouped data is 28.285.

The Standard Deviation

Simply stated, the standard deviation is the most useful measure of variation. It can be defined as the square root of the variance for a given set of data.

Thus,

Standard deviation = S = $\sqrt{S^2}$

Or,

$S = \sqrt{\underline{\sum X^2-(\sum X)^2/n}}$, for ungrouped data.

n-1

The standard deviation for the last example is:

$S = \sqrt{S^2} = \sqrt{28.285} = 5.318$

Variance And Standard Deviation For A Grouped Data

The computation of variance and standard deviation for a grouped data is illustrated with the following example.

The Variance and Standard Deviation for a grouped data are defined by the following formulations:

$$\text{Variance} = S^2 = \frac{\sum fx^2 - (\Sigma fx)^2/n}{n-1}$$

$$\text{Standard deviation} = \sqrt{S^2} = \sqrt{\frac{(\sum fx^2 - (\sum fx)2/n}{n-1}}$$

Example.

The following data presents the profit ranges of 100 firms in a given industry.

Profits (N'millions)	No. of Firms (f)
10-15	8
16-21	18
22-27	20
28-33	12
34-39	15
40-45	17
46-51	10
	$\sum f = n = 100$

We are required to compute the variance and standard deviation of profits within the industry.

Solutions.

By definition,

$$\text{Variance} = S^2 = \frac{\sum fx^2 - (\sum fx)^2/n}{n-1}$$

Standard Deviation = $\sqrt{S^2} = \sqrt{\frac{\sum fx^2 - (\sum fx)^2/n}{n-1}}$

The computational process is as follows:

Profits (N millions)	Frequency (f)	Mid-Value (x)	fx	x^2	fx^2
10-15	8	12.5	100	156.25	1250
16-21	18	18.5	333	342.25	6160.5
22-27	20	24.5	490	600.25	12005
28-33	12	30.5	366	930.25	11163
34-39	15	36.5	547.5	1332.25	19983.75
40-45	17	42.5	722.5	1806.25	30706.25
46-51	10	48.5	485	2352.25	23522.50
	$\sum f=n=100$		$\sum fx=3044$		

SUMMARY:

$\sum fx^2 = 104791$

$\sum fx=3044$

$\frac{(\sum fx)^2}{n} = \frac{(3044)^2}{100} = 92659.36$

$\sum fx^2 = 104791$

It follows that:

Variance = $S^2 = \underline{\sum fx^2-(\sum fx)^2/n} = \underline{104791-92659.36}$

$n-1 \qquad 100-1$

$= \underline{12131.64} = 122.54$

99

Standard Deviation = $\sqrt{S^2} = \sqrt{122.54} = 11.07$

Thus, the required variance and standard deviation are 122.54 and 11.07 respectively.

The Coefficient of Variation

Unlike other measures of variability, the coefficient of variation is a relative measure. It is particularly useful when comparing the variability of two or more sets of data that are expressed in different units of measurements.

The coefficient of variation measures the standard deviation relative to the mean and is computed by:

Coefficient of Variation = CV = $\underline{S}$ x 100%

$\bar{X}$

The coefficient of variation is also useful in the comparison of two or more sets of data which are measured in the same units but differ to such an extent that a direct comparison of the respective standard deviations is not very helpful. As an example, suppose a potential investor is considering the purchase of shares in one of two companies, A or B, which are listed on the Nigerian Stock Exchange (NSE). If neither company offered dividends to its shareholders and if both companies were rated equally high in terms of potential growth, the potential investor might want to consider the volatility of the two stocks to aid in the investment decision.

Now, suppose each share of stock in Company A has averaged N50 over the past months with a standard deviation of N10. In addition, suppose that in this same time period, the price per share for Company B's stock averaged N12 with a standard deviation of N4. Observe that in terms of actual standard deviations, the price of Company A's shares seems to be more volatile than that of Company B. However, since the average prices per share for the two stocks are so different, it would be more appropriate for the potential investor to consider the

variability in price relative to the average price in order to examine the volatility/stability of two stocks.

The coefficient of variation of company A's stock is

$$CV_A = \frac{S_A}{\overline{X}_A} \times 100\% = \frac{N10}{N50} \times 100\% = 20\%$$

That of Company B's is

$$CVB = \frac{S_B}{\overline{X}_B} \times 100\% = \frac{N4}{N12} \times 100\% = 33.3\%$$

It follows that relative to the average, the share price of company B's stock is much more variable/unstable than that of Company A.

3.2 MEASURES OF SKEWNESS

The measures of skewness are generally called Pearson's first coefficient of skewness and Pearson's second coefficient of skewness. Measures of skewness are used in determining the degree of asymmetry of a distribution; a distribution which is not symmetrical is said to be skewed.

The Pearson's No. 1 Coefficient of skewness: The formula used in calculating Pearson's No. 1 coefficient is:

$$Sk = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Notice that the mean, the mode, and the standard deviation are all expressed in the units of the original data. When the difference between the mean and the mode is computed as a

fraction as a fraction of the standard deviation (or average spread of the data around the mean), the original units cancel out in the fraction. The result will be a coefficient of skewness, a number which tells you the extent of the skewness in the distribution.

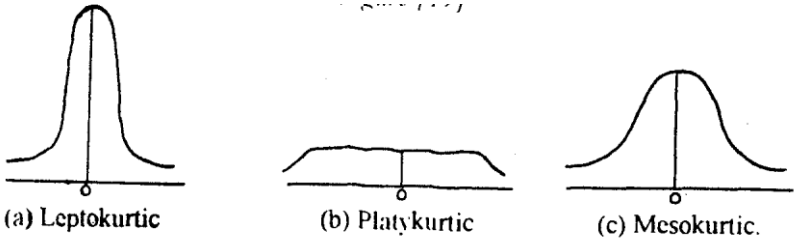
Example: Consider a set of data on monthly sales of a company's product, the mean of which was found to be N240,000; the mode found to be N135,000; and the standard deviation found to be N85,000. The Pearson's No. 1 Coefficient of skewness would be calculated as follows:

$$Sk = \frac{\text{mean} - \text{mode}}{\sigma} = \frac{240{,}000 - 135{,}000}{85{,}000}$$

$$= 1.24$$

3.2 KURTOSIS

Kurtosis measures the degree of peakedness of a distribution. It is usually taken relative to a normal distribution. There are usually three types of kurtosis namely. **LEPTOKURTIC, PLATYKURTIC and MESOKURTIC**. The mesokurtic is otherwise known as normal distribution curve i.e. the curve that is moderately distributed. The figures below show the relative peakedness of distribution of data.



(a) Leptokurtic (b) Platykurtic (c) Mesokurtic.

The moment coefficient of kurlosis is used to calculate the peakedness of a distribution. However, for normal distribution (mesokurtic). The moment coefficient is given as b = a = 3. If moment coefficient of kurtosis a > 3 it is said

to be leptokurtic: If a < 3 ii is equal to platykurtic and it is a1led mesokurtic when a = 3.

Example calculates the first four moments about the means for the weight distribution of the students in National Open University of Nigeria given below:

Class(X)	51-53	54-56	57-59	60-62	63-65
f	4	17	41	26	7

Solution:

Class(X)	**MD**	**F**	**U=xA/c**	**FU**	**F(U^2)**	**F(U^3)**	**F(U^4)**
51-53	52	4	-2	-8	16	-32	64
54-56	55	17	-1	-17	17	-17	17
57-59	58	41	0	0	0	0	0
60-62	61	26	1	26	26	26	26
63-65	64	7	2	14	28	56	112
		95		**15**	**87**	**33**	**219**

A= Average Mean of 'X'=58, C= class interval= 53.5-50.5=3.

M_1= (€fu/€f) c = 15/95 x 3 = 0.474

M_2 = [€f(u^2)/€f] C^2 = 87/95 x 3^2 =8.242

M_3 = [€f(u^3)/€f] C^3 = 33/95 x 3^3 = 9.379

M_4 **= [€f(**$\mathbf{u^4}$**)/€f]** C^4 = 219/95 x 3^4 = 186.73

Thus m_1 = 0

$M_2 = m_2-(m_1)^2 = 8.242 –(0.474)^2 = 8.017$

$M_3 = m_3 – 3m_1m_2 + (m_1)^3 = 9.379 – 3(0.474)(8.242) + (0.474)^3$

= 9.379 – 11.720124 + 0.1065 = -2.235

$M_4 = m_4 - 4m_1m_3 + 6(m_1)^2m_2 - 3(m_1)^4$

$M_4 = 186.73 - 4(0.474)(9.379) + 6(0.474)^2(8.242) - 3(0.474)^4$

$= 186.73 - 17.783 + 11.1107 - 0.1514$

$M_4 = 179.91$

Then moment coefficient of kurtosis is

$A_4 = m_4/5^4 = m_4/(m_2)^2 = 179.91/(8.017)^2 = 2.799.$

Since $a_4 = 2.799<3$ it means that the distribution is playkurtic in relation to the normal distribution.

4.0 CONCLUSION

Generally, a complete absence of skewness would have a coefficient of skewness equal to zero. In our example, since the mean was larger than the mode, we obtained a positive coefficient of skewness to the extent of 124% of the standard deviation.

The Pearson's No. 2 Coefficient of Skewness: This type of the Pearson's coefficient of skewness came as a result of the fact that a precise calculation of mode is difficult in many distributions. Hence, Pearson's No. 2 coefficient of skewness uses the difference between the mean and the median of the distribution instead of the difference between the mean and the mode. In this calculation, you have the formula:

$$sk = \frac{3(\text{mean} - \text{median})}{\sigma}$$

This formula should give you a more accurate measure of skewness than that of the Pearson's No. 1 formula.

5.0 SUMMARY

Easily now, you can comprehend that the dispersion and skewness can be described completely by the two parameters $\mu and \sigma$. As always, the mean is the center of the distribution and the standard deviation is the measure of the variation around the mean.

6.0 TUTOR-MARKED ASSIGNMENT

1. Consider the monthly sales revenue of 90 sales representatives of a conglomerate:

Sales (N'000s)	No. of Sales Reps
10 – 15	10
16 – 21	36
22 – 27	28
28 – 33	10
34 – 39	6

Using this distribution, compute:

(a) The mean, modal, and median sales for the sales reps.

(b) The standard deviation and coefficient of variation of the sales distribution

(c) The coefficient of skewness of the sales distribution

2. A distribution of data about the sales reps' salaries per month is found to have an arithmetic mean of N60,000, with a standard deviation of N15,000, and a coefficient of skewness of 0.92. Explain what these terms mean in describing the distribution of the sales reps' salaries.

3. A certain set of data about the weight of female typists in the 25 – 32 age group gives a mean weight of 51 kg, a standard deviation of 7.3 kg, and a median weight of 49.6 kg. Compute and explain the coefficient of skewness

7.0 REFERENCES/FURTHER READINGS

ONWE J.O. NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OTOKOTI O.S. Contemporary Statistics

JUDE, MICAN & EDITH N. Statistical & Quantitative Methods for Construction & Business Managers

TAIWO S. O. Statistics for Undergraduates

UNIT 4: ADMINISTRATION AND DECISION THEORY

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

- 3.1 Administrative and Decision Analysis
- **3.2** Certainty and Uncertainty in Decision
- **3.3** Expected Monetary Value Decisions

4.0 Conclusion

5.0 Summary

6.0 Assignment

7.0 References / Further Reading

1.0 INTRODUCTION

DECISION ANALYSIS

Decision analysis is the modern approach to decision making both in economics and in business. It can be defined as the logical and quantitative analysis of all the factors influencing a decision. The analysis forces decision makers to assume some active roles in the decision-making process. By so doing, they rely more on rules that are consistent with their logic and personal behaviour than on the mechanical use of a set of formulas and tabulated probabilities.

2.0 Objective

The primary aim of decision analysis is to increase the likelihood of good outcomes by making good and effective decisions. A good decision must be consistent with the information and preferences of the decision maker. It follows that decision analysis provides decision-making framework based on available information on the business environment, be it a sample information, judgmental information, or a combination of both.

3.0 Main Content

3.1 ADMINISTRATIVE AND DECISION PROCESS

The art of problem solving and decision making is base on common sense. It is to ensure that better quality decisions are made for approaching problem solution.

There are two (2) main ways of approaching problems and obtaining solutions.

1. Analytical Thinking
2. Creative Thinking.

ANALYTICAL THINKING: Seeks to improve on a given situation

CREATIVE THINKING: This considers end rather than means

The above two (2) approaches may further be subdivided unto various methods

- Critical Examinations
- Brain Storming or Group Creativity
- Analogies
- Morphological Approach or Attribute Listening
- Heuristic Approach

Critical Examinations: Is the logical approach, it answers questions like What, Who, Where, How, When and Why. The result of ''why'' investigation are supposed to indicate possible alternatives or choices from which an acceptable solution may be derived.

Brain Storming: This is a method base on two head is better than one. Brain Storming involves conference techniques by which a group of people attempts to find solution for specified problems by amazing all ideas spontaneously contributed by its members. It is a free thinking meeting.

Steps on Brain Storming are as follows

- Orientation
- Consideration
- Speculation (opinion)
- Recommendation

Analogies: This is the comparison of one thing with another that has similar features

Types of Analogous

- Personal Analogy: Is putting self in place of object.
- Direct Analogy: This to compare things in nature with the situation and use nature's solution to provide a lead to a suitable solution.
- Symbolic Analogy: Uses objects and impersonal images.
- Fantancy Analogy: The problem is transformed into the realism of fantancy.

Morphological Approach or Attribute Listening: This method looks for the attributes or qualities of the product .i.e. comparison of the best one.

3.2 CERTAINTY AND UNCERTAINTY IN DECISION ANALYSIS

Most decision-making situations involve the choice of one among several alternatives actions. The alternative actions and their corresponding payoffs are usually known to the decision-maker in advance. A prospective investor choosing one investment from several alternative investment opportunities, a store owner determining how many of a certain type of commodity to stock, and a company executive making capital-budgeting decisions are some examples of a business decision maker selecting from a multitude of a multitude of alternatives. The decision maker however, does not know which alternative which alternative will be best in each case, unless he/she also knows with certainty the values of the economic variables that affect profit. These economic variables are referred to, in decision analysis, as *states of nature* as they represent different events that may occur, over which the decision maker has no control.

The states of nature in decision problems are generally denoted by s_i (i = 1, 2, 3, …, k), where k is the number of or different states of nature in a given business and economic environment. It is assumed here that the states of nature are mutually exclusive, so that no two states can be in effect at the same time, and collectively exhaustive, so that all possible states are included within the decision analysis.

The alternatives available to the decision maker are denoted by

a_i (i = 1, 2, 3, …, n), where n is the number of available alternatives. It is also generally assumed that the alternatives constitute a mutually exclusive, collectively exhaustive set.

When the state if nature, s_i, whether known or unknown, has no influence on the outcomes of given alternatives, we say that the decision maker is operating under *certainty*. Otherwise, he/she is operating under *uncertainty*.

Decision making *under certainty* appears to be simpler than that under uncertainty. Under certainty, the decision maker simply appraises the outcome of each alternative and selects the one that best meets his/her objective. If the number of alternatives is very high however, even in the absence of uncertainty, the best alternative may be difficult to identify. Consider, for example, the problem of a delivery agent who must make 100 deliveries to different residences scattered over Lagos metropolis. There may literally be thousands of different alternative routes the agent could choose. However, if the agent had only 3 stops to make, he/she could easily find the least-cost route.

Decision making *under uncertainty* is always complicated. It is the probability theory and mathematical expectations that offer tools for establishing logical procedures for selecting the best decision alternatives. Though statistics provides the structure for reaching the decision, the decision maker has to inject his/her intuition and knowledge of the problem into the decision-making framework to arrive at the decision that is both theoretically justifiable and intuitively appealing. A good theoretical framework and commonsense approach are both essential ingredients for decision making under uncertainty.

To understand these concepts, consider an investor wishing to invest N100,000 in one of three possible investment alternatives, A, B, and C. Investment A is a Savings Plan with returns of 6 percent annual interest. Investment B is a government bond with 4.5 percent annual interest. Investments A and B involve no risks. Investment C consists of shares of mutual fund with a wide diversity of available holdings from the securities market. The annual return from an investment in C depends on the uncertain behaviour of the mutual fund under varying economic conditions.

The investors available actions (a_i; I = 1, 2, 3, 4) are as follows

a_1: Do not invest

a_2: Select investment A the 6% bank savings plan.

a_3: Select investment B, the 4.5 % government bond.

a_4: Select investment C, the uncertain mutual fund

Observe that actions a_1 to a_3 do not involve uncertainty as the outcomes associated with them do not depend on uncertain market conditions. Observe also that action a 2 dominates actions a_1 and a_3. In addition, action a_1 is clearly inferior to the risk-free positive growth investment alternatives a_2 and a_3 as it provides for no growth of the principal amount.

Action a_4 is associated with an uncertain outcome that, depending on the state of the economy, may produce either a negative return or a positive return. Thus there exists no apparent dominance relationship between action a_4 and action a_2, the best among the actions involving no uncertainty.

Suppose the investor believes that if the market is down in the next year, an investment in the mutual fund would lose 10 percent returns; if the market stays the same, the investment would stay the same; and if the market is up, the investment would gain 20 percent returns. The investor has thus defined the states of nature for his/her investment decision-making problem as follows:

s_1: The market is down.

s_2: The market remains unchanged.

s_3: The market is up.

A study of the market combined with economic expectations for the coming year may lead the investor to attach subjective probabilities of 0.25, 0.25, and 0.50, respectively, the the states of nature, s_1, s_2, and s_3. The major question is then, how can the investor use the foregoing information regarding investments A, B, and C, and the expected market behaviour serves as an aid in selecting the investment that best satisfies his/her objectives? This question will be considered in the sections that follow.

3.2 ANALYSIS OF THE DECISION PROBLEM

In problems involving choices from many alternatives, one must identify all the actions that may be taken and all the states of nature whose occurrence may influence decisions. The action to take none of the listed alternatives whose outcome is known with certainty may also be included in the list of actions. Associated with each action is a list of payoffs. If an action does not involve risk, the payoff will be the same no matter which state of nature occurs.

The payoffs associated with each possible outcome in a decision problem should be listed in a *payoff table*, defined as a listing, in tabular form, of the value payoffs associated with all possible actions under every state of nature in a decision problem.

The payoff table is usually displayed in grid form, with the states of nature indicated in the columns and the actions in the rows. If the actions are labeled a_1, a_2, …, a_n, and the states of nature labeled s_1, s_2, …, s_k, a payoff table for a decision problem appears as in table 10.1 below. Note that a payoff is entered in each of the nk cells of the payoff table, one for the payoff associated with each action under every possible state of nature.

Table 3.1: The Payoff Table

STATE OF NATURE

ACTION	s_1	s_2	s_3	...	s_k
a_1					
a_2					
a_3					
. . .					
a_n					

Example

The managing director of a large manufacturing company is considering three potential locations as sites at which to build a subsidiary plant. To decide which location to select for the subsidiary plant, the managing director will determine the degree to which each location satisfies the company's objectives of minimising transportation costs, minimising the effect of local taxation, and having access to an ample pool of available semi-skilled workers. Construct a payoff table and payoff measures that effectively rank each potential location according to the degree to which each satisfies the company's objectives.

Solution

Let the three potential locations be sites A, B, and C. To determine a payoff measure to associate with each of the company's objectives under each alternative, the managing director subjectively assigns a rating on a 0 – to – 10 scale to measure the degree to which each location satisfies the company's objectives. For each objective, a 0 rating indicates complete dissatisfaction, while a 10 rating indicates complete dissatisfaction. The results are presented in table 3.2 below:

Table 3.2: Ratings for three alternative plant sites for a Manufacturing Company

	ALTERNATIVE		
COMPANY OBJECTIVE	Site A	Site B	Site C
Transportation Costs	6	4	10
Taxation Costs	6	9	5
Workforce Pool	7	6	4

To combine the components of payoff, the managing director asks himself, what are the relative measures of importance of the three company objectives I have considered as components of payoff? Suppose the managing director decides that minimising transportation costs is most important and twice as important as either the minimization of local taxation or the size of workforce available. He/she thus assigns a weight of 2 to the transportation costs and weights of 1 each to taxation costs and workforce. This will give rise to the following payoff measures:

Payoff (Site A) = 6(2) + 6(1) + 7(1) = 25

Payoff (Site B) = 4(2) + 9(1) + 6(1) = 23

Payoff (Site C) = 10(2) + 5(1) + 4(1) = 29

3.3 EXPECTED MONETARY VALUE DECISIONS

A decision-making procedure, which employs both the payoff table and prior probabilities associated with the states of nature to arrive at a decision is referred to as the *Expected Monetary Value* decision procedure. Note that by *prior probability* we mean probabilities representing the chances of occurrence of the identifiable states of nature in a decision problem prior to gathering any sample information. The *expected monetary value decision* refers to the selection of available action based on either the expected opportunity loss or the expected profit of the action.

Decision makers are generally interested in the *optimal monetary value decisions*. The optimal expected monetary value decision involves the selection of the action associated with

the minimum *expected opportunity loss* or the action associated with the maximum *expected profit*, depending on the objective of the decision maker.

The concept of expected monetary value applies mathematical expectation, where opportunity loss or profit is the random variable and the prior probabilities represent the probability distribution associated with the random variable.

The ***expected opportunity loss*** is computed by:

$$E(L_i) = \sum_{all\ j} L_{ij}P(s_j), \qquad (i = 1, 2, \ldots, n)$$

where L_{ij} is the opportunity loss for selecting action a_i given that the state of nature, s_j, occurs and $P(s_j)$ is the prior probability assigned to the state of nature, s_j.

The ***expected profits*** for each action is computed in a similar way:

$$E(\pi_i) = \sum_{all\ j} \pi_{ij}P(s_j)$$

where π_{ij} represents profits for selecting action a_i

Example

By recording the daily demand for a perishable commodity over a period of time, a retailer was able to construct the following probability distribution for the daily demand levels:

Table 3.3: Probability Distribution for the Daily Demand

s_j	$P(s_j)$
1	0.5
2	0.3
3	0.2
4 or more	0.0

The opportunity loss table for this demand-inventory situation is as follows:

Table 3.4: The Opportunity Loss Table

	State of Nature, Demand		
Action, Inventory	$s_1(1)$	$s_2(2)$	$s_3(3)$
$a_1(1)$	0	3	6
$a_2(2)$	2	0	3
$a_3(3)$	4	2	0

We are required to find the inventory level that minimises the expected opportunity loss.

Solution

Given the prior probabilities in the first table, the expected opportunity loss are computed as follows:

$E(L_i) = \sum_{j=1}^{3} L_{ij}P(s_j)$, *for each inventory level, I = 1, 2, 3.*

The expected opportunity losses at each inventory level become:

$E(L_1) = 0(0.5) + 3(0.3) + 6(0.2) =$ N2.10

$E(L_2) = 2(0.5) + 0(0.3) + 3(0.2) =$ N1.60

$E(L_3) = 4(0.5) + 2(0.3) + 0(0.2) =$ N2.60

It follows that in order to minimize the expected opportunity loss, the retailer should stock 2 units of the perishable commodity. This is the optimal decision.

4.0 CONCLUSION

In conclusion Decision Analysis is a limiting case of Administration problems, it can be applied in cases when the number is very large tending towards infinity and the probability of success is very low.

5.0 SUMMARY

In this unit, student must have learnt the rudiments and applications of Administrative and Decision Analysis. Students are must have learnt how to solve problems using Decision Analysis.

6.0 TUTOR-MARKED ASSIGNMENT

1. Give the justification for using an expected monetary value objective in decision problems.

2. Write short note on the following:

- Critical Examinations
- Brain Storming or Group Creativity
- Analogies
- Morphological Approach or Attribute Listening
- Heuristic Approach

3. The following table shows a set of utility values that have been assessed for the associated Naira-valued outcomes by a decision maker. If the decision maker wishes to maximise his/her expected utility, how should he/she act on each of the following investment problems?

Naira-Valued Outcome	Utility
- N10,000	0
- N5,000	0.45
- N1,000	0.50
N0.00	0.55
N5,000	0.70
N10,000	0.80
N25,000	1.0

(a) The investment of N1,000 is an Oil drilling venture returning either a N10,000 profit or nothing. The probability of success in the Oil drilling venture is estimated to be 10 percent.

(b) The investment of N10,000 is in a new Hotel-Restaurant facility. Depending on the success of the project, the investment is expected to return a N25,000 profit with a probability of 20 percent, a N5,000 profit with a probability of 30 percent, a N5,000 loss with a probability of 40 percent, or a loss of the entire N10,000 investment with a probability of 10 percent.

(c) In both of the above investment problems, compare the optimal decision using a maximum expected utility objective with the optimal decision using a maximum expected payoff objective. How do you account for any differences in the selection of an optimal decision between these two objectives?

7.0 REFERENCES / FURTHER READING

ONWE J.O. NOUN TEXT BOOK, MBF 839: Quantitative Methods for Banking & Finance.

JUDE, MICAN & EDITH N. Statistical & Quantitative Methods for Construction & Business Managers

MODULE 2 Index Numbers and Introduction to Research Methods in Management Sciences

Unit 1: Index Number
Unit 2: Statistical Data
Unit 3: Sample and Sampling Techniques
Unit 4: Estimation Theory

UNIT 1: INDEX NUMBER

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Uses of index numbers

3.2 Types of index number

3.3 Problems encountered in the construction of index numbers

3.4 Methods of constructing index numbers

4.0 Conclusion
5.0 Summary
6.0 Assignment
7.0 References / Further Reading

1.0 INTRODUCTION

Index numbers are indicators which reflect the relative changes in the level of certain phenomenon in any given period (or over a specified period of time) called the current period with respect to its value in some fixed period called the base period selected for comparison. The phenomenon or variable under consideration may be price, volume of trade, factory production, agricultural production, imports or exports, shares, sales, national income, wage

structure, bank deposits, foreign exchange reserves, cost of living of people of a particular community etc.

2.0 OBJECTIVE

The main objective of this unit is to provide students with good understanding of index numbers and its applications in statistics and business management.

3.0 MAIN CONTENT

3.1 Uses of Index Number

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

3.2 Types of Index Numbers

Index number may be classified in terms of the variables they measure. They are generally classified into three categories:

1. **Price Index Number:** The most common index numbers are the price index numbers which study changes in price level of commodities over a period of time. They are of two types:
   (a) **Wholesale price index number** – They depict changes in the general price level of the economy.
   (b) **Retail Price Index Number –** The reflect changes in the retail prices of different commodities. The are normally constructed for different classes of consumers.
2. **Quantity Index Number –** They reflect changes in the volume of goods produced or consumed
3. **Value Index Number** – They study changes in the total value (price X quantity) e.g index number of profit or sales.

3.3 Problems in the construction of Index Numbers

1. The purpose of index number – This must be carefully defined as there is no genera purpose index number.

2. Selection of base period – The base period is the previous period with which comparison of some later period is made. The index of the base period is taken to be 100. The following points should be borne in mind while selecting a base period:
   (a) Base period should be a normal period devoid of natural disaster, economic boom, depression, political instability, famine etc.
   (b) The base period should not be too distant from the given period. This is because circumstances such as tastes customs, habits and fashion keep changing.
   (c) One must determine whether to use fixed-base or chain-base method

(1) Selection of commodities – Commodities to be selected must be relevant to the study; must not be too large nor too small and must be of the same quality in different periods.

(2) Data for the index number- Data to be used must be reliable.

(3) Type of average to be used – ie, arithmetic, geometric, harmonic etc.

(4) Choice of formula – There are different types of formulas and the choice is mostly dependent on available data.

(5) System of weighting – Different weights should be assigned to different commodities according to their relative importance in the group.

3.4 Methods of constructing index numbers

(1) **Simple (unweighted) Aggregate Method** – Aggregate of prices (of all the selected commodities) in the current year as a percentage of the aggregate of prices in the base year.

$P_{01} \rightarrow$ Price index number in the current year with respect to the base year

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

Limitations of the Simple Aggregate Method

(a) The prices of various commodities may be quoted in different units

(b) Commodities are weighted according to the magnitude of their price. Therefore, highly priced commodity exerts a greater influence than lowly priced commodity. Therefore, the method is dominated by commodities with higher prices.

(c) The relative importance of various commodities is not taken into consideration

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

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

Exercise: From the following data calculate Index Number by Simple Aggregate method.

Commodity	**A**	**B**	**C**	**D**
Price 2011	81	128	127	66
Price 2012	85	82	95	73

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

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

$$= 83.3\ \%$$

(2) Weighted Aggregate Method - In this method, appropriate weights are assigned to various commodities to reflect their relative importance in the group. The weights can be production figures, consumption figure or distribution figure

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

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

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

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

This formula was invented by French economist Laspeyre in 1817.

(ii) **Paasche's Price Index –** Here, the current year quantities are taken as weights and we obtain:

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

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

(i) Dorbish-Bowley Price Index – This index is given by the arithmetic mean of Laspeyre's and Paasche's price index numbers.It is also sometimes known as L-P formula:

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

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

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

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

Example 1: Consider the table below which gives the details of price and consumption of four commodities for 2010 and 2012. Using an appropriate formula calculate an index number for 2012 prices with 2010 as base year.

Commodities	**Price per unit 2010 (₦)**	**Price per unit 2012 (₦)**	**Consumption value 2010 (₦)**
Yam flour	70	85	1400
Vegetable oil	45	50	720
Beans	90	110	900
Beef	100	125	600

Solution: In the above problem, we are given the base year (2010) consumption values ($p_o q_o$) and current year quantities (q_1) are not given, the appropriate formula for index number here is the Laspeyre's Price Index.

Commodities	**Price per unit 2010 (₦) p_0 (1)**	**Consumption value 2010 (₦) p_oq_o(2)**	**Price per unit 2012 (₦) p_1 (3)**	**2010 quantities $q_0 = = \frac{(2)}{(1)}$**	P_1q_0
Yam flour	70	1400	85	20	1700
Vegetable oil	45	720	50	16	800
Beans	90	900	110	10	1100
Beef	100	600	125	6	750
		$\sum P_0q_0 = 3620$			$\sum P_1q_0 = 4350$

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

$$P_{01}^{La} = \frac{\sum P_1q_0}{\sum P_0q_0} X \frac{100}{1}$$

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

$$= = 120.1657 \cong 120.17$$

Example 2: From the following data calculate price index for 2012 with 2007 as the base year by (i) Laspeyre's method (ii) Pasche's method (iii) Fisher's method and

(iii) Dowbish-Bowley price index methods

Commodities	***2007***		***2012***	
	Price	Quantity	Price	Quantity
Gaari	20	8	40	6
Rice	50	10	60	5
Fish	40	15	50	15
Palm-oil	20	20	20	25

Solution:

Commodities	***2007***		***2012***					
	Price (p_o)	Quantity (q_o)	Price (p_1)	Quantity (q_1)	p_oq_o	p_oq_1	p_1q_o	p_1q_1
Gaari	20	8	40	6	160	120	320	240
Rice	50	10	60	5	500	250	600	300
Fish	40	15	50	15	600	600	750	750
Palm-oil	20	20	20	25	400	500	400	500
Total					$\boldsymbol{p_oq_o}$= ***1660***	$\boldsymbol{p_oq_1}$= ***1470***	$\boldsymbol{p_1q_o}$= ***2070***	$\boldsymbol{p_1q_1}$= ***1790***

Laspeyre's Price Index

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

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

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

$$= 124.7$$

(i) Pasche's Price Index

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

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

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

$$= 121.77$$

(ii) Fisher's Price Index

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

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

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

$$= 123.23$$

(iii) Dorbish-Bowley Price Index

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

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

= ½ [1.247 + 1.2177] X100

= 1.23235 X 100

= 123.24

4.0 CONCLUSION

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

5.0 SUMMARY

In this unit, we have been able to introduce students to the concept of index numbers, its uses and methods of calculation. Students are now expected to be proficient in the calculation, use and interpretation of index numbers. This is useful in the study and interpretation of inflation, cost of living, trends of economic variables among others.

6.0 TUTOR-MARKED ASSIGNMENT

1. Calculate Price index number of the year2010 with 2000 as the base year from the following data using:
   (i) Laspeyre's
   (ii) Pasche's
   (iii) Dorbish-Bowley and
   (iv) Fisher's formulae

Commodity		**2000**		**2010**	
	Unit	**Price (₦)**	**Value (₦)**	**Quantity consume**	**Value (₦)**
Rice	Kg	20	3000	320	3520
Gari	Kg	24	2160	200	2600
Cloth	Yards	30	1800	120	1920
Sugar	Packets	18	900	80	960

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

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

7.0 REFERENCES/FURTHER READING

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

Lucey T. (2002). *Quantitative Techniques*.(6thed.). BookPower

UNIT 2: STATISTICAL DATA

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/Further Reading

1.0 INTRODUCTION

Statistics deals with the theories and methods of collection, presentation, analysis, and interpretation of numerical data.

2.0 OBJECTIVES

In general, the objective for you the student here is to make you appreciate the purpose of statistical tests and data, in the determinationof whether some hypotheses are extremely unlikely given observed data.

3.0 MAIN CONTENT

Types of Data

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

(1) Based on sources – Data can be classified base on the sources from which they are obtained. In this regards, we have:

(a) **Primary data** – These are data collected directly from the field of enquiries by the user(s) themselves.

Advantages – They are always relevant to the subject under study because they are collected primarily for the purpose.

- They are more accurate and reliable
- Provide opportunity for the researcher to interact with study population.
- Information on other relevant issues can be obtained

Disadvantages – Always costly to collect

- Inadequate cooperation from the study population
- Wastes a lot of time and energy

(b) **Secondary Data:** These are data which have been collected by someone else or some organization either in published or unpublished forms.

Advantages: - It is easier to get

- It is less expensive

Disadvantages:-May not completely meet the need of the research at hand because it was not collected primarily for that purpose

- There is always a problem of missing periods

(2) **Classification based on form of the data:** Sometimes, data are classified based on the form of the data at hand and may be classified as:

(a) **Cross-sectional data –** These are data collected for cross-section of subjects (population under study) at a time. For example, data collected on a cross-section of household on demand for recharge card for the month of August 2013.

(b) **Time-series data –** These are data collected on a particular variable or set of variables over time e.g a set Nigeria's Gross Domestic Product (GDP) values form 1970 to2012.

(c) **Panel Data –** These combine the features of cross-sectional and time-series data. They are type of data collected from the same subjects over time. For example, a set of data collected on monthly recharge card expenditure from about 100 households in Lagos from January to December 2013 will form a panel data.

Note that Social and Economic data of national importance are collected routinely as by-product of governmental activities e.g. information on trade, wages, prices, education, health, crime, aids and grants etc.

Sources of Data

1. **Source of Primary data:**
   - (i) Census
   - (ii) Surveys
2. **Sources of Secondary data:**

(i) Publications of the Federal Bureau of statistics

(ii) Publications of Central Bank of Nigeria

(iii) Publications of National population commission

(iv) Nigerian Custom Service

(v) Nigeria Immigration Service

(vi) Nigerian Port Authority

(iv) Federal and State Ministries, Departments and Agencies

Some of the publications referred to above are:

(i) Annual Digest of statistics (by NBS)

(ii) Annual Abstract of statistics (by NBS)

(iii) Economic and Financial Review (by CBN)

(iv) Population of Nigeria (by NPC)

4.0 CONCLUSION

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

5.0 SUMMARY

This unit has acquainted you with the transformation of the processed data into statistics and steps in the statistical cycle. The transformation involves analysis and interpretation of data to identify important characteristics of a population and provide insights into the topic being investigated.

6.0 TUTOR-MARKED ASSIGNMENT

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

7.0 REFERENCES / FURTHER READINGS

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

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

UNIT 3: SAMPLE AND SAMPLING TECHNIQUES

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

Researchers collect data in order to test hypotheses and to provide empirical support for explanations and predictions. Once investigators have constructed their measuring instrument in order to collect sufficient data pertinent to the research problem, the subsequent explanations and predictions must be capable of being generalised to be of scientific value. Generalizations are important not only for testing hypotheses but also for descriptive purposes. Typically, generalizations are not based on data collected from all the observations, all the respondents, or the events that are defined by the research problem as this is always not possible or where possible too expensive to undertake. Instead, researchers use a relatively small number of cases (a sample) as the bases for making inferences for all the cases (a population).

2.0 OBJECTIVES

The objective here is to make an awareness of how the art of sampling is a very valuable tool in collecting of data for planning and decision making.

3.0 MAIN CONTENT

Empirically supported generalizations are usually based on partial information because it is often impossible, impractical, or extremely expensive to collect data from all the potential units of analysis covered by the research problem. Researchers can draw precise inferences on all the units (a set) based on relatively small number of units (a subset) when the subsets accurately represent the relevant attributes of the whole sets. For example, in a study of patronage of campus photographer among students in a university, it may be very expensive and time consuming to reach out to all students (some universities have as high as 40,000

students). A careful selection of relatively small number of students across faculties, departments and levels will possibly give a representation of the entire student population.

The entire set of relevant units of analysis, or data is called the population. When the data serving as the basis for generalizations is comprised of a subset of the population, that subset is called a **sample**. A particular value of the population, such as the mean income or the level of formal education, is called **a parameter**; its counterpart in the sample is termed the **statistic**. The major objective of sampling theory is to provide accurate estimates of unknown values of the parameters from sample statistics that can be easily calculated. To accurately estimate unknown parameters from known statistics, researchers have to effectively deal with three major problems:

(1) the definition of the population,

(2) the sample design, and

(3) the size of the sample.

Population

Methodologically, a population is the "aggregate of all cases that conform to some designated set of specifications". For example, a population may be composed of all the residents in a specific neighbourhood, legislators, houses, records, and so on. The specific nature of the population depends on the research problem. If you are investigating consumer behaviour in a particular city, you might define the population as all the households in that city. Therefore, one of the first problems facing a researcher who wishes to estimate a population value from a sample value is how to determine the population involved.

The Sampling Unit

A single member of a sampling population (e.g a household) is referred to as a sampling unit. Usually sampling units have numerous attributes, one or more of which are relevant to the research problem. The major attribute is that it must possess the typical characteristics of the

study population. A sampling unit is not necessarily an individual. It can be an event, a university, a city or a nation.

Sampling Frame

Once researchers have defined the population, they draw a sample that adequately represents that population. The actual procedures involve in selecting a sample from a sample frame comprised of a complete listing of sampling units. Ideally, the sampling frame should include all the sampling units in the population. In practice, a physical list rarely exists; researchers usually compile a substitute list and they should ensure that there is a high degree of correspondence between a sampling frame and the sampling population. The accuracy of a sample depends, first and foremost, on the sampling frame. Indeed, every aspect of the sample design – the population covered, the stages of sampling, and the actual selection process – is influenced by the sampling frame. Prior to selecting a sample, the researcher has to evaluate the sampling frame for potential problems.

Sample Design

The essential requirement of any sample is that it be as representative as possible of the population from which it is drawn. A sample is considered to be representative if the analyses made using the researcher's sampling units produce results similar to those that would be obtained had the researcher analysed the entire population.

Probability and Non-probability Sampling

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

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

population. When a researcher is using a probability sample design, it is possible for him or her to estimate the population's parameters on the basis of the sample statistics calculated.

Non-probability Sample Designs

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

Convenience samples: Researchers obtain a convenience sample by selecting whatever sampling units are conveniently available. Thus a University professor may select students in a class; or a researcher may take the first 200 people encountered on the street who are willing to be interviewed. The researcher has no way of estimating the representativeness of convenience sample, and therefore cannot estimate the population's parameters.

Purposive samples: With purposive samples (occasionally referred to as judgement samples), researchers select sampling units subjectively in an attempt to obtain a sample that appears to be representative of the population. In order words, the chance that a particular sampling unit will selected for the sample depends on the subjective judgement of the researcher. At times, the main reason for selecting a unit in purposive sampling is the possession of pre-determined characteristic(s) which may be different from that the main population. For example, in a study of demand preference for cigarette brands in a city, researcher will need to select smokers purposively.

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

Probability Sample Designs

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

Simple random sampling – is the basic probability sampling design, and it is incorporated into all the more elaborate probability sampling designs. Simple random sampling is a procedure that gives each of the total sampling units of the population an equal and known nonzero probability of being selected. For example, when you toss a perfect coin, the

probability that you will get a head or a tail is equal and known (50 percent), and each subsequent outcome is independent of the previous outcomes.

Random selection procedures ensure that every sampling unit of the population has an equal and known probability of being included in the sample; this probability is n/N, where n stands for the size of the sample and N for the size of the population. For example if we are interested in selection 60 household from a population of 300 households using simple random sampling, the probability of a particular household being selected is 60/300 = 1/5.

Systematic Sampling: It consists of selecting every k^{th} sampling unit of the population after the first sampling unit is selected at random from the total of sampling units. Thus if you wish to select a sample of 100 persons from total population of 10,000, you would take every hundredth individual (K=N/n = 10,000/100 = 100). Suppose that the fourteenth person were selected; the sample would then consist of individuals numbered 14,114, 214, 314, 414, and so on. Systematic sampling is more convenient than simple random sampling. Systematic samples are also more amenable for use with very large populations or when large samples are to be selected.

Stratified Sampling:Researchers use this method primarily to ensure that different groups of population are adequately represented in the sample. This is to increase their level of accuracy when estimating parameters. Furthermore, all other things being equal, stratified sampling considerably reduces the cost of execution. The underlying idea in stratified sampling is to use available information on the population "to divide it into groups such that the elements within each group are more alike than are the elements in the population as a whole. That is, you create a set of homogeneous samples based on the variables you are interested in studying. If a series of homogenous groups can be sampled in such a way when the samples are combined they constitute a sample of a more heterogeneous population, you will increase the accuracy of your parameter estimates.

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

Sample size

A sample is any subset of sampling units from a population. A subset is any combination of sampling units that does not include the entire set of sampling units that has been defined as the population. A sample may include only one sampling unit, or any number in between.

There are several misconceptions about the necessary size of a sample. One is that the sample size must be certain proportion (often set as 5 percent) of the population; another is that the sample should total about 2000; still another is that any increase in the sample size will increase the precision of the sample results. These are faulty notions because they do not derive from the ***sampling theory***. To estimate the adequate size of the sample properly, researchers need to determine what level of accuracy is expected of their estimates; that is, how large a standard error is acceptable.

Standard error

Some people called it *error margin*or*sampling error.* The concept of standard error is central to sampling theory and to determining the size of a sample. It is one of the statistical measures that indicate how closely the sample results reflect the true value of a parameter.

Methods of data collection

There are three methods of data collection with survey and these are mail questionnaires, personal interviews, and telephone interviews.

Mail questionnaire: It is an impersonal survey method. Here, survey instrument (the questionnaire) is mailed to the selected respondents and the questionnaires are mailed back to the researcher after the respondents must have filled it up. This is very common in developed countries where the citizens appreciate the relevance of data and research. Under certain conditions and for a number of research purposes, an impersonal method of data collection can be useful.

Advantages and disadvantages of mail questionnaires

Advantages

- The cost is low compared to others
- Biasing error is reduced because respondents are not influenced by interviewed characteristics or techniques.
- Questionnaires provide a high degree of anonymity for respondents. This is especially important when sensitive issues are involved.

- Respondents have time to think about their answers and /or consult other sources.
- Questionnaires provide wide access to geographically dispersed samples at low cost

Disadvantages

- Questionnaires require simple, easily understood questions and instructions
- Mail questionnaires do not offer researchers the opportunity to probe for additional information or to clarify answers.
- Researchers cannot control who fills out the questionnaire.
- Response rate are low

Factors affecting the response rate of mail questionnaires

Researchers use various strategies to overcome the difficulty of securing an acceptable response rate to mail questionnaires and to increase the response rate.

- Sponsorship: The sponsorship of a questionnaire motivates the respondents to fill the questionnaires and return them. Therefore, investigators must include information on sponsorship, usually in the cover letter accompanying the questionnaire.
- Inducement to response: Researchers who use mail surveys must appeal to the respondents and persuade them that they should participate by filling out the questionnaires and mailing them back. For example, a student conducting a survey for a class project may mention that his or her grade may be affected by the response to the questionnaire.
- Questionnaire format and methods of mailing- Designing a mail questionnaire involves several considerations: typography, colour, and length and type of cover letter.

Personal interview

The personal interview is a face-to-face, interpersonal role situation in which an interviewer asks respondents question designed to elicits answers pertinent to the research hypotheses. The questions, their wording, and their sequence define the structure of the interview.

Advantages of personal interview

- Flexibility: The interview allows great flexibility in the questioning process, and the greater the flexibility, the less structure the interview. Some interviews allow the interviewer to determine the wording of the questions, to clarify terms that are unclear, to control the order in which the question are presented, and to probe for additional information and details.
- Control of the interview situation: An interviewer can ensure that the respondents answer the questions in the appropriate sequence or that they answer certain questions before they ask subsequent questions.
- High response rate: The personal interview results in a higher response rate than the mail questionnaire.
- Fuller information: An interviewer can collect supplementary information about respondents. This may include background information, personal characteristics and their environment that can aid the researcher in interpreting the results.

Disadvantages of the personal interview

- Higher cost: The cost of interview studies is significantly higher than that of mail survey. Costs are involved in selecting, training, and supervising interviewers; in paying them; and in the travel and time required to conduct interviews.
- Interviewer bias: The very flexibility that is the chief advantage of interviews leaves room for the interviewer's personal influence and bias.
- Lack of anonymity: The interview lacks the anonymity of the mail questionnaire. Often the interviewer knows all or many of the potential respondents (their names, addresses, and telephone numbers). Thus respondents may feel threatened or intimidated by the interviewer, especially if a respondent is sensitive to the topic or some of the questions.

Telephone interview

It is also called telephone survey, and can be characterised as a semi-personal method of collecting information. In comparison, the telephone is convenient, and it produces a very significant cost saving.

Advantages of Telephone interview

- Moderate cost
- Speed: Telephone interviews can reach a large of respondents in a short time. Interviewers can code data directly into computers, which can later compile the data.
- High response rate: Telephone interviews provide access to people who might be unlikely to reply to a mail questionnaire or refuse a personal interview.
- Quality: High quality data can be collected when interviewers are centrally located and supervisors can ensure that questions are being asked correctly and answers are recorded properly.

Disadvantages of Telephone interview

- Reluctant to discuss sensitive topics: Respondents may be resistant to discuss some issues over the phone.
- The "broken off" interview: Respondents can terminate the interview before it is completed.
- Less information Interviewers cannot provide supplemental information about the respondents' characteristics or environment.

4.0 CONCLUSION

This unit has relayed to you that a well-chosen sample can usually provide reliable information about the whole of the population to any desired degree of accuracy. In some instances sampling is an alternative to a complete census, and may be preferable mainly because of its cheapness and convenience.

5.0 SUMMARY

You now would be able to discern that a sample is a subset of a population selected to meet specific objectives. And also familiar with the guiding principle and sampling techniques in selecting a sample, is that it must, as far as possible have the essential characteristics of the target population.

6.0 TUTOR-MARKED ASSIGNMENT

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

7.0 REFERENCES/FURTHER READINGS

OKOJIE, DANIEL E. NOUN TEXT BOOK, Eco 203: Statistics for Economists

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

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

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

Unit 4: ESTIMATION THEORY

CONTENTS

1.0 Introduction
2.0 Objective
3.0 Main Content
    3.1 Methods of Point Estimation
    3.2 Method of Maximum likelihood
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 Introduction

Point Estimation when a single numerical value of the statistic is used as an estimate of the exact population value, we have a point or target estimate. An estimate is value of the sample statistic which is taken as an approximation of the parameter value. An estimator refers to the formula or statistic which has been chosen to provide an estimate of the population value. The mean, mode, medium, variance etc are examples of point estimates. In any population distribution with mean j and variance o the corresponding estimators are the sample mean and sample variance given as

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

Note that the estimators are functions of the random samples which do not depend on the parameters.

2.0 OBJECTIVES

The main objective of this unit is to enable students understand the theory behind and the application of estimation in statistics. Students are expected at the end of this unit to be able to apply estimation theory to solving day-to-day business and economic problems

3.0 Main Content

3.1 Methods of Point Estimation

The following are methods of obtaining point estimators of the population parameter.

Method of maximum likelihood

Method of least squares

Method of moments

Method of moment generating function.

3.2 Method of Maximum Likelihood: Let x x x,, be a random sample of size n from a population ,vith pdf f(x, 0). The likelihood function is the function of the sample values x 2' ...x, which expresses the joint probability of occurrence of the sample values. That is, the likelihood of the random samples is the product of their respective probability distribution.

$$L(\theta; x_1, x_2, \cdots x_n) = f(x_1, \theta)\, f(x_2, \theta)\, f(x_3, \theta) \ldots\ldots f(x_n, \theta)$$
$$= \prod_{i=1}^{n} f(x_i, \theta).$$

The Maximum Likelihood Estimator (MLE) of 0 based on a random sample x x ... x is the value of 0 which maximizes the likelihood function L(0; x x ...

Since any positively valued function attains a maximum at the same point as its logarithm function, we obtain the m.l.e usually by maximizing the natural logarithm of the likelihood.

Given

$$L(\theta; x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} f(x_i, \theta)$$

then

$$\ln L(\theta; x_1, x_2, \cdots x_n) = \Sigma \ln f(x_i, \theta).$$

We then maximize the log likelihood function by differentiating partially with respect to θ and equating to zero.

Example 13.1.1

Given the Bernoulli distribution defined as:

$$f(x, p) = P^x(1-P)^{1-x} \qquad x = 0, 1 \quad \text{and} \quad 0 < P < 1$$

P is the probability of success.

The m.l.e of P is obtained as followed

$$L(p, x) = \prod_{i=1}^{n} f(x_i, p)$$

$$= p^{\Sigma x}(1-p)^{n-\Sigma x}$$

$$\text{and} \quad \ln L(p, x) = \Sigma x \log p + (n - \Sigma x)\log(1-p).$$

$$\frac{\partial \ln L}{\partial p} = \frac{\Sigma x}{p} - \frac{n - \Sigma x}{1-p} = 0$$

multiplying by the Lcm we have

$$(1-p)\Sigma x - p(n - \Sigma x) = 0$$
$$\Sigma x - p\Sigma x - np + p\Sigma x = 0$$
$$\Sigma x - np = 0$$
$$\Sigma x = np$$
$$\hat{p} = \frac{\Sigma x}{n}.$$

Example 13.1.2.

Suppose in a Bernoulli trial of hitting a target, the following results were obtained x_i = 0, 1, 0, 0, 1, 1, 1, 1, 0, where 1 is hitting the target and 0 is missing the target. Then P, the probability of hitting the target is estimated as

$$\hat{p} = \frac{\Sigma x}{n} = \frac{1+1+1+1+1}{9} = \underline{5/9}$$

Estimation in A Binomial Population

Let us define x as the number of success in n-trials such that **p** is the probability of success. To obtain the maximum likelihood of **p**, then in the **n** trials, the likelihood is

$$B(x, n, p) = L(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$x = 0, 1, 2, \cdots, n.$$

$$\therefore \ln L = \log\binom{n}{x} + x\log P + (n-x)\log(1-P)$$

$$\frac{\partial \ln L}{\partial P} = \frac{x}{p} - \frac{n-x}{1-p} = 0$$

$$x(1-p) - p(n-x) = 0$$

$$x - xp - np + xp = 0$$

$$x - np = 0$$

$$np = x$$

$$\sum y - n\hat{\alpha} - \hat{\beta}\Sigma x = 0$$

$$\Rightarrow \quad n\hat{\alpha} = \Sigma y - \hat{\beta}\Sigma x$$

$$\Rightarrow \hat{\alpha} = \frac{\Sigma y - \hat{\beta}\Sigma x}{n}$$

Also

$$\frac{\partial Q}{\partial \beta} = -2\Sigma\left(y - \hat{\alpha} - \hat{\beta}x\right)x = 0$$

$$\Rightarrow \Sigma yx - \hat{\alpha}\Sigma x - \hat{\beta}\Sigma x^2 = 0$$

Substituting $\hat{\alpha}$ we have

$$\Rightarrow \quad \Sigma yx - \left(\frac{\Sigma y - \hat{\beta}\Sigma x}{n}\right)\Sigma x - \hat{\beta}\Sigma x^2 = 0$$

$$\therefore \quad \hat{\beta} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

13.1.5. Method of Moments

The rth moment of a random variable about the origin **A** is given as

$$\mu_r^1 = \frac{\Sigma(X-A)^r}{N}$$

The rth moment about the origin **O** is give as

$$\mu_r^1 = \frac{\Sigma X^r}{N}$$

Note $\mu_1^1 = \frac{\Sigma X}{N} = \bar{X} = \mu =$ population mean.

The rth moment about the mean is given as

$$\mu_r = \frac{\Sigma\left(x-\bar{x}\right)^r}{N} = \frac{\Sigma(x-\mu)^r}{N}$$

Note for $\mathbf{r = 2}$

$$\mu_2 = \frac{\Sigma\left(x-\bar{x}\right)^2}{N} = \sigma^2 = \textit{population variance.}$$

Also $\mu_2 = \mu_2^1 - \mu_1^2$

$$= \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2 = \underline{\sigma^2}$$

13.1.5. Method of Moments

The rth moment of a random variable about the origin **A** is given as

$$\mu_r^1 = \frac{\Sigma(X-A)^r}{N}$$

The rth moment about the origin **O** is give as

$$\mu_r^1 = \frac{\Sigma X^r}{N}$$

Note $\mu_1^1 = \frac{\Sigma X}{N} = \bar{X} = \mu =$ population mean.

The rth moment about the mean is given as

$$\mu_r = \frac{\Sigma\left(x-\bar{x}\right)^r}{N} = \frac{\Sigma(x-\mu)^r}{N}$$

Note for $\mathbf{r = 2}$

$$\mu_2 = \frac{\Sigma\left(x-\bar{x}\right)^2}{N} = \sigma^2 = \textit{population variance.}$$

Also $\mu_2 = \mu_2^1 - \mu_1^2$

$$= \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2 = \underline{\sigma^2}$$

Example: Given the data, 5, 8, 3, 4, 6, 1. Obtain: (i) first and second noncentral moment (ii) second central moment, (iii) 4 moment about zero, (second moment about 5.

Solution:

(i) *First moment about zero is given as*

$$\mu_1^1 = \frac{\Sigma X^1}{N} = \frac{5+8+3+4+6+1}{6} = \frac{27}{6} = 4.5$$

$= \mu$ = mean.

(ii) *Second moment about zero is given as*

$$\mu_2^1 = \frac{\Sigma X^2}{N} = \frac{5^2+8^2+3^2+4^2+1^2+6^2}{6} = \frac{151}{6} = 25.167$$

(iii) *The second central moment is obtained as*

$$\mu_2 = \frac{\Sigma(X-\bar{X})^2}{N} = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$$

$$= \mu_2^1 - (\mu_1^1)^2$$

$$= 25.167 - (4.5)^2$$

$$= 4.9167$$

Note that μ_2 is the variance.

(iv) The 4th moment about zero is obtained as:

$$\mu_4^1 = \frac{\Sigma X^4}{N} = \frac{5^4+8^4+3^4+4^4+1^4+6^4}{6} = \frac{6355}{6}$$

$$= 1059.167$$

(v) *Second moment about 5 is*

$$\mu_2^1 = \frac{\Sigma(X-5)^2}{N} = \frac{(5-5)^2+(8-5)^2+(3-5)^2+(4-5)^2+(1-5)^2+(6-5)^2}{6}$$

$$= \frac{0+9+4+1+16+1}{6} = \frac{31}{6} = 5.1667$$

Generally, the rth moment of a random variable X about the mean or the rth central moment is given as:

That is if the random variable X has pdf f(x), then

$$\mu_r = \sum_{j=1} (x_j - \mu)^r f(x_j) \qquad \text{if } \mathbf{X} \text{ is discrete}$$

$$= \int_{-\infty}^{\infty} (x_j - \mu)^r f(x_j)dx \qquad \text{if } \mathbf{X} \text{ is continuous}$$

If **X** is discrete and $f(x) = \frac{1}{N}$, then

$$\mu_r = E\left[(x-\mu)^r\right]$$

$$= \sum_{j=1}^{n} (x_j - \mu)^r \left(\frac{1}{N}\right)$$

$$= \frac{\sum_{j=1}^{n} (x-\mu)^r}{N}$$

4.0 CONCLUSION

This unit has relayed to you that a well-chosen estimation can usually provide reliable information about the whole of the population to any desired degree of accuracy. In some

instances estimation is an alternative to a complete census, and may be preferable mainly because of its cheapness and convenience.

5.0 SUMMARY

You now would be able to discern that a estimation theory is a subset selected to meet specific objectives. And also familiar with the guiding principle and estimation techniques in selecting formula, is that it must, as far as possible have the essential characteristics of the target estimation.

6.0TUTOR-MARKED ASSIGNMENT

Given the data, 3, 8, 5, 1, 6, 4. Obtain: (i) first and second non-central moment (ii) second central moment.

7.0 References/ Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

JUDE I. EZE Statistics &Quantitative Methods for Construction &Business Managers

MODULE 3: CORRELATION AND REGRESSION ANALYSIS

UNIT 1: CORRELATION THEORY

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
    3.1 Perfect Positive Correlation
    3.2 Perfect Negative Correlation
    3.3 Strong Positive Correlation
    3.4 Strong Negative Correlation
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

Correlation can be defined as the branches of statistics that deals with mutual dependence or inter-relationship of two or more variables. If the value of two variables such that when one changes, the other changes too, then the variable are said to be correlated.

Generally, correlation implies that variation in one variable, when there is a variation in other variable.

Note that the degree of relationship which exist between two variables. The degree of relationship existing between two variables is called simple correlation. While the degree of relationship that connected three or more variables together is called Multiple correlation.

2.0

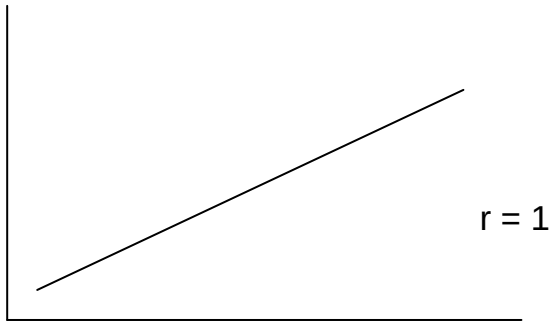
2.0 OBJECTIVES

The main objective of this unit is to enable students understand the theory behind and the application of correlation in statistics. Students are expected at the end of this unit to be able to apply correlation theory to solving day-to-day business and economic problems.

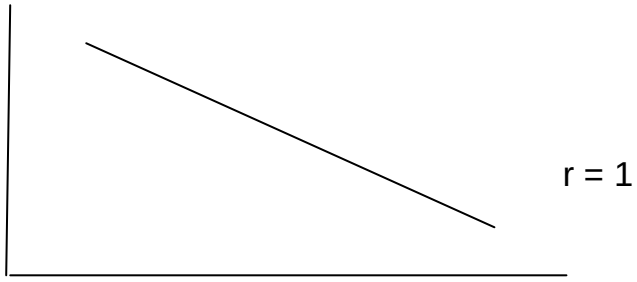
3.0 MAIN CONTENT

3.1 Perfect Positive Correlation

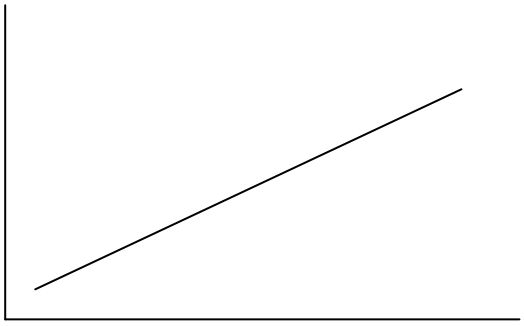
This can be defined as the situation where all the scatter points passes through a straight line none of the points deviated from the normal curve and positive slope.



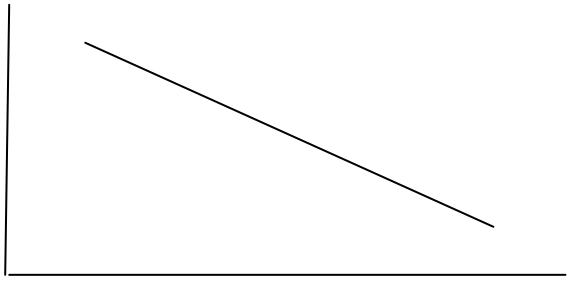
3.2 **Perfect Negative Correlation:** This indicates that all the points passes through the normal straight line and non deviated from the line. The curve shown downward slope of units.



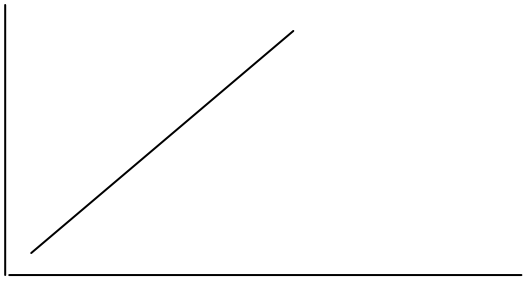
3.3 **Strong Positive Correlation:** In these case, most of the scatter points passes through the straight line, although there are few deviation from the straight line, but the deviation are very close to each other.



3.4 **Strong Negative Correlation:** In a strong negative correlation, some of the points passes through the straight line and all other scatter point are very close to the straight line, it has a negative slope which is very close to unity.



3.5 **Weak positive correlation:** In these case the points are deviated from each other so that each of the scatter points are for the depart from each other and the association is weak. The slope is positive and not close, to unity.



3.6 **Weak negative correlation:** In a weak negative correlation, there are serious deviations of scatter points and the points slope downward. It has a negative slope and not close to unity.

3.7 **No Correlation:** The scatter point at random and did not form any regular pattern for recognization by any straight line. There is no association between the variables.

4. Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5. Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter 〉 referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

6. Tutor-Marked Assignment

1. Explain with the use of diagram different types of correlation
2. Differentiate between strong positive correlation and negative correlation.

7. References /Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 2: PEARSON'S CORRELATION COEFFICIENT

CONTENTS

1.0 Introduction
2.0 Objective
3.0 Main Content
4.0 Conclusion
5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Reading

1.0 INTRODUCTION

Coefficient of correlation refers as the ratio of covariance between the related variables to the square root of the product of individual variance.

2.0 OBJECTIVE

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

- describe the computation of linear correlation coefficients
- apply the concept of correlations in business decisions.

3.0 MAIN CONTENT

Given a bivariate set of data x, y; x, y; y2 ---- x, y,

To obtain the general representations of product moment correlation coefficient as

$$r : \quad r = \frac{Sxy}{Sx\ Sx}$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 - \sum y^2}}$$

Where,

$X = x - x$

$Y = y - y$ respectively

From above equation, substitutes for x and y

$$r = \frac{\sum (x - x)(y - y)}{\sqrt{\sum (x - x)^2 - \sum (y - y)^2}}$$

$$r = \frac{\sum (x - x)(y - y)}{\sqrt{\sum (x - x)(x - x) - \sum (y - y)(y - y)}}$$

From numerator above,

$r = \sum (x - x)(y - y)$

$r = \sum (xy - xy - yx + xy)$

$r = \sum (xy - \frac{\sum y}{\cap} x - \frac{\sum xy}{\cap} + \frac{\sum x}{\cap} . \frac{\sum y}{\cap}$

$r = \sum xy - \frac{\sum y}{\cap} \sum x - \frac{\sum x}{\cap} . \sum y + \cap . \frac{\sum x}{\cap} . \frac{\sum y}{\cap}$

$= \sum xy - \frac{\sum x \sum y}{\cap}$

$= \frac{\cap \sum xy - \sum x . \sum y}{\cap}$ ----------(i)

From renominator,

$\sum (x - x)(x - x)$

$\sum (x^2 - x\,x - xx\ f\ xx)$

$\sum (x^2 - \frac{\sum x}{\cap} x - \frac{\sum y}{\cap} x + \frac{\sum x}{\cap} . \frac{\sum x}{\cap}$

$\sum x^2 - \frac{(\sum xy)^2}{\cap} - \frac{(\sum x)}{\cap} + \frac{(\sum x)^2}{\cap}$

$= \sum x^2 - (\sum X)^2$

$= \sum x^2 - (\sum x)^2$ ----------(ii)

Mathematically,

$\sum (y - y)(y - y)$

$\sum (y^2 - y\,y - yy\ f\ yy)$

$$\sum (y^2 - \underline{\sum y}\ \ y - \underline{\sum y}\ x + \underline{\sum y}\ \ . \underline{\sum y}$$
$$\cap \qquad \cap \qquad \cap \qquad \cap$$

$$\sum y^2 - \underline{(\sum y)^2} - \underline{(\sum y)} + \underline{(\sum y)^2}$$
$$\cap \qquad \cap \qquad \cap$$

$$= \quad \sum y^2 - \underline{(\sum y)^2}$$
$$\cap$$

$$= \sum x^2 - (\sum x)^2 \text{ ----------(ii)}$$

Thus, equate (i) and (ii)

$$= \sum xy^2 - \sum y \sum x \div \sqrt{\underline{\sum x^2 - (\sum x)^2}}\ . \sqrt{\underline{\sum y^2 - (\sum y)^2}}$$
$$\cap \qquad \cap$$

$$\sum xy - \sum y \sum x \div \sqrt{\underline{\sum x^2 - (\sum x)^2}}\ . \sqrt{\underline{\sum y^2 - (\sum y)^2}}$$
$$\cap \qquad \cap$$

$$= \sum xy - \sum y \sum x\ \ x\ \ \sqrt{\underline{\sum x^2 - (\sum x)^2}}\ \ . \sqrt{\underline{\sum y^2 - (\sum y)^2}}$$
$$\cap \qquad \cap$$

$$r = \underline{\sum xy - \sum y \sum x}$$
$$\sqrt{\underline{\sum x^2 - (\sum x)^2}}\ \ . \underline{\sum y^2 - (\sum y)^2}$$

Or

$$r = \underline{\quad \sum xy}$$
$$\sum x^2 . \sum y^2$$

Remarks:

The value of r can be expressed in 3 ways of interpretation of relationship between x and y.

i. When r = +1, i.e. perfect (positive) linear relationship

ii. When r = -1 i.e. perfect (negative) linear relationship

iii. When r = 0 i.e. no relationship.

Note: The straight of relationship between x and y depends on how close r is to zero. And the coefficient of determination will be given as (r^2).

Illustration: Relationship between money spent on research and development and chemical firm's annual report profit. The information for proceeding 6 years was as recorded. Calculate product moment correlation coefficient.

Years	1994	1995	1992	1991	1990	1989	
Money (N) res and Dev.	5	11	4	5	3	2	x
Annual profit (N)	31	40	30	34	25	20	y

Data

$\mu = 6$; $\sum x = 30$; $\sum y = 180$; $\sum xy = 1000$; $\sum x^2 = 250$; $\sum y^2 = 5642$

Yrs	X	Y	xy	X^2	Y^2
1994	5	31	155	25	961
1993	11	40	440	121	1605
1992	4	30	120	16	900
1991	5	34	170	25	1156
1990	3	25	75	9	625
1989	2	20	40	4	400

$$r^1 = \frac{\sum xy - (\sum x)(\sum y)}{\sqrt{(\sum x^2 - (\sum x)^2)(\sum y^2 - (\sum y^2)}}$$

$$r^1 = \frac{6(1000) - (30)(180)}{\sqrt{(6(200) - (30)^2 - (6(5642) - (180)^2}}$$

$$r^1 = \frac{6000) - 5400}{\sqrt{(1200 - (900) - (33852 - 32400)}}$$

$$r^1 = \frac{600}{\sqrt{(300)\ (1452)}}$$

$$r^1 = \frac{600}{\sqrt{435600}}$$

$$r^1 = \frac{600}{\sqrt{660}} = 0.9091$$

Remarks: They are highly perfect / related.

Illustration: Lasu Campus stores has been selling the believe it or not. Wonders of statistics study guide for 12 Semester and would like to estimates the relationship between sales and number of sections of elementary statistics taught in each Semester. The data below have been collected.

Sales (units)	33	35	24	61	52	45	65	82	29	63	50	79
No of sections	3	7	6	6	10	12	12	13	12	13	14	15

a. Obtain the coefficient of correlation
b. Comment on your result?

sales (x)	No. of Section (y)	Xy	x^2	y^2
33	3	99	1089	9
38	7	226	1444	49
24	6	144	576	36
61	6	366	3721	36
52	10	520	2704	100
45	12	540	2025	144
65	12	780	4225	144
82	13	1066	6724	169
29	12	348	841	144

63	13	819	3969	169
50	14	700	2500	196
79	15	1185	6241	225

Data

$\sum x = 621$

$\sum y = 123$

$\sum xy = 6833$

$\sum x^2 = 385641$

$\sum y^2 = 15129$

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

$$r = \frac{12\,(6833) - (621)(123)}{\sqrt{(12(385641) - (621)^2 - (12\,(151291) - (123)^2}}$$

$$r = \frac{81996 - 76383}{\sqrt{(432706 - 385641) - (17052 - 15129)}}$$

$$r = \frac{5613}{\sqrt{(47067\,(1923)}}$$

$$r = \frac{5613}{\sqrt{9513.7}} = 0.59$$

Remarks: The sales units and the number of section are particularly correlates or related. The relationship is weak. The unit sales may not necessary be determined or depend on the number of sections.

Illustration: Find the correlation coefficient between the following series. Calculate the correlation of beer consumption as regards the accident in our high ways between 2001 – 2010.

Hence, calculate the dependent variables between bear consumption and road accident.

Year	Road accident	Beer consumption
2001	155	70
2002	150	63
2003	180	72
2004	135	60
2005	156	66
2006	165	70
2007	178	74
2008	160	65
2009	132	62
2010	145	67

Year	Beer consumption (x)	Road accident (y)	xy	x^2	y^2
2001	70	155	10850	4900	24025
2002	63	150	9450	3969	22560
2003	72	180	12960	5084	32400
2004	60	135	5100	3600	18225
2005	66	156	10296	4356	24336
2006	70	165	11760	4900	28224
2007	74	178	13172	5476	31684
2008	65	160	10400	4225	25600
2009	62	132	8184	3844	17424
2010	67	145	9715	4489	21025

Data

V = 10

$\sum x = 669$

$\sum y = 1559$

$\sum xy = 10,4887$

$\sum x^2 = 44943$

$\sum y^2 = 245443$

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

$$r = \frac{10\,(104887) - (660)\,(15590}{\sqrt{(10(44943) - (669)^2 - (10\,(245443) - (1559)^2}}$$

$$r = \frac{1048870 - 1042971}{\sqrt{(449430 - 447651)\,(2454430 - 2436481)}}$$

$$r = \frac{5899}{\sqrt{(779)\,(23949)}}$$

$$r = \frac{5899}{\sqrt{42605271}}$$

$$r = \frac{5899}{6227.27} = 0.9037$$

= 0.9037

But r = 0.8169 = 0.82

The coefficient determination shows the variation in the independent variable (y) as a result of corresponding variation in the explanatory variables (x).

This shows that 90% of beer consumption belong to road accident and of is thus RA = F (BO). The interpretation of coefficient correlation means that 0.82% road accident is brought about 90% of the beer consumption.

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter ⟩ referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

6.0 Tutor-Marked Assignment

Determine the correlation between X and Y in the table below.

Period	1	2	3	4	5	6	7	8	9	10
Qty Supply	10	20	50	40	50	60	80	90	90	120
Unit Price (N)	2	4	6	8	10	12	14	16	18	20

7.0 References /Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT THREE: SPEARMAN'S RANK CORRELATION

RANK CORRELATION OR TIED IN RANK CORRELATION

Contents

1.0 Introduction
2.0 Objectives
3.0 Main content
3.1 Analysis of Rank Correlation
4.0 Summary and Conclusion
5.0 Tutor-Marked Assignment
6.0 Further Reading
7.0 References

1.0 INTRODUCTION

It is found very difficult to quantity a data or set of data that has big values. Rank correlation is used to determine the extent at which the variable are correlated. This idea was employed by Spearman's rank correlation coefficient, which is computed by using this formula.

$$r = \frac{1 - 6\sum d^2}{(x^2 - 1)}$$

2.0 OBJECTIVE

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

- explain the computation of rank correlation coefficients
- apply the concept of correlations in business decisions.

3.0 MAIN CONTENT

Where,

O = number of observation

d = difference between the pairs of rank (x & y) values

r = rank correlation

note: In a cases where there tied or tied in ranking of variables x and y, other representation is applicable.

$$r^1 = \frac{1 - 6 (\sum d^2 + t^3 - t)}{(\pi + 1)(\pi - 1)}$$

Where, t = number of ties in variable x & x respectively, but when coefficient of no determination occur, we equate $1 - r^2$

Illustration: The following data refer to the students scores. The general level of their intelligent in 9 selected courses. Using Spearman's correlated techniques to determine the straight of the relationship between the students cadres and their intelligent.

Sales (units)	y	16	14	15	13	31	16	10	17	20
Intelligent	x	38	41	48	22	64	64	26	53	30

Y	X	rx	Ry	d = (rx – ry)	d^2
16	38	6	4.5	1.5	2.25
14	41	5	7	-2	4
15	48	4	6.	-2	4
13	22	9	8	1	1
31	64	1.5	1	0.5	0.25
16	64	1.5	4.5	-3.0	9
10	26	8	9	-1	1

17	53	3	3	0	0
20	30	7	2	5	25

Data

$X = 9$

$\sum F^2 = 46.5$

$$r = \frac{1 - 6\sum d^2}{x\,(x^2 - 1)}$$

$$1 - \frac{6\,(46.5)}{9\,(9^2 - 1)}$$

$$1 - \frac{6\,(46.5)}{9\,(9^2 - 1)}$$

$$1 - \frac{279}{720} = 1 - 0.3875 = 0.6125 = 0.61$$

Illustrate: A market research asked two (2) smoker to express their difference for 12 difference brands of cigarettes. The reply as shown in the following table.

Brand of cigarette	A	B	C	D	E	F	G	H	I	J	K	L
Smoker z (v)	9	10	4	1	8	11	3	2	5	7	12	6
Smoker W (x)	7	8	3	2	10	12	1	6	5	4	11	9

Requirement: Use Spearman's rank correlation technique to evaluate the straight of relationship between the smokers.

Y	X	rx	Ry	d	d^2
9	7	6	4	2	4
10	8	5	3	2	4
4	3	10	9	1	1
1	2	11	12	-1	1
8	10	3	5	-2	4
11	12	1	2	-1	1
3	1	12	10	2	4
2	6	7	11	-4	16
5	5	8	8	0	0
7	4	9	6	3	9
12	11	2	1	1	1
6	9	4	7	-3	9

Data

$\cap = 12$

$\sum d^2 = 54$

$r1 = \underline{1 \cap \sum d^2}$

$n(n^2 - 1)$

$= 1 - \underline{12\ (54)} \qquad 1 - \underline{324}$

$12\ (12^2 - 1) \qquad 1716$

$= 1 - 0.1888$

$r = \quad 0.89$

Illustration: Assuming that 10 men assign to a particular job or task were given two aptitude test. After they have been on the job for some period of time. The production manager was ask to rank the employees from 1^{st} to 10^{th} in regard to their value to the company. You, as the particular manager, should use the Spearman's technique to determine the relationship between the 2 test.

Workers	A	B	C	D	E	F	G	H	I	J
Test 1	96	98	79	78	84	84	76	79	62	44
Test 2	78	72	60	72	64	84	72	56	78	40

Y	X	rx	ry	d	d^2
96	78	2	2.5	-0.15	0.25
98	72	1	4.5	-8.5	12.25
79	60	5.5	7	-2.5	6.25
78	72	7	4.5	2.5	6.25
84	64	3.5	7	-3.5	12.25
84	84	3.5	1	2.5	16.25
76	72	8	4.5	3.5	12.25
79	56	5.5	9	-3.5	12.25
62	78	9	2.5	6.5	42.25
44	40	10	10	0	0

Data

$\cap = 10$

$\sum d^2 = 110.25$

$r1 = \underline{1\ 6\sum d^2}$

$n(n^2 - 1)$

$= 1 - \underline{6 \times 110.25} \qquad 1 - \underline{661.5}$

$10\ (10^2 - 1) \qquad 990$

$= 1 - 0.6682$

$r = \quad 0.3318$

Illustration: The debits in international business transactions (current transfer in million) of United Kingdom from personal sector (x) and central government (y) for the quarters in the period of 1970 to 1972 is given as below:

X	56	57	55	58	51	56	56	58	57	57	57	57
Y	52	40	37	43	57	45	47	51	68	49	43	48

a. Rank in data
b. Compute Spearman's coefficient of rank correlation

X	Y	rx	Ry	d	d^2
56	52	9	3	6	36
57	40	5	11	6	36
55	37	11	12	1	1
58	43	1.5	9.5	8	64
51	57	1.2	2	10	100
56	45	9	8	1	1
56	47	9	7	2	4
58	51	1.5	4	2.5	6.25
57	68	5	1	4	16
57	49	5	5	0	0
57	43	5	9.5	4.5	20.25
57	48	5	6	1	1

Data

$\cap = 12$

$\sum d^2 = 285.5$

$r1 = \underline{1\ 6\sum d^2}$

$n(n^2 - 1)$

$= 1 - \underline{6 \times 285.5} \qquad 1 - \underline{1713}$

$$12\ (12^2 - 1) \qquad 1716$$

$= 1 - 0.9983$

$r = \quad 0.0001$

Comment: The value of r1 shows that x and y are not correlated i.e. they are not in agreement

3.1 ANALYSIS OF RANKED DATA

Spearman's coefficient of correlation assumes the data to be at least interval scale. Chalse – Spearman, a British statistician, introduced a measure of correlation for ordinal- level data known as Spearman's rank-order correlation coefficient (i.e. A measure of relationship between two sets of ranked data). Its designated as r_s, may range between -1.0 and +1.0 inclusive with -1.0 and +1.0 representing perfect rank correlation. Zero indicates no rank correlation.

The general representation can be given as

$$r_s = 1 - \frac{6\ (\sum d2)}{n\ (n^2 - 1)}$$

where,

n = number of paired observations

d = difference between the ranks for each pair.

NB: for large-sample where n is 10 or more, the student's "t" distribution can be used as the test of statistic. And the degree of freedom is given as (n -2)

The general computed formular is given as

$$t = r_s \quad \frac{\pi - 2}{1 - r_s^2}$$

Example: A sample of 12 auto mechanics was ranked by the supervisor regarding their mechanical ability and their social compatibility. The results are as follows:

Worker	Mechanical Ability	Social compatibility
1	1	4
2	2	3
3	3	2
4	4	6
5	5	1
6	6	5
7	7	8
8	8	12
9	9	11
10	10	9
11	11	7
12	12	10

Compute the coefficient of rank correlation can we conclude that there is a positive association in the population between the ranks of mechanical ability and social compatibility?

Use the 0.05 significance level.

Worker	Mechanical Ability	Social compatibility	d	d^2
1	1	4	-3	9
2	2	3	-1	1
3	3	2	1	1
4	4	6	-2	4
5	5	1	4	16
6	6	5	1	1

7	7	8	-1	1
8	8	12	-4	16
9	9	11	-2	4
10	10	9	1	1
11	11	7	4	16
12	12	10	2	4

$\sum d^2 = d\ 4$

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

$1 - \frac{6(74)}{12(12-1)} = 1 = 0.741$

Decision: The value 0.741 indicate fairly strong positive association between the ranks of mechanical ability and social compatibility.

$\alpha = 0.05$

H_0: The rank correlation in the population is zero

H_1: the rank correlation in the population is greater than zero.

Using one tailed test.

d.f = n -2 = 12 – 2 = 10

To obtain “t” test

$t = rs\sqrt{\frac{n-2}{1-r^2_s}}$

$= 0.741\ \sqrt{\frac{12-2}{1-(0.74)^2}} \qquad 0.741\ \sqrt{\frac{10}{1-(0.74)^2}}$

$0.741\ (22.177) \qquad = 0.0741\ \sqrt{(4.7092)}$

$= 3.4895 = 3.49$

Decision: Since computer value exceed critical value of 1.812, then H_0 is rejected, and H_1 is accepted. It is concluded that there is a positive association between the ranks of social compatibility and mechanical ability among auto mechanics.

4.0 Summary and Conclusion

5.0 Tutor-Marked Assignment

1. Twelve persons whose IQs were measured in cottage between 1960 and 1965 were located recently and retested with an equivalent IQ test. The information is given below.

Student	Recent score	Original score
John Barr	119	112
Bill Sedwick	103	108
Morica Elephant	115	115
Ginge Tale	109	100
Larry Clark	131	120
Jim Redding	110	108
Carol Papalia	109	113
Victor Soppa	113	126
Dallae Paul	94	95
Carol Kozoloski	119	110
Jok Sass	118	117
P.S Sundar	112	102

At the 0.05 significance level can we conclude that the IQ scores have increased in over 20 years. Compute the coefficient of rank correlation.

6.0 Further Reading

NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

7.0 References/ Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 4: ORDINARY LEAST SQUARE ESTIMATION (REGRESSION ANALYSIS)

Contents

1.0 Introduction

4.0 Objectives

5.0 Main content

6.0 Summary and Conclusion

7.0 Tutor-Marked Assignment

6.0 Further Reading

7.0 References

1.0 INTRODUCTION

Regression analysis can be defined as the relationship between two or more variables. This relationship has to do with the changes that result from a change in one of the related variables.

2.0 OBJECTIVE

The main objective of this unit is to enable students understand the theory behind and the application of regression analysis in statistics. Students are expected at the end of this unit to be able to apply regression analysis to solving day-to-day business and economic problems.

3.0 MAIN CONTENT

Uses and the Types of Regression

i. It is used for prediction
ii. It is used for description of relationship
iii. To improve on knowledge of variable of interest

Basically, there are two types:

i. Simple (linear) regression
ii. Multiple (non linear) regression

Simple (linear) regression

This involve only two variables and the relationship between them tends towards a fixed direction.

Multiple (non linear) regression

This also involved more than two variables in the regression model or equation.

Mathematically, let us assume that x_1 and x_2 as independent variable (factor) and y as dependent variable. The independent variable may be more than two, i.e. rt can be obtainable as x_1, x_2, x_3….x

Recall, $y = a + bx$ = simple regression

Similarly, we can have $y = a + b_1x_1 + b_2x_2$ (for multiple regression)

Method of Calculating Regression Line

Regression line of any form can be titled to a bivariate data by any of the following methods.

1. Freehand method

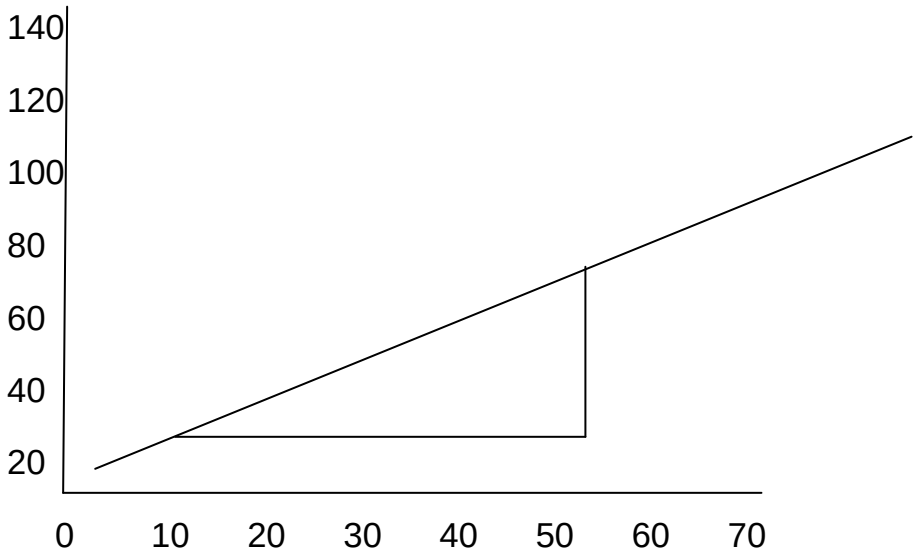
In this method, regression line is fitted into the scatter diagram

This scatter diagram is the graphically representation of relationship which exists between two variables by drawing a lines of best fit through the various points which are estimate from the relationship x and y.

Illustration: Given/estimate the regression equation by using the scatter diagram from the data below. The marks scored by a group of philosophy students and mathematics students are as follows.

Philosophy marks	38	51	19	53	39	38	66
Mathematics marks	50	32	36	54	52	56	80

By scatter diagram:



Limitation:

i. It does not give a unique regression line
ii. It also does not give unique regression coefficient

2. Least Square Method
This is the mathematical method of determined the points estimate of 'a' and 'b' from the available sample points. This method is the most reliable of all the methods. A gives a unique regression line and a unique regression coefficient. The method of least square provides two set of equation called (Normal Equations) which can solved simultaneously for two unknown.

By representation,
Y = a+bx; b = the coefficient of x and x = independent variable.

From the line of fit from y = a + bx
$\sum y = an + b\sum x$----------i
$\sum xy = a\sum x + b\sum x^2$ -----ii

Both equation are regretted a normal equation.
$ax + b\sum x$
$a\sum x + b\sum x^2$

From the equation above,

$ax + b\sum x - \sum y$

$a \sum x + b\sum x^2 = \sum xy$

by determinant method

$$\begin{Bmatrix} x & \sum x \\ \sum x & \sum x^2 \end{Bmatrix} \quad \begin{Bmatrix} a \\ b \end{Bmatrix} = \begin{Bmatrix} \sum y \\ \sum xy \end{Bmatrix}$$

$$\begin{Bmatrix} x & \sum x \\ \sum x & \sum x^2 \end{Bmatrix}$$

To obtain

$$\begin{Bmatrix} x & \sum x \\ \sum x & \sum x^2 \end{Bmatrix} \quad = N \quad \sum x^2 - \sum x \, . \, \sum x$$
$$= \sum x^2 - \sum x^2$$

$= \sum x^2 – (\sum x)^2$

$(A) = \sum x^2 – (\sum x)^2$ -------3

From (A_1) $\begin{Bmatrix} x & \sum x \\ \sum xy & \sum x^2 \end{Bmatrix}$ $= (A1) = (\sum x^2 \, . \, \sum y - \sum x \, . \, \sum xy)$

$(A^1) = \sum x^2 \, . \, \sum y - \sum x \, . \, \sum xy$ -------4

From other Equation

(A_2) $\begin{Bmatrix} x & \sum x \\ \sum x & \sum xy \end{Bmatrix}$ $= \sum xy - \sum x \, . \, \sum y$ ----------5

Mathematically, (By determinant)

$(A1) = \underline{\Delta A^1} \, . \, \underline{\Delta A^2}$

$\Delta 1 \qquad \Delta$

$$= \frac{\sum x^2 . \sum y - \sum x . \sum xy}{\sum x^2 - (\sum x)^2} \text{.....a}$$

$$(A2) = \frac{\sum x^2 . \sum y - \sum x . \sum xy}{\sum x^2 - (\sum x)^2} \text{----------b}$$

Both equation should be memorized.

Non-Linear Model

On most occasion, the simple linear model and in particular the multiple linear model will not be satisfactory. A plot or scatter diagram on the dominant variable may suggest that the relationship is not linear. We consider non-linear model, which involves:

i. Different type of curve
ii. Linearization

Types of Curve

There are 3 main types of curve

1. Exponential (Growth) curve: This is a situation whereby when a data is expected to grow by some proportion or percentage in each period.
   An exponential curve have:
   $Y = ab^{ru}$ and in particular or
   $Y = ac^{ru}$ or ab^{u}
   When abr are constant
   Where: y = variable to be predicted
   Axb = constant
   X = number of period
   Now, to linearise the above equation
   $Y = ab^{u}$
   Obtain log of both sides
   $\text{Log } y = \log A + \log b^{x}$
   $\text{Log } y = \log A + x \log B$
   Equate $\log_u$ to both sizes
   Log y = A + Bx ----------x

2. Hyperbolic Model (curve): This has a formular

$Y = a + b/x$ or $y = 1/a+bx$

To linearise y, take $x = 1/x$, then we have

$Y = a + bx$

$i/y = a + bx$

since $y = 1/y$

therefore, $y = a + bx$

3. Power curve mor model: This power model have the form of $y = axb$. Otherwise known as logarithms functions. The general representation can be given as:
   $y = ax^b$
   to linearise: obtain $\log_{10}$ to both sides
   $\log_{10} y = \log_{10} a + b\log_{10}x$
   since a and b are constant.
   $\text{Log}_{10}y = y^1$ and $\log_{10} a = A$
   Therefore, $y^1 = A + bx$

   Illustration: Draw the scatter diagram and fit an exponential curbe in the following data

Years	1983	1984	1985	1986	1987
Sales	100	150	225	337.5	506.25

X years	Y sales
O	100
1	150
2	225
3	337.5
	506.25

X	Y	Log y	Xlogy	X^2
0	100	2.000	0	0
1	150	2.1761	2.1761	1
2	225	2.3522	4.7044	4

3	337.5	2.5282	7.5846	9
4	506.25	2.7045	10.8180	16

Data

$\sum x = 10$

$\sum y = 11.7610$

$\sum xy = 25.2831$

$\sum x^2 = 30$

From general representation

$y = abx$

$\log y = \log a + x \log b$

to find a and b

firstly, to find b = ?

$$b = \frac{n\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{5(25.283) - (10)(11.7610)}{5(30) - 100}$$

$$b = \frac{8.8056}{50} = 0.17611$$

Then, $\log b = 0.17611$

$b^{-1} (0.17611) = 1.5$

to obtain a = ?

$A = y - bx$

$A = y - bx$

$$A = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$a = \frac{12.6532}{5} - \frac{(2)(2.0792)}{5}$

$A = 1.6989$

$A^{-1}(1.6989) = 50$

Therefore, $y = ax^b$

$Y = 50(x^2)$

But when x = 1, 2, 3, 4, -----

Therefore, $y = 50(1^2) = 5$

$Y = 50(2^2) = 200$

$Y = 50(3^2) = 450$

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter ⟩ referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

6.0 Tutor-Marked Assignment

1. Illustration: Given/estimate the regression equation by using the scatter diagram from the data below. The marks scored by a group of philosophy students and mathematics students are as follows.

Philosophy marks	3	5	9	5	9	8	6
Mathematics marks	5	2	3	4	5	6	8

2.
3. Illustration: Draw the scatter diagram and fit an exponential curbe in the following data

Years	1983	1984	1985	1986	1987
Sales	10	15	25	33.75	50.625

7.0 References/Further Reading

ONWE J.O. NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 5: MULTIPLE REGRESSION ANALYSIS

Content

1.0 Introduction

2.0 Objectives

3.0 Main content

4.0 Summary and Conclusion

5.0 Tutor-Marked Assignment

6.0 Further Reading

7.0 References

1.0 INTRODUCTION

Recall, the degree of relationship that connect three or more variables together are called multiple correlation regression.

e.g. $y = a+b_1x_1+b_2x_2+b_3x_3\text{------}b_x\ x_n$

2.0 OBJECTIVE

The objective of this unit is to introduce students to multiple regression analysis and emphasize its applications in statistics.

3.0 MAIN CONTENT

The above expression can be solved by the normal equation of the three variables.

$y = a+b_1x_1+b_2x_2+b_3x_3\text{------}b_x\ x_n$

$\sum y = ax + b_1\sum x_1+b_2\sum x_2\text{----------(i)}$

$\sum x_1y = a\sum x_1+b_1\sum x_1^2+b_2\sum x_1x_2\text{----------(ii)}$

$\sum x_2y = a\sum x_2+b_1\sum x_1^2,x_2+b_2\sum x^2{}_2\text{--------(iii)}$

But the coefficient of multiple determination r^2 can be expressed as:

$$r^2 = \frac{a\sum y+b_1\sum x_1y+b_2\sum x_2y - (\sum y)^2}{\sum y^2 - (\sum y/x)^2}$$

Illustration: The Faculty of Management Science (HMS) has investigating the relationship between some students performance in their various courses and lecture received per each semester and also the quality of some lecturers. The faculty has a data of ten candidates which are:

Student	1	2	3	4	5	6	7	8	9	10
No of lecturer	9	6	12	14	11	6	19	16	3	9
Quality of lecturers	99	100	119	95	110	117	98	101	100	115
Exams scores	56	45	80	73	71	55	95	86	34	66

From general representation

$y = a+b_1x_1+b_2x_2+b_3x_3$------$b_x\ x_n$

Students	y	y^2	x_1	x_1^2	x_2	x^2_2	x_1y	x_2y	x_1x_2
1	56	3136	9	81	99	9801	5044	5544	891
2	45	2025	6	36	100	10,000	270	4500	600
3	50	6400	12	144	119	14161	960	9520	1428
4	73	5329	14	196	95	9025	1022	6935	1330
5	71	5041	11	121	110	12100	781	7810	1210
6	55	3025	6	36	117	13689	330	6435	702
7	95	9025	19	361	98	9604	1805	9310	1862
8	86	7396	16	256	101	10201	1376	8656	1616
9	34	1156	3	9	100	10000	102	3400	300
10	66	4356	9	81	115	13225	594	7590	1035

Data:

$\sum y = 661$

$\sum y^2 = 46889$

$\sum x_1 = 105$

$\sum x_1^2 = 1321$

$\sum x2 = 1054$

$\sum x_2^2 = 111,806$

$\sum x_1 y = 7744$

$\sum x_2 y = 69730$

$\sum x_1 x_2 = 10,974$

To find b = p

$$bx_1 = \frac{n \sum x_1 y - \sum x_1 y}{n \sum x_1^2 - (\sum x_1)^2}$$

$$= \frac{10\,(7744) - (105)\,(661)}{10\,(1321) - (105)^2}$$

$$= \frac{77440 - 69405}{13210 - 11025} = \frac{8035}{2185} = 3.6773 = 3.68$$

Therefore: bx = 3.68

To find a = ?

$$ax_1 = y - \frac{bx1 \sum x1}{n}$$

$$ax1 = \frac{\sum y}{n} - \frac{bx_1 \sum x_1}{n}$$

$$= \frac{661}{10} - \frac{(3.68)\,(105)}{10}$$

$= 66.1 - 38.64 = 27.46$

Therefore, $ax_1 = 27.64$

But, $y = ax_1+bx_1x_1$

$y =27.64+3.68$

$y = 27.64+ 3.68x_1$

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.
A **partial correlation coefficient** measures the relationship between any two variables, keeping other variables constant.
The limitations of linear correlations as a technique for the study of economic relations are as follows

The formula for correlation coefficient applies only to linear relationships between variables.

That correlation coefficient as a measure of co-variability of variables does not imply any functional relationship between the variables concerned.

6.0 Tutor Mark Assignment

Illustration: calculate the coefficient of linear multiple regression of the data below: the association of accountants is investigating the relationship between performance in Quantitative methods and how studied per week and the general level of intelligence of candidates. The Association has data on ten students which are:

Students	1	2	3	4	5	6	7	8	9	10
Hours studied x_1	9	6	12	14	11	6	19	16	3	9
T.Q (x_2)	99	100	119	95	110	117	98	101	100	115

Hence, predict the expected score of a candidate.

7.0 References/Further Reading

NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques.

MODULE FOUR: STATISTICAL TEST

UNIT 1: HYPOTHESIS AND T–TEST

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

- 3.1 Application of Hypothesis and *t*-distribution
- 3.2 Test for single mean
- 3.3 Assumptions for Student's test
- 3.4 t-Test for difference of means

4.0 Conclusion

5.0 Summary

6.0 Assignment

7.0 References / Further Reading

1.0 INTRODUCTION

A hypothesis can be defined as a conjectural statement a postulate, or a proposition about an assumed relationship between two or more variables.

Hypothesis testing or testing a hypothesis are used interchangeably. Hypothesis testing starts with a statement about population parameters such as mean. But, in an attempt to reach a decision, statistician often make an assumption or proposition about the population involve. Such assumption which is subject to testing either may be true or may not be true is called statistical hypothesis.

The Student's t-test

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

If the population variance is unknown then for the large samples, its estimates provided by sample variance S^2 is used and normal test is applied. For small samples an unbiased estimate of population variance σ^2 is given by:

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

It is quite conventional to replace σ^2 byS^2 (for small samples) and then apply the normal test even for small samples. W.S Goset, who wrote under the pen name of Student, obtained the sampling distribution of the statistic $\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$ for small samples and showed that it is far from normality. This discovery started a new field, viz '*Exact Sample Test*' in the history of statistical inference.

Note: If x_1, x_2.....x_n is a random sample of size *n* from a normal population with mean μ and variance σ^2 then the Student's *t* statistic is defined as:

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

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

2.0 OBJECTIVES

The objective of this unit is to introduce students to *t*-distribution and emphasize its application in statistics.

3.0 MAIN CONTENT

3.1 Applications of *t*-distribution

(i) *t*-test for the significance of single mean, population variance being unknown

(ii) *t*-test for the significance of the difference between two sample means, the population variances being equal but unknown

(iii) *t*-test for the significance of an observed sample correlation coefficient

3.2 Test for Single Mean

Sometimes, we may be interested in testing if:

(i) The given normal population has a specified value of the population mean, say μ_o.

(ii) Thesample mean $\bar{x}$ differ significantly from specified value of population mean.

(iii) A given random sample x_1, x_2.....$x_n$of size n has been drawn from a normal population with specified meanμ_o.

Basically, all the three problems are the same. We set up the corresponding null hypothesis thus:

(a) H_o: $\mu = \mu_o$ i.e the population mean is μ_o

(b) H_o: There is no significant difference between the sample mean and the population mean. In order words, the difference between $\bar{x}$ and μ is due to fluctuations of sampling.

(c) H_o: The given random sample has been drawn from the normal population with mean μ_o. Under H_o the test-statistic is:

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

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

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

We compute the test-statistic using the formula above under H_o and compare it with the tabulated value of *t* for (n-1) *d.f.*at the given level of significance. If the absolute value of the calculated *t* is greater than tabulated *t*, we say it is significant and the null hypothesis is rejected. But if the calculated *t* is less than tabulated *t*, H_o may be accepted at the level of significance adopted.

3.3 Assumptions for Student's test

(i) The parent population from which the sample is drawn is normal

(ii) The sample observations are independent i.e. the given sample is random.

(iii) The population standard deviation σ is unknown

Example: Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 11.8kg and standard deviation is 0.15kg. Does the sample mean differ significantly from the intended weight of 12kg, α=0.05

Hint: You are given that for *d.f.* =9, $t_{0.05} = 2.26$

Solution: n= 10, $\bar{x}$= 11.8kg, s = 0.15kg

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

*Alternative Hypothesis.*Ho: ≠ 12kg (Two tailed)

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

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

The tabulated value of t for 9 d.f. at 5% level of significance is 2.26. Since the calculated t is much greater than the tabulated t, it is highly significant. Hence, null hypothesis is rejected at 5% level of significance and we conclude that the sample mean differ significantly.

3.4 t-Test for difference of means

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

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

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

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

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

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

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

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

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

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

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

4.0 CONCLUSION

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

5.0 SUMMARY

Here, you would have learnt how to apply t-test in solving statistical problems such as test to confirm if mean is a certain value, to test significance of the difference between two mean among others.

6.0 TUTOR-MARKED ASSIGNMENT

1.The mean weekly sale of the chocolate bar in candy stores was 146.3 bars per store. After advertising campaign the mean weekly sales in 22 stores for typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?

2.Prices of shares of a company on the different days in a month were found to be: 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68. Discuss whether the mean price of the price of the shares in the month is 65.

3.Two salesmen A and B are working in certain district. From a Sample Survey Conducted by the Head Office the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

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

7.0 REFERENCES / FURTHER READING

OKOJIE, DANIEL E. NOUN TEXT BOOK, Eco 203: Statistics for Economists

OTOKOTI O.S. Contemporary Statistics

UNIT 2:

F-TEST

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Applications of the F-distribution

3.2 For testing equality of population variances

4.0 Conclusion

5.0 Summary

6.0 Assignment

7.0 References/Further Reading

1.0 INTRODUCTION

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

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

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

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

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

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

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

$$P\left[F > F\alpha\left(v_1, v_2\right)\right] = \alpha$$

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

2.0 OBJECTIVE

The main objective of this section is to introduce student to the world of F-distribution and learn its theories and application to day-to-day business and economic problems.

3.0 MAIN CONTENT

3.1 Applications of the F-distribution

F-distribution has a number of applications in the field of statistics. This includes but not limited to the following:

(1) To test for equality of population variances

(2) To the equality of several population means i.e. for testing Ho: $\mu_1 = \mu_2 = ... = \mu_k$. This is by far the most important application of F-statistic and is done through the technique of Analysis of Variance (ANOVA). This shall be treated as a separate unit later.

(3) For testing the significance of an observed sample multiple correlation

(4) For testing the significance of an observed sample correlation ratio

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

Under Ho, the test statistic is

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

Where, S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 and are given by:

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

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

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

Assumption for F-test for equality of variances

1. The samples are simple random samples
2. The samples are independent of each other

3. The parent populations from which the samples are drawn are normal

N.B (1) since, the most available tables of the significant values of F are for the right-tail test, i.e. against the alternative H_o: $\sigma_1^2 > \sigma_2^2$, in numerical problems we will tale greater of the variances S_1^2 or S_2^2 as the numerator and adjust for the degree of freedom accordingly. Thus, in $F \sim (v_1, v_2)$, v_1 refers to the degree of freedom of the larger variance, which must be taken as the numerator while computing *F*.

If Ho is true i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ the value of F should be around 1, otherwise, it should be greater than 1. If the value of *F* is far greater than 1 the Ho should be rejected. Finally, if we take larger of S_1^2 or S_2^2 as the numerator, all the tests based on the F-statistic become right tailed tests.

- All one tailed tests for Ho at level of significance "α" will be right tailed tests only with area "α" in the right.
- For two-tailed tests, the critical valuesare located in the right tail of F-distribution with area ($\alpha/_2$) in the right tail.

Example 1: The time taken (in minutes) by drivers to drive from Town A to Town B driving two different types of cars X and Y is given below

Car Type X: 20 16 26 27 23 22

Car Type Y: 27 33 42 35 32 34 38

Do the data show that the variances of time distribution from population from which the samples are drawn do not differ significantly?

Solution:

X	***d = x – 22***	***d^2***	***Y***	***d = y -35***	***D^2***
20	-2	4	27	-8	64
16	-6	36	33	-2	4
26	4	16	42	7	49

25	5	9	35	0	0
23	1	1	32	-3	9
22	0	0	34	-1	1
			38	3	9
Total	**2**	**d^2 = 82**		**-4**	**ΣD^2 =136**

$$S_1^2 = \frac{1}{n_1-1}\sum(x - \bar{x})^2 = \frac{1}{n_1-1}\left[\sum d^2 - \frac{(\sum d)^2}{n_1}\right]$$

$$= \frac{1}{5}\left[82 - \frac{4}{6}\right] = \frac{1}{5}[82 - 0.67] = 16.266$$

$$S_2^2 = \frac{1}{n_2-1}\sum(y - \bar{y})^2 = \frac{1}{n_2-1}\left[\sum D^2 - \frac{(\sum d)^2}{n_1}\right]$$

$$= \frac{1}{6}\left[136 - \frac{16}{7}\right] = \frac{1}{6}[136 - 2.286] = 22.286$$

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

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

$$F = \frac{22.286}{16.266} = 1.37$$

Tabulated $F_{0.05(6,5)} = 4.95$

Since the calculated F is less than tabulated F, it is not significant. Hence Ho may be accepted at 5% level of significance or risk level. We may therefore conclude that variability of the time distribution in the two populations is same.

4.0 CONCLUSION

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

5.0 SUMMARY

Students have learnt the theories and application of the F-test

6.0 TUTOR-MARKED ASSIGNMENT

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

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

7.0 REFERENCE/FURTHER READING

OKOJIE, DANIEL E. NOUN TEXT BOOK, Eco 203: Statistics for Economists

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

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

UNIT 3:

CHI-SQUARE TEST

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Application of Chi-Square Distribution

3.2 Chi-squared test of goodness of fit

3.3 Steps for computing χ^2 and drawing conclusions

3.4 Chi-Square test for independence of attributes

4.0 Conclusion

5.0 Summary

6.0 Assignment

7.0 References/ Further Reading

1.0 INTRODCUTION

The square of a standard normal variable is called a Chi-square variate with 1 degree of freedom, abbreviated as *d.f*. Thus if x is a random variable following normal distribution with mean μ and standard deviation σ, then $(X- \mu)/ \sigma$ is a standard normal variate.

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

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

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

$$= \sum_{i=1}^{v} (\frac{x_1-\mu_1}{\sigma 1})^2$$

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

2.0 OBJECTIVE

The main objective of this unit is to enable students understand the theory behind and the application of chi-square statistics. Students are expected at the end of this unit to be able to apply chi-square analysis to solving day-to-day business and economic problems.

3.0 MAIN CONTENTS

3.1 Applications of the χ^2-Distribution

Chi-square distribution has a number of applications, some of which are enumerated below:

(i) Chi-square test of goodness of fit.

(ii) χ^2-test for independence of attributes

(iii) To test if the population has a specified value of variance σ^2.

(iv) To test the equality of several population proportions

Observed and Theoretical Frequencies

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

Definition of χ^2

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

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

3.2 Chi-Square test of goodness of fit

The chi-square test can be used to determine how well theoretical distributions 9such as the normal and binomial distributions) fit empirical distributions (i.e. those obtained from sample data). Suppose we are given a set of observed frequencies obtained under some experiment and we want to test if the experimental results support a particular hypothesis or theory. Karl Pearson in 1900, developed a test for testing the significance of the discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis. This test is known as χ^2-test of goodness of fit and is used to test if the deviation between

observation (experiment) and theory may be attributed to chance (fluctuations of sampling) or if it is really due to the inadequacy of the theory to fit the observed data.

Under the null hypothesis that there is no significant difference between the observed (experimental and the theoretical or hypothetical values i.e. there is good compatibility between theory and experiment.

Karl Pearson proved that the statistic

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

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots \ldots\ldots\ldots\ldots\ldots \frac{(O_n - E_n)^2}{E_n}$$

Follows χ^2-distribution with $v = n-1$, *d.f.* where $O_1, O_2, \ldots\ldots\ldots\ldots O_n$ are the observed frequencies and $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ are the corresponding expected or theoretical frequencies obtained under some theory or hypothesis.

3.3 Steps for computing χ^2 and drawing conclusions

(i) Compute the expected frequencies $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ corresponding to the observed frequencies $O_1, O_2, \ldots\ldots\ldots\ldots O_n$ under some theory or hypothesis

(ii) Compute the deviations *(O-E)* for each frequency and then square them to obtain $(O-E)^2$.

(iii) Divide the square of the deviations $(O-E)^2$ by the corresponding expected frequency to obtain $(O-E)^2/E$.

(iv) Add values obtained in step (iii) to compute $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

(v) Under the null hypothesis that the theory fits the data well, the statistic follows χ^2-distribution with v = n-1 *d.f.*

(vi) Look for the tabulated (critical) values of χ^2 for (n-1) d.f. at certain level of significance, usually 5% or 1%, from any Chi-square distribution table.

If calculated value of χ^2 obtained in step (iv) is less than the corresponding tabulated value obtained in step (vi), then it is said to be non-significant at the required level of significance. This implies that the discrepancy between observed values (experiment) and the expected values (theory) may be attributed to chance, i.e. fluctuations of sampling. In other words, data do not provide us any evidence against the null hypothesis [given in step (v)] which may, therefore, be accepted at

the required level of significance and we may conclude that there is good correspondence (fit) between theory and experiment.

(vii) On the other hand, if calculated value of χ^2 is greater than the tabulated value, it is said to be significant. In other words, discrepancy between observed and expected frequencies cannot be attributed to chance and we reject the null hypothesis. Thus, we conclude that the experiment does not support the theory.

Example 1: A pair of dice is rolled 500 times with the sums in the table below

Sum (x)	**Observed Frequency**
2	15
3	35
4	49
5	58
6	65
7	76
8	72
9	60
10	35
11	29
12	6

Take $\alpha = 5\%$

It should be noted that the expected sums if the dice are fair, are determined from the distribution of *x* as in the table below:

Sum (x)	***P(x)***
2	$^1/_{36}$
3	$^2/_{36}$
4	$^3/_{36}$
5	$^4/_{36}$
6	$^5/_{36}$

7	$^{6}/_{36}$
8	$^{5}/_{36}$
9	$^{4}/_{36}$
10	$^{3}/_{36}$
11	$^{2}/_{36}$
12	$^{1}/_{36}$

To obtain the expected frequencies, the *P(x)* is multiplied by the total number of trials

Sum (x)	**Observed frequency (O)**	***P(x)***	***Expected Frequency (P(x).500)***
2	15	$^{1}/_{36}$	13.9
3	35	$^{2}/_{36}$	27.8
4	49	$^{3}/_{36}$	41.7
5	58	$^{4}/_{36}$	55.6
6	65	$^{5}/_{36}$	69.5

7	76	$^{6}/_{36}$	83.4
8	72	$^{5}/_{36}$	69.5
9	60	$^{4}/_{36}$	55.6
10	35	$^{3}/_{36}$	41.7
11	29	$^{2}/_{36}$	27.8
12	6	$^{1}/_{36}$	13.9

Recall that $\chi_i^2 = (O_i - E_i)^2/E_i$

Therefore $\chi_1^2 = (O_1 - E_1)^2/E_1 = (15 - 13.9)^2/13.9 = 0.09$

$\chi_2^2 = (O_2 - E_2)^2/E_2 = (35 - 27.8)^2/27.8 = 1.86$

$\chi_3^2 = (O_3 - E_3)^2/E_3 = (49 - 41.7)^2/41.7 = 1.28$

$\chi_4^2 = (O_4 - E_4)^2/E_4 = (58 - 55.6)^2/55.6 = 0.10$

$\chi_5^2 = (O_5 - E_5)^2/E_5 = (65 - 69.5)^2/69.5 = 0.29$

$\chi_6^2 = (O_6 - E_6)^2/E_6 = (76 - 83.4)^2/83.4 = 0.66$

$\chi_7^2 = (O_7 - E_7)^2/E_7 = (72 - 69.5)^2/69.5 = 0.09$

$\chi_8^2 = (O_8 - E_8)^2/E_8 = (60 - 55.6)^2/55.6 = 0.35$

$\chi_9^2 = (O_9 - E_9)^2/E_9 = (35 - 41.7)^2/41.7 = 1.08$

$\chi_{10}^2 = (O_{10} - E_{10})^2/E_{10} = (29 - 27.8)^2/27.8 = 0.05$

$\chi_{11}^2 = (O_{11} - E_{11})^2/E_{11} = (6 - 13.9)^2/13.9 = 4.49$

To calculate the overall Chi-squared value, recall that $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ i.e. we add the individual χ^2 value.

Therefore, $\chi^2 = 0.09 + 1.86 + 1.28 + 0.10 + 0.29 + 0.66 + 0.09 + 0.35 + 1.08 + 0.05 + 4.49$

$\chi^2 = 10.34$

For the critical value, since n=11, *d.f.* = 10

Therefore, table value = 18.3

Decision: since the calculated value which is 10.34 is less than table (critical) value the null hypothesis is accepted.

Conclusion: There is no significant difference between observed and expected frequencies. The slight observed differences occurred due to chance.

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

Digit	0	1	2	3	4	5	6	7	8	9	Total
Frequency	1,02	1,107	997	966	1,075	933	1,107	972	964	853	10,000

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

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

1.4 Chi-Square test for independence of attributes

Consider a given population consisting of *N* items divided into *r* mutually disjoint (exclusive) and exhaustive classes A_1, A_2, ...*Ar* with respect to (*w.r.t*) the attribute A, so that randomly selected item belongs to one and only one of the attributes A_1, A_2, ...*Ar*. Similarly, let us suppose that the same population is divided into *s* mutually disjoint and exhaustive classes B_1, B_2, ...B_s*w.r.t* another attribute B_s so that an item selected at random possesses one and only one of the attributes B_1, B_2, ...B_s can be represented in the following *r x s* manifold contingency e.g like below:

B	B_1	B_2	.....	B_j	.....	B_s	*Total*

A							
A_1	$(A_1\ B_1)$	$(A_1\ B_2)$		(A_1B_j)	.....	(A_1B_s)	(A_1)
A_2	$(A_2\ B_1)$	$(A_2\ B_2)$	.....	(A_2B_j)	.....	(A_2B_s)	(A_2)
:	:	:		.....	.....	:	:
A_i	$(A_i\ B_1)$	$(A_i\ B_2)$	.....	(A_iB_j)	.....	(A_iB_s)	(A_i)
:	:	:		.....	.....	:	:
A_r	$(A_r\ B_1)$	$(A_r\ B_2)$	.....	A_rB_j	.....	(A_rB_s)	(A_r)
Total	(B_1)	(B_2)	.....	(B_j)	.....	(B_s)	$\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$

Where (A_i) is the frequency of the *ith* attribute A_i,i.e, it is, number of persons possessing the attribute A_i , i=1,2, ...r; (B_j) is the number of persons possessing the attribute B_j, j=1,2.....s; and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j ; (i: 1, 2, ...r; j: 1, 2, ..., s)

Under the hypothesis that the two attributes A and B are independent, the expected frequency for (A_i, B_i) is given by

$$E[(A_iB_j)] = N.P\ [A_iB_j] = N.P[A_i \cap B_j] = N.P\ [A_i].\ P[B_j]$$

[By compound probability theorem, since attributes are independent]

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

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

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

Thus, under the null hypothesis of independence of attributes, the expected frequencies for each of the cell frequencies of the above table can be obtained on using this last equation. The rule in the last can be stated in the words as follows:

"*Under the hypothesis of independence of attributes the expected frequency for any of the cell frequencies can be obtained by multiplying the row totals and the column totals in which the frequency occurs and dividing the product by the total frequency N".*

Here, we have a set of *r* x *s* observed frequencies (A_iB_j) and the corresponding expected frequencies $(A_iB_j)_o$. Applying χ^2–test of goodness of fit , the statistic

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

follows χ^2-distribution with (r-1)X(s-1) degrees of freedom.

Comparing this calculated value of χ^2 with the tabulated value for (r-1)X(s-1) *d.f.*and at certain level of significance, we reject or retain the null hypothesis of independence of attributes at that level of significance.

Note: For the contingency table data, the null hypothesis is always set up that the attributes under consideration are independent. It is only under this hypothesis that formula $(A_iB_j)_o = \frac{(A_i)(B_i)}{N}$; *(i = 1, 2, ...,r; j=1,2, ...s)* can be used for computing expected frequencies.

Example: A movie producer is bringing out a new movie. In order to map out her advertising, she wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equally to all age groups. The producer takes a random sample from persons attending a pre-reviewing show of the new movie and obtained the result in the table below. Use Chi-square (χ^2) test to arrive at the conclusion (α=0.05).

	Age-groups (in years)				
Persons	***Under 20***	***20-39***	***40– 59***	***60& over***	***Total***
Liked the movie	320	80	110	200	**710**
Dislikedthe movie	50	15	70	60	**195**
Indifferent	30	5	20	40	**95**
Total	**400**	**100**	**200**	**300**	**1,000**

Solution:

It should be noted that the two attributes being considered here are the age groups of the people and their level of likeness of the new movie. Our concern here is to determine whether the two attributes are independent or not.

Null hypothesis (H_o): Likeness of the of the movie is independent of age group (i.e. the movie appeals the same way to different age group)

Alternative hypothesis (H_a): Likeness of the of the movie depends on age group (i.e. the movie appeals differently across age group)

As earlier explained, to calculate the expected value in the cell of row 1 column 1, we divide the product of row 1 total and column 1 total by the grand total (N) i.e.

$E_{ij} = (A_iB_j)/N$

Therefore, $E_{11} = \frac{710\ X\ 400}{1000} = 284$

$E_{12} = \frac{710\ X\ 100}{1000} = 71$

$E_{13} = \frac{710\ X\ 200}{1000} = 142$

$E_{14} = \frac{710\ X\ 300}{1000} = 213$

$E_{21} = \frac{195\ X\ 400}{1000} = 78$

$E_{22} = \frac{195\ X\ 100}{1000} = 19.5$

$E_{23} = \frac{195\ X\ 200}{1000} = 39$

$E_{24} = \frac{195\ X\ 300}{1000} = 58.5$

$E_{31} = \frac{95\ X\ 400}{1000} = 38$

$E_{32} = \frac{95\ X100}{1000} = 9.5$

$E_{33} = \frac{95\ X200}{1000} = 19$

$E_{34} = \frac{95\ X300}{1000} = 28.5$

We can get a table of expected values from the above computations

Table of expected values

	Under 20	20-39	40-59	60 &above
Like	284	71	142	213
Dislike	78	19.5	39	58.5
Indifferent	38	9.5	19	28.5

χ^2 value = $\sum_i \sum_j [\frac{[(A_i B_i)-(AiBj)o]^2}{(AiBj)o}]$ = χ^2 = $\sum_i \sum_j \frac{(Oij-Eij)^2}{Eij}$ where O_{ij} are the observed frequencies while theE_{ij} are the expected values.

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

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

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

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

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

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

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

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

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

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

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

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

$\chi^2_{\text{calculated}} =$ $\sum_i \sum_j \frac{(Oij-Eij)^2}{Eij}$

=4.56+1.14+7.12+0.79+10.05+1.04+24.64+0.04+1.68+2.13+0.05+4.64 = **57.97**

Recall, that the *d.f.* is (number of row minus one) X (number of column minus one)

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

Decision: Since the calculated χ^2 value is greater than the table (critical value) we shall reject the null hypothesis and accept the alternative.

Conclusion: It can be concluded that the movie appealed differently to different age groups (i.e. likeness of the movie is dependent on age).

4.0 CONCLUSION

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

5.0 SUMMARY

In this unit, we have examined the concept of chi-square and its scope. We also look at its methodology and applications. It has been emphasized that it is not just an ordinary statistical exercise but a practical tool for solving day-to-day business and economic problems.

6.0 TUTOR-MARKED ASSIGNMENT

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

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

H_o: The distribution of test scores is the same for private and public high school students at α=0.05

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

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

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

7.0 REFERENCES/FURTHER READING

OKOJIE, DANIEL E. NOUN TEXT BOOK, Eco 203: Statistics for Economists

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

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

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

UNIT 4:

ANALYSIS OF VARIANCE (ANOVA)

CONTENTS

5.0 Introduction

6.0 Objectives

7.0 Main Content

- 7.1 Assumption for ANOVA test
- 7.2 The one-way classification
- 7.3 Bernoulli Distribution

8.0 Conclusion

9.0 Summary

10.0 Assignment

11.0 References/Further Reading

1.0 INTRODUCTION

In day-to-day business management and in sciences, instances may arise where we need to compare means. If there are only two means e.g. average recharge card expenditure between male and female students in a faculty of a University, the typical t-test for the difference of two means becomes handy to solve this type of problem. However in real life situation man is always confronted with situation where we need to compare more than two means at the same time. The typical t-test for the difference of two means is not capable of handling this type of problem; otherwise, the obvious method is to compare two means at a time by using the t-test earlier treated. This process is very time consuming, since as few as 4 sample means would require ${}^4C_2 = 6$, different tests to compare 6 possible pairs of sample means. Therefore, there must be a procedure that can compare all means simultaneously. One such procedure is the analysis of variance (ANOVA). For instance, we may be interested in the mean telephone recharge expenditures of various groups of students in the university such as student in the faculty of Science, Arts, Social Sciences, Medicine, and Engineering. We may be interested in testing if the average monthly expenditure of students in the five faculties are equal or not or whether they are drawn from the same normal population. The answer to this problem is provided by the technique of analysis of variance. It should be noted that the basic purpose of the analysis of variance is to test the homogeneity of several means.

The term Analysis of Variance was introduced by Prof. R.A Fisher in 1920s to deal with problem s in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes and (ii) chance causes

The variation due to assignable causes can be detected and measured whereas the variation due to chances is beyond the control of human and cannot be traced separately.

2.0 OBJECTIVE

The main objective of this unit is to teach students the theories and application of Analysis of Variance (ANOVA). It is hoped that students should after taking this unit be able to apply ANOVA in solving business and economic problem especially as it concern multiple comparison of means

3.0 MAIN CONTENT

3.1 Assumption for ANOVA test

ANOVA test is based on the test statistic *F* (or variance ratio). For the validity of the *F*-test in ANOVA, the following assumptions are made:

(i) The observations are independent.
(ii) Parent population from which observation are taken are normal.
(iii) Various treatment and environmental effects are additive in nature.

ANOVA as a tool has different dimensions and complexities. ANOVA can be (a) One-way classification or (b) two-way classification. However, the one-way ANOVA we will deal with in this course material.

Note

(i) ANOVA technique enables us to compare several population means simultaneously and thus results in lot of saving in terms of time and money as compared to several experiments required for comparing two populations means at a time.

(ii) The origin of the ANOVA technique lies in agricultural experiments and as such its language is loaded with such terms as treatments, blocks, plots etc. However, ANOVA technique is so versatile that it finds applications in almost all types of design of experiments in various diverse fields such as industry, education, psychology, business, economics etc.

(iii) It should be clearly understood that ANOVA technique is not designed to test equality of several population variances. Rather, its objective is to test the equality of several population means or the homogeneity of several independent sample means.

(iv) In addition to testing the homogeneity of several sample means, the ANOVA technique is now frequently applied in testing the linearity of the fitted regression line or the significance of the correlation ratio.

3.2 The one-way classification

Assuming n sample observations of random variable X are divided into k classes on the basis of some criterion or factor of classification. Let the $i$thclass consist of $n_i$observations and let:

X_{ij} = jth member of the ith class; {j=1,2,.....n_i; i= 1,2, ...k}

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

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

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

Such scheme of classification according to a single criterion is called one-way classification and its analysis of variance is known as one-way analysis of variance.

The total variation in the observations X_{ij}can be split into the following two components:

(i) The variation between the classes or the variation due to different bases of classification (commonly known as treatments in pure sciences, medicine and agriculture). This type of variation is due to assignable causes which can be detected and controlled by human endeavour.

(ii) The variation within the classes, i.e. the inherent variation of the random variable within the observations of a class. This type of variation is due to chance causes which are beyond the control of man.

The main objective of the analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate classes.

Steps for testing hypothesis for more than two means (ANOVA): Here, we adopt the rejection region method and the steps are as follows:

Step1: Set up the hypothesis:

Null Hypothesis: Ho: $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots\ldots = \mu_k$i.e, all means are equal

Alternative hypothesis: H_1 : At least two means are different.

Step 2: Compute the means and standard deviations for each of the by the formular:

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

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

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

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

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

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

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

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

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

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

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

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

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

Which follows *F*-distribution with (v_1 = k-1, v_2 = n-k)*d.f* (This implies that the degrees of freedom are two in number. The first one is the number of classes (treatment) less one, while the second *d.f* is number of observations less number of classes)

Step 8: Find the critical value of the test statistic *F* for the degree of freedom and at desired level of significance in any standard statistical table.

If computed value of test-statistic F is greater than the critical (tabulated) value, reject (H_o, otherwise H_o may be regarded as true.

Step 9: Write the conclusion in simple language.

Example 1: To test the hypothesis that the average number of days a patient is kept in the three local hospitals A, B and C is the same, a random check on the number of days that seven patients stayed in each hospital reveals the following:

Hospital A:	8	5	9	2	7	8	2
Hospital A:	4	3	8	7	7	1	5
Hospital A:	1	4	9	8	7	2	3

Test the hypothesis at 5 percent level of significance.

Solution: Let X_{1j}, X_{2j}, X_{3j} denote the number of days the jth patient stays in the hospitals A, B and C respectively

Calculations for various Sum of Squares

X_{1j}	X_{2j}	X_{3j}	$(X_{1j} - \bar{X}_1)^2$	$(X_{2j} - \bar{X}_2)^2$	$(X_{3j} - \bar{X}_3)^2$
8	4	1	4.5796	1	14.8996
5	3	4	0.7396	4	0.7396
9	8	9	9.8596	9	17.1396
2	7	8	14.8996	4	9.8596
7	7	7	1.2996	4	4.5796
8	1	2	4.5796	16	8.1796
2	5	3	14.8996	0	3.4596
Total=ΣX_{1j} = T_1 = 41	$\Sigma X_{2j} = T_2$ = 35	$\Sigma X_{3j} = T_3$ = 41	$\sum_{j=1}^{7}(X_{1j} - \bar{X}_1)^2$ =50.8572	$\sum_{j=1}^{7}(X_{2j} - \bar{X}_2)^2$ =38	=58.8572

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

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

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

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

Within Sample Sum of Square: To find the variation within the sample, we compute the sum of the square of the deviations of the observations in each sample from the mean values of the respective samples (see the table above)

Sum of Squares within Samples =

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

= 50.8572 + 38 + 58.8572 = 147.7144 ~ 147.71

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

To obtain the variation between samples, we compute the sum of the squares of the deviations of the various sample means from the overall (grand) mean.

$$(\overline{X_1} - \bar{X})^2 = (5.86 - 5.24)^2 = (0.62)^2 = 0.3844;$$

$$(\overline{X_2} - \bar{X})^2 = (5 - 5.24)^2 = (-0.24)^2 = 0.0576;$$

$$(\overline{X_3} - \bar{X})^2 = (4.86 - 5.24)^2 = (-0.38)^2 = 0.1444;$$

Sum of square Between Samples (hospitals):

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

= 7(0.3844) + 7(0.0576) + 7(0.1444)

= 2.6908 + 0.4032 + 1.0108 = 4.1048 = 4.10

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

The total variation in the sample data is obtained on calculating the sum of the squares of the deviations of each sample observation from the grand mean, for all the samples as in the table below:

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

	$(X_{1j} - 5.24)^2$		$(X_{2j} - 5.24)^2$		$(X_{3j} - 5.24)^2$
8	7.6176	4	1.5376	1	17.9776
5	0.0576	3	5.0176	4	1.5376
9	14.1376	8	7.6176	9	14.1376
2	10.4976	7	3.0976	8	7.6176
7	3.0976	7	3.0976	7	3.0976
8	7.6176	1	17.9776	2	10.4976
2	10.4976	5	0.0576	3	5.0176
Total = 41	**53.5232**	**35**	**38.4032**	**34**	**59.8832**

Total sum of squares (TSS) = $\sum (X_{1j} - \bar{X})^2 + \sum (X_{2j} - \bar{X})^2 + \sum (X_{3j} - \bar{X})^2$

= 53.5232 + 38.4032 + 59.8832 = 151.81

Note: Sum of Squares Within Samples + S.S Between Samples = 147.71 + 4.10 =151.81

= Total Sum of Squares

Ordinarily, there is no need to find the sum of squares within the samples (i.e, the error sum of squares), the calculations of which are quite tedious and time consuming. In practice, we find the total sum of squares and between samples sum of squares which are relatively simple to calculate. Finally within samples sum of squares is obtained by subtracting Between Samples Sum of Squares from the Total Sum of Squares:

W.S.S.S = T.S.S – B.S.S.S

Therefore, Within Sample (Error) Sum of Square = 151.8096 – 4.1048 = 147.7044

Degrees of freedom for:

Between classes (hospitals) Sum of Squares = k-1 = 3-1=2

Total Sum of Squares = n-1 = 21-1 = 20

Within Classes (or Error) Sum of Squares = n-k = 21 – 3= 18

ANOVA TABLE

Sources of variation(1)	*d.f*(2)	Sum of Squares(S.S) (3)	Mean Sum of Squares(4) = $\frac{(3)}{(2)}$	Variance Ratio(F)
Between Samples (Hospitals)	3-1 =2	4.10	$\frac{4.10}{2} = 2.05$	$\frac{2.05}{8.21} = 0.25$
Within Sample (Error)	20-2=18	147.71	$\frac{147.71}{18} = 8.21$	
Total	21-1=20	151.81		

Critical Value: The tabulated (critical) value of *F* for *d.f* (v_1=2, v_2=18) *d.f* at 5% level of significance is 3.55

Since the calculated *F* = 0.25 is less than the critical value 3.55, it is not significant. Hence we fail to accept Ho.

However, in cases like this when MSS between classes is less than the MSS within classes, we need not calculate *F* and we may conclude that the means $\overline{X_1}$, $\overline{X_2}$ and $\overline{X_3}$ do not differ significantly. Hence, H_o may be regarded as true.

Conclusion: H_o : $\mu_1 = \mu_2 = \mu_3$, may be regarded as true and we may conclude that there is no significant difference in the average stay at each of the three hospitals.

Critical Difference: If the classes (called treatments in pure sciences) show significant effect then we would be interested to find out which pair(s) of treatment differ significantly. Instead of calculating Student's *t* for different pairs of classes (treatments) means, we calculate the Least Significant Difference (LSD) at the given level of significance. This LSD is also known as Critical Difference (CD).

The LSD between any two classes (treatments) means, say $\overline{X}_i$ and $\overline{X}_j$ at level of significance 'α' is given by:

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

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

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

MSSE means sum of squares due to Error

If the difference $|\overline{X}_i - \overline{X}_j|$ between any two classes (treatments) means is greater than the LSD or CD, it is said to be significant.

Another Method for the computation of various sums of squares

Step 1: Compute: $G = \sum_i \sum_j X_{ij} = Grand\ Total\ of\ all\ observations$

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

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

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

Step 5: Compute

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

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

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

Step 7: Within Classes or Error Sum of Squares = Total S.S – Between Classes S.S

The calculations here are much simpler and shorter than in the first method

Application: Let us now apply this alternative method to solve the same problem treated earlier.

n= Total number of observation = 7 + 7 + 7 = 21

Grand Total (*G)* =

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

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

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

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

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

$= 291 + 213 + 224 = 728$

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

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

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

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

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

Therefore, Within Classes (hospitals) Sum of Squares or Error S.S = TSS – BCSS

= 151.8095 – 4.0957 = 147.7138

Having arrived at the same Sums of Squares figures, computations can proceed as done earlier.

Example 2: The table below gives the retail prices of a commodity in some shops selected at random in four cities of Lagos, Calabar, Kano and Abuja. Carry out the Analysis of Variance

(ANOVA) to test the significance of the differences between the mean prices of the commodity in the four cities.

City	***Price per unit of the commodity in different shops***			
Lagos	9	7	10	8
Calabar	5	4	5	6
Kano	10	8	9	9
Abuja	7	8	9	8

If significant difference is established, calculate the Least Significant Difference (LSD) and use it to compare all the possible combinations of two means (α=0.05).

Solution:

Using the alternative method of obtaining the sum of square

City	***Price per unit of the commodity in different shops***				***Total***	***Means***
Lagos	9	7	10	8	34	8.5
Calabar	5	4	5	6	20	5
Kano	10	8	9	9	36	9
Abuja	7	8	9	8	32	8

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

= 34 + 20 + 36 +32

= 122

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

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

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

= 930.25

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

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

$(7^2 + 8^2 + 9^2 + 8^2)$

= 294 + 102 + 326 + 258

RSS = 980

Total Sum of Square (TSS) = RSS – CF

= 980 – 930.5

TSS= 49.75

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

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

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

BCSS = 289 + 100 + 324 + 256 – 930.25

= 969 – 930.25

BCSS = 38.75

Within Class (cities) or Error Sum of Squares = TSS – BCSS

= TSS –BCSS

= 49.75 – 38.75

WSS = 11

Between Class Mean Sum of Square Error = $\frac{BCSS}{k-1}$; where k is the number of classes

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

= 12.92

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

= 0.92

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

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

$$F_{calculated} = 14.04$$

F-table (critical value) = $F_{(v1, v2, \alpha)} = F_{(3, 12, 0.05)} = 3.49$

Decision: Sincethe computed *F* is greater than the table value $F_{(v1, v2, \alpha)}$, the null hypothesis is rejected and the alternative is accepted.

Conclusion: At least one of the means is significantly different from others.

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

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

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

= 2.18 X $\sqrt{0.46}$

= 2.18 X 0.678

LSD = 1.48

Comparison between different means

Cities	**Absolute Difference**	**Comparison**	**Conclusion**
Lagos and Calabar	$\|8.5 - 5\| = 3.5$	> LSD	Significant

Lagos and Kano	$\|8.5 - 9\| = 0.5$	< LSD	Not Significant
Lagos and Abuja	$\|8.5 - 8\| = 0.5$	< LSD	Not Significant
Calabar and Kano	$\|5 - 9\| = 4$	> LSD	Significant
Calabar and Abuja	$\|5 - 8\| = 3$	> LSD	Significant
Kano and Abuja	$\|9 - 8\| = 1$	< LSD	Not Significant

4.0 CONCLUSION

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

5.0 SUMMARY

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

6.0 TUTOR-MARKED ASSIGNMENT

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

Brand 1: 10, 12, 9, 9

Brand 2: 9, 8, 11, 8, 10

Brand 3: 11, 10, 10, 8, 7

Brand 4: 8, 9, 13, 9

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

7.0 REFERENCES / FURTHER READINGS

OKOJIE, DANIEL E. NOUN TEXT BOOK, Eco 203: Statistics for EconomistsGupta S.C. (2011). *Fundamentals of Statistics*.**(**6th Rev.& Enlarged ed.).Mumbai India, Himalayan Publishing House.

UNIT FIVE:FORECASTING AND TIME SERIES ANALYSIS

CONTENTS

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Steps in Forecasting

3.2 Types of Forecasts

3.3 Methods of Forecasting

3.4 Lease Squares or Trend Lines

3.5 Lease Square Method

4.0 Conclusion

5.0 Summary

6.0 Assignment

7.0 References / Further Reading

1.0 Introduction

Assumptions in Forecasting

Forecasts are based on past performances. In other words, future values are predicted from past values. This assumes that the future will be basically the same as the past and present, implying that the relationships underlying the phenomenon of interest are stable overtime.

Forecasting can be performed at different levels, depending on the use to which it will be put. Simple guessing, based on previous figures, is occasionally adequate. However, where there is a large investment at stake, structured forecasting is essential.

2.0 Objective

Any forecasts made, however technical or structured, should be treated with caution, since the analysis is based on past data and there could be unknown factors present in the future. However it is often reasonable to assume that patterns that have been identified in the analysis of past data will be broadly continued, at least into the short-term future.

3.0 Main Content

3.1 Steps In Forecasting.

We outline the basic steps in forecasting as follows:
Step 1. Garther past data: daily, weekly, monthly, yearly.

Step 2. Adjust or clean up the raw data against inflationary factors. Index numbers can be used in deflating inflationary factors.

Step 3. Make forecasts from the "refined" data

Step 4. When the future data (which is been forecast) becomes available, compare forecasts with actual values, By so doing, one will be able to establish the error due to forecasting.

3.2 Types of Forecasts.

The basic types forecast are outlined below:

1. *Short-term Forecasts.* These are forecasts concerning the near future. They, are characterized by few uncertainties and therefore more accurate then distant future forecasts
2. Long – term forecasts. These concern the distant future. They are characterized by more uncertainties than short – term forecasts.

3. Extrapolation. These are forecasts based solely on past and present values of the variable to be forecast. In this case, future values are extrapolated from past and present values.

4. Forecasts Based on Established Relationships between the variable to be forecast and other variables.

3.3 METHODS OF FORECASTIG

There are generally used methods of forecasting:

I. Moving Averages
II. Trend lines or least squares.

Moving Averages

Moving averages can be used to generate the general picture (or trend) behind a set of data or time series. The general pattern generated can be used to forecast future values.

Note that a time series is a name given to numerical data described over a uniform set of time points. Time series occur naturally in all spheres of business activity.

The method of moving Averages can be illustrated by the following example.

A monthly sales data is given:

Sales (N)	Jan.	Feb.	March.	April.	May.	June
Past (Actual)	50	55	70	50	45	90

Using a 3 – period moving averages, the forecast values are:

$\frac{50 + 55 + 70}{3} = 58$ (Feb)

$\frac{55 + 70 + 50}{3} = 58$ (March)

$\frac{70 + 50 + 45}{3} = 55$ (April)

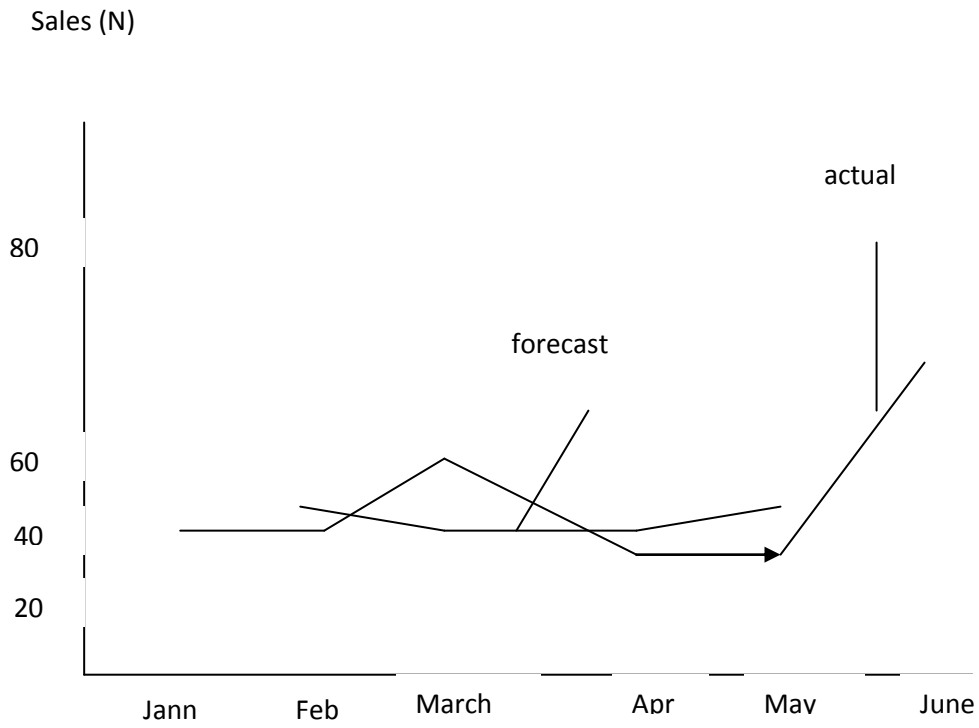
$\frac{50 + 45 + 90}{3} = 62$ (May)

We can thus summarize the forecast sales as follows:

Forecast sales (N)	Jan	Feb	March	April	May	June
Future (forecast)	-	58	58	55	62	-

These can be presented in a graph as in figure 12.1 below:

Figure 3.3 Graph of Sales Forecasts.



*3.4 **Least Squares or Trend Lines***

The idea behind the use of trend lines in forecasting is based on the assumption that the general picture underlying a given set of data can be reasonably approximated by a straight line. Such a straight line can be extended backwards or forward for predicting past or future values.

Example

Suppose the line AB in the following straight line reasonably approximates a set of data for 1995 – 2000

Figure 3.4: Trend line

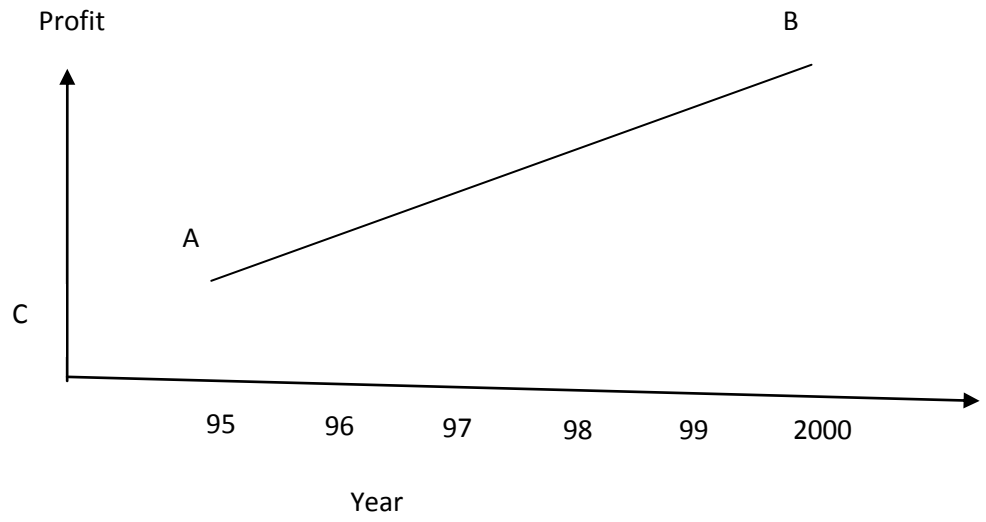


Figure 3.4 indicates that we can forecast profit backwards for years below 1995, using the dotted line AC. Similarly, profits can be forecast for years beyond 2000, using the dotted line BD.

The basic task in using a trend line for forecasting is to determine a line similar to line AB in figure 3.4: then forecasting backwards or forwards is a straight forward activity. The most effective way of determining such a line is the Least-Squares method.

3.5 The Least – Squares Method.

The least – squares method provides a sound mathematical basis for choosing the best trend line; of all possible trend lines for a given set of time series. This method provides an equation (with its numerical coefficients) so that the value corresponding to any given year (or period) can be determined by substituting the given year (or period) into the equation.

For example, consider the following periodic data:

Table 3.1 **Time Series Data**

Year (Period) (t)	**Output** (y)
1990	50
1991	80
1992	90
1993	49
1994	75
1995	58
1996	82
1997	73
1998	95

With t representing period and y representing output, the equation showing the relationship between time and output (or the estimated trend line) is given below:

$$Y = \hat{a} + b\,t$$

The Least – Squares method is then used to determine the numerical values of the parameters, â and b

We assume:

t = 1 in 1990
t = 2 in 1991
t = 3 in 1992
t = 4 in 1993
t = 5 in 1994
t = 6 in 1995
t = 7 in 1996
t = 8 in 1997
t = 9 in 1998

Table 3.1 can thus be rewritten as:

t	**y**
1	50
2	80
3	90
4	49
5	75
6	58
7	82
8	73
9	95

The formulas for the least – squares estimates are as follows:

$$\hat{a} = \bar{Y} - b\,\bar{t}$$

where $\bar{Y} = \frac{\sum Y}{n}$; $\bar{t} = \frac{\sum t}{n}$; n = number of pairs of observations

$$\hat{b} = \frac{n\sum ty - \sum t \sum y}{n\sum t^2 - (\sum t)^2}$$

Using the given data and second formula, we get:

t	y	ty	t^2
1	50	50	1
2	80	160	4
3	90	270	9
4	49	196	16
5	75	375	25
6	58	348	36
7	82	574	49
8	73	584	64
9	95	855	81
45	652	3412	285

Thus, $n \sum t y = 9(3412) = 30708$

$\sum t \sum y = 45(652) = 29340$

$n\sum t^2 = 9(285) = 2565$

$(n\sum t)^2 = (45)^2 = 2025$

$\bar{y} = \frac{\sum y}{n} = \frac{652}{9} = 72.44$

$\bar{t} = \frac{\sum t}{n} = \frac{45}{9} = 5$

It follows that:

$$\hat{b} = \frac{30708 - 29340}{2565 - 2025} = \frac{1368}{540}$$

$= 2.53$

$\hat{a} = y - b\,t$

$= 72.44 - 2.53\,(5)$

$= 72.44 - 12.65$

$= 59.79$

The least – squares line becomes:

$\hat{}$

$Y = 59.79 + 2.53\,t.$

This equation can be used any time to forecast the value of any given year, provided the numerical value of the year is appropriately identified.

For example, let use forecast the value of output, Y, for year 2003.
Following the systematic process, the year 2003 is associated with the numerical value, t = 14, so that for t = 14,

$Y_{2003} = 59.79 + 2.53\,(14)$ (by substitution)

$= 59.79 + 35.4$

$= 95.21$

therefore, the forecast value for output in year 2003 is 95.21

4.0 CONCLUSION

The unit has espoused the theory and application of Forecasting and Time Series Analysis in statistics with special emphasis on its application in the comparison of more than two means.

5.0 SUMMARY

In summary, Time Series is very useful in the multiple comparison of mean among other important uses in both managementl and applied sciences.

6.0 TUTOR-MARKED ASSIGNMENT

1. Calculate a set of moving averages of period:

   (a) 3 (b) 5 for the following time series data:

   8 , 11, 10, 21, 4, 9, 12, 10, 23, 5, 10, 13, 11, 26, 6

   which set of moving averages is the correct one to use for obtaining a trend for the series.

2. The data given in Table below represent the annual gross revenue (in N' millions) obtained by a Telephone company over the periods 1997 – 2006:

Table: Annual Gross Revenues

Year	Gross Revenue (N' million)
1997	13.0
1998	14.1
1999	15.7
2000	17.0
2001	18.4
2002	20.9
2003	23.5
2004	26.2
2005	29.0
2006	32.8

a) Plot the data on a graph paper
b) Fit a least – squares trend line to the data and plot the line on your graph
c) What are your trend forecasts for the years 2009, 2010, 2013, and 2014?

7.0 REFERENCES / FURTHER READINGS

ONWE J.O. NOUN TEXT BOOK, ENT 735: Quantitative Methods for Banking and Finance

OKOJIE, DANIEL E. NOUN TEXT BOOK, Eco 203: Statistics for Economists