INTRODUCTION TO STATISTICAL METHOD

Salihu H. Muhammed Gani
&
Muhammad Kabiru Dahiru

INTRODUCTION TO STATISTICAL METHOD

COPY RIGHT (c) 2014 S.H. Muhammed Gani
M.K. Dahiru

First published 2014

ISBN: 978-978-083-299-5

ALL RIGHTS RESERVED

No part of this publication may be reproduced or transmitted in any form by any means, electronic or mechanical, including photocopying, recording, or by any means or otherwise, without prior permission to the copyright owner.

First published in 2014 by
Azaben Press,
David Mark Bye-pass, High Level, Makurdi
08069313617, 08026731891

Reprinted by:
Onaivi Printing & Publishing Co. Ltd.,
Keffi, Nasarawa State
08036274617

DEDICATION

To Almighty God and Our Country Nigeria.

ACKNOWLEDGEMENT

I acknowledge the teaching of God and his messenger in every moment of my life; I believe that, for Him to keep me among the living is a way of teaching me about life.

I also believe that my students are also my teachers because they tought me the appropriate method of teaching them. They have contributed a lot in the development and the writing of this text. To them I say thank you.

I appreciate the support (both morally and otherwise) given to me by my family and friends. They are the pillars that hold me together and for this, I owe them great respect.

The efforts of my colleagues /teachers in the department are commendable since they are the people that put me through whatever I do in the department. Also commendable is the effort of my publisher as well as all those who put together all my schedules for this work.

PREFACE

The basic idea behind this writing is to free the students and researchers from certain complete arising from the use of foreign and other authors that uses abstract approaches.

Statistics can better be understood when is treated in line with the method to he applied

This text is reader centered, therefore it uses an African approach to statistics and in a language that an average Nigerian reader can understand

Salihu H Mohammed-Gani
Nasarawa State University Keffi.

FORWARD

Most students and researchers alike often have problem with the understanding of statistics and its applications due to certain level of complexes brought in the nature of the language, approach and the notations used in most statistics texts. In most cases, the illustration used are foreign and hence, not quite fitting to our environment.

The approach in this text is such that it closes the gap between abstraction and the application of statistic. Statistics is better understood when treated in connection with the method to which it can be applied. Apart from the endogenous approach used, the language used is more friendly and most convenient for self study with minimal reference to the dictionary. The text takes the reader through most of the course contents and their applications (with relevant illustrations) in a gradual and adaptative . At each stage the work tries to marry abstraction with application.

While the text tries to simplify statistics for the understanding of students and researchers, it does not compromise standard.

This text will go a long way in reducing the phobia in students and researcher, associated with statistics and its methods. The text will therefore make for an important reading.

J.M. Ibbih, Ph.D

TABLE OF CONTENTS

Dedication - - - - - - - iii
Acknowledgments - - - - - - iv
Preface - - - - - - - v
Forward - - - - - - - vi
Table of contents - - - - - - viii

CHAPTER ONE
DEFINITION AND SCOPE OF STATISTICS
Uses of Statistics - - - - - - - 1
Abuse of statistics - - - - - - - 2

CHAPTER TWO
LEVEL OF MEASUREMENT IN SOCIAL SCIENCE
Measurement - - - - - - - - 3
Types of scale or level or measurement - - - - 3
Measurement Techniques - - - - - - 5

CHAPTER THREE
RESEARCH METHODOLOGY
Types of Research - - - - - - - 7
Research Methodology - - - - - - 9
Methodology of Economic Research - - - - - 10
Criterias for verification of a model - - - - - 16
The desirable properties of good model - - - - 18

CHAPTER FOUR
SOURCES AND METHOD OF DATA COLLECTION
Source of Data - - - - - - - 20
Methods of Data Collection - - - - - - 20
Observation method - - - - - - 20
Guidelines for Conduction an observation - - - - 22
Questionnaire Method - - - - - - 22
The essentials of a Good Questionnaire - - - - 23
Questionnaire Design - - - - - - 24

Formation of Questions - - - - - - 24
Personal Interview - - - - - - - 26
The essential properties of a good interviewer - - - 27

CHAPTER FIVE
TECHNIQUES OF DATA COLLECTION
Data Presentation - - - - - - - 29
Frequency Distribution - - - - - - 29
Cumulative Frequency Distribution - - - - - 32
Relative Frequency - - - - - - - 33
Charts - - - - - - - - - 34
Pie Charts - - - - - - - - 34
Bar Chart - - - - - - - - 35
Histogram - - - - - - - - 39
Frequency Polgon - - - - - - - 40

CHAPTER SIX
MEASURES 0F' CENTRAL TENDENCY ANI) DISPERSIONS
Measures of Central Tendency - - - - - 43
Arithmetic mean - - - - - - - 43
The Median - - - - - - - - 46
Mode - - - - - - - - 47
Range - - - - - - - - 48
Quartile - - - - - - - - 49
Quartile deviation - - - - - - - 49
Mean Deviation or Average Deviation - - - - 51
Variance - - - - - - - - 52

CHAPTER SEVEN
PROBABILITY THEORY
Review of set theory - - - - - - 56
Probability - - - - - - - - 60
Fundamentals of Probability - - - - - 61
Approaches or types of Probability - - - - 62
The Axioms of Probability - - - - - 63
Theorems on probability (based on the above Axioms) - - 63

Assignment of Probability - - - - - - 64
Calculations of Probability - - - - - - 64
Unconditional Probability or Marginal Probability - - - 65
Relationship between Conditional and joint Probability - - 66
Joint Probability - - - - - - - 68
Calculation of Conditional Probability - - - - 70
Bayers' Thorem or Rule - - - - - - 71
Counting or Sophisticated ways of Counting - - - 73
Factorial Rule - - - - - - - - 74
Permutation - - - - - - - - 74
Combination Rule - - - - - - - 77

CHAPTER EIGHT
PROBABILITY DISTRIBUTIONS
Random Variable - - - - - - - 80
Probability Distribution - - - - - - 80
Binomial Probability - - - - - - - 85
Poison Distribution - - - - - - - 89
Normal Distribution - - - - - - - 93
Non Standard Distribution - - - - - - 102
Central Limit Theorem - - - - - - 105

CHAPTER NINE
SAMPLING THEORY
Population, Sample Size and Sample Design - - - - 110
Rationale for Sampling - - - - - - 110
Sampling Design - - - - - - - 111
The Determination of a Sample Size - - - - 112
Sampling Techniques - - - - - - 118
Non Probability Sampling - - - - - - 126

CHAPTER TEN
ESTIMATION
Estimator - - - - - - - - 135
Desirable Properties of a good estimator - - - - 135
Critical Value - - - - - - - - 137
Margins of Error (E) - - - - - - 138

CHAPTER ELEVEN
LOGIC OF HYPOTHESES TESTING - - - - 140
Steps involved in hypothesis testing - - - - - 141
Some aspects of hypothesis Testing - - - - - 149

CHAPTER TWELVE
ANALYSIS OF VARIANCE (ANOVA) - - - - 156
ANOVA Test of Difference between Means - - - - 157
ANOVA Summary Table - - - - - 162

CHAPTER THIRTEEN
CURVE FITTING - - - - - - - 164
Steps involved in Curve Fitting - - - - - 164

CHAPTER FOURTEEN
CORRELATION THEORY - - - - - 169
The Rank correlation Coefficient - - - - - 174
Limitation of Analysis of correlation - - - - - 175

CHAPTER FIFTEEN
SIMPLE REGRESSION MODEL - - - - - 176
Conditional Model - - - - - - - 176
The Underlying assumption regarding the error term and - - 179
Assumption of the ordinary least square (OLS) - - - 181
Method regarding the CEF - - - - - 189
Total, explained and residual sums of squares - - - 191
Mean and variance of OLS estimates - - - - 196
Test of Significance of the estimates - - - - - 202
The Gauss Markov Theorem - - - - - 205

CHAPTER SIXTEEN
SOME ASPECTS OF SIMPLE REGRESION MODEL
Time as a Regressor: the Linear trend Model - - - - 212
The lin-log model - - - - - - 222
The log linear model: the Log-log model - - - - 224

CHAPTER SEVENTEEN
MULTIPLE REGRESSIONS MODEL - - - - 228
The Multiple coefficient of multiple determination - - - 234
Hypothesis testing in multiple regression - - - - 239
Hypothesis testing about individual Regression coefficients - 239
Test for Parameter Restriction - - - - - 240

CHAPTER EIGHTEEN
STATISTICAL BIAS - - - - - - 245
Omission of relevant variable or model misspecification - - - 245
Consequence of omission of important variable(under
fitting or mis-specification - - - - - - 248

CHAPTER NINETEEN
THE X^2 CHI-SQUARE - - - - - - 249
X TEST OF GOODNESS OF FIT - - - - - 249
X^2 of independence of events - - - - - 250

CHAPTER TWENTY
NONE PARAMETRIC TESTS - - - - - 256
Sign Test - - - - - - - - 257
Kruska Wallis test - - - - - - - 262

CHAPTER ONE

DEFINATION AND SCOPE OF STATISTICS

The term statistics (when used in singular) refers to actual or specific numbers derived from data. That is to say that statistics can be said to be the science of collection, classification, and the use of data. Therefore it can be referred to as the technique of data analysis. However, the term statistics (when used in plural) is refers to as method of analysis.

Statistics involves much more than the compilation of data and simply drawing graphs and calculating averages. It is a collection of methods for planning experiments, obtaining data, organizing, summarizing, presenting, analyzing, interpreting and drawing conclusions based on the data. (Triola 1998). Clearly the term has two connotations, it can be used to refer to a specific data so collected or a technique. It can also refer to a method of analysis.

Statistical method is the science of reasoning which follows a systematic pattern of reasoning with adequate applications of statistical tools and techniques that involves the use of reasoning which make up scientific method in general. Therefore, adequate use of statistical tools as well as techniques is essential in this regards. Hence, knowledge of statistics and a perfect know how of how to handle statistical tool and techniques is required.

The knowledge and the good handling of statistics methods is not only essential for those who present statistical arguments alone, but it is also essential for those at the receiving end of the arguments. Administrators, Politicians and of course, Social Scientists requires statistics and statistical method to support their arguments, as well as illustrate the problem under study.

1.2 Uses of Statistics

People all over the world and from all works of life tend to appreciate the importance of statistics in their daily activities such that it found application in almost all fields of endeavours. Some of the benefits of the use of statistical methods are as follows:

1. Statistics has found prominence in the health sector such that diseases are controlled through certain statistical observation and analysis that are designed to anticipate epidemics.

2. The use of statistical methods provides better products at lower cost through the use of statistical quality control techniques thereby improving the performance of the manufacturing sector.
3. Statistical estimation of the population size of certain endangered species and other wild life provided for the regulations and laws designed to protect them from extinction.
4. Statistical estimate of the rate of accidents on the high ways can be used by legislators to justify laws such as those of vehicles inspection seat belt, air bag use, drunk driving etc.
5. Statistical method is also used to obtain estimates that can be the basis for the allocation of scarce resource to the various sectors of the economy.
6. This method is also important such that politicians can better allocate their campaign promises on basis of areas needs.
7. Statistical evidence is also important in passing certain critical judgment by judges in the law courts.
8. It is also important in the area of environmental and social studies, such that it can be used for analysis and prediction of possible outcomes of events, estimation of values, making of inferences, generalization and drawing of dependable conclusions.

1.3 Abuse of Statistics

Statistics can be abused through the following ways

1. The use of inappropriate method of collecting data is a source of deceptive statistical analysis.
2. The use of small sample size may lead to a misleading statistical enquiry or results.
3. A survey respondent may give responses that are favourable to his self image if placed under pressure. Therefore acting under pressure by the researcher may yield misleading results.
4. Distorted or unclear percentages may give misleading research result, some researcher corner their research result by deliberately distorting percentage to favour their work.
5. Loaded survey question can make respondents to give misleading responses. Survey questions can be worded to elicit desired responses.
6. Misleading graphs or visual devices can be used to exaggerate the true nature of data used in a research.

CHAPTER TWO

LEVELS OF MEASUREMENT IN SOCIAL SCIENCE

2.1 Measurement:

Notations are in most cases used by researchers to represent the attributes or characteristics of objects, persons, events, concepts or state. Therefore, measurement can be said to be the way and manner in which these notations are defined in respect of the attributes and characteristics consistent of an object, persons, events, concepts or state.

Elements of Measuring Attributes, or Characteristics.

a) Persons, objects, events, or state to be observed.
b) Environmental conditions of the study.
c) Steps to be followed in collecting the observation.
d) Instruments to be used to perform the steps.
e) Observation to be collected.

2.2 Types of Scale or Levels of Measurement

The types of scale or levels of measurement are closely related to the categories of data in question. There are four levels of measurement or categories of data which in most cases are arranged in a hierarchical manner

1. Nominal scale or data
2. Ordinal scale or data
3. Interval scale or data
4. Ratio scale

Nominal Scale or Level of Measurement

This level of measurement is characterized by data that is consistent with names, labels or categories. The scale has no order, distance or origin, hence, the measurement value merely separate the data into different classes or categories. Basic to this scale is to determine if objects or persons belong to the same set, class, or whether they have common properties on the basis of which they can be classified; e.g. male and female, literate and illiterate.

In this scale, the data used cannot be arranged in an ordering scheme like low and high, therefore, tags or symbols are used to identify or categories objects, names etc. For instance, telephone numbers card can be used to

identify those on MTN network or other networks and also numbers are assigned to categories, persons on the same network.

Because the categories lack ordering and significance, the preceding data cannot be used for meaningful calculation. For instance, the number assigned to a telephone subscriber lack a real computational significance and the average calculated with it are meaningless because they are mere code number or symbols. The researcher can only count their numbers and make inference.

Nominal data are qualitative in nature, because they have no inherent numerical properties; they only indicate or identify non-numerical features or qualities.

Ordinal Scale

This level of measurement provides information about relative comparisons such that it is called the ranking scale which possesses the properties of order. It involves data that may be arranged in some order such that the difference between the data value cannot be determined and therefore, meaningless.

In this type of scale, the researcher can rank objects based on their characteristics or attributes, for example, ranking of students according to their performance.

Interval Scale

This type of scale has the properties of nominal order and distance, such that a meaningful amount of difference between data can be determined. Hence, on this basis, one can judge the difference between two objects. Although the scale does not have the inherent (natural) zero starting point, it has arbitrary zero point and constant unit of measurement, the numbers on the scale are placed at equi-distance. The mathematical operations of addition, subtraction and the calculation of mean are possible with any interval scale data. However the operation of multiplication and division are not possible, for instance, in the performance of students between 0-100 percent, the numbers are arranged in equi-distance. Therefore, it would not be rational to say that the person that score zero is completely not intelligent and in the case of two students with 40% and 80%, it is rational to say that the person with 80% has performed better than the person with 40%, but it is not rational to say that the person with 80% is twice as intelligent as the other.

Ratio Scale

This scale can be used for the purpose of categorization, ranking, comparison and the making of rational statement such that it is said to be an interval scale modified to include the inherent (natural) zero starting point. Therefore, values at this level or scale, differences and ratios are both meaningful.

Ratio measurement scale is the highest and the most precised level of measurement such that all mathematical and statistical operations can be performed, therefore, variable like distance, time, weight, and height can be measured.

Since the ratio scale has a unique zero point, the idea of zero distance, zero time, zero height, and zero weight are meaningful in this regard. For instance if athlete A covers *40* kilometres within an hour, athlete B covers 80 kilometres within the same time frame then:

a) Both of them started from the initial zero kilometres.
b) Both of them are fast *(nominal scale)*.
c) Athlete *B* is faster than A *(ordinal scale)*.
d) The difference between them is *(80-40) 40* kilometres (interval scale).
e) Athlete *B* is twice as fast as athlete A (ratio scale) or athlete A covered half the distance covered by athlete *B*, that is 40 kilometres divided by *80* kilometres is 0.5 (ratio).

2.3 Measurement Techniques in Statistics

In statistics, we can classify measurement technique into two, the parametric and the non parametric measurement techniques.

Parametric measures are measures that depend on a particular type of distribution of variables. Example, *the z, t, x^2, f, and h* distributions, or statistics, etc. In parametric measurement, it is assumed that data are normally distributed. That is the assumption of normality is adequately observed.

A non parametric measure does not depend on a particular type of distribution. Both parametric and non parametric measures are used to measure economic variable and to determine the extent and magnitude of the relationship between economic variables. The sign test, rank correlation and kruska wallis test are good examples of non parametric measures of relationships.

In this relationship between economic variable, one variable depends

on the other; that is, one of them is a dependent variable that depends on the other. In other words its value is determine within the equation which defined the relationship, while the other variable is an independent variable whose value is not determined within the equation. (rather, it is determined outside the equation) and hence it becomes an exogenous variables.

Therefore casual relation exist between exogenous (i.e. *the independent variable*) and exogenous variable (i.e. *the dependent variable*), for example, $C = F(Y)$ as stated by economic theory only tell us that there is a relationship between two variables. One is causing the other to change. Mathematical economics carries further explanation on the variables by explaining the level, as well as the type of relationships (that is, either positive or negative), whether the relationship starts from the origin or not. E.g. $C = B_0 + B_1Y$. This equation tells us that there is a linear but positive relationship between consumption and disposable income.

Economists try to go further than this, by attempting to measure or quantify the relationship between these variables and test their magnitude to establish if there is a significant relationship between such variables. E.g $C = B_0 + B_1Y + U$, where U stands for error term which is introduced to capture other variables that are not captured in the model, but are likely to affect the test of the hypothesis given that they have certain influence on the dependent variable.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Researches is a systematic way of seeking for undiscovered information that may lead to the development of knowledge It involves the search processes to examine ideas and to verify their validity leading to the development of generation.

A research methodology is refers to the step by step procedure involve in research seeking for undiscovered information. It is a system of models, procedure and techniques used in the development of knowledge and therefore generalization.

3.2 Types of Research

Research can be classified into fundamental research, applied research, conclusive oriented research, decision oriented research, explanatory research, modeling research and algorithmic research.

Fundamental Research

This type of research is primarily concerned with formulation and testing of theories for generalization. This research has found application in most societies.

Applied Research

This type of research is aimed at providing answers to certain problems facing the society. The main objective of this research is to provide solution to the problems at hand.

Exploratory Research

This research is aimed at establishing and explaining relationship between variables under study without a specific end-objective except to establish these relationships (as many as possible) between variables. Like the fundamental research, this research provides the basis for general findings, which can be used by other researchers in the future. This type of research also provides the foundation for the formulation of different hypothesis of research problems.

This type of research can be further classified as literature survey, experience survey, as well as the study of problem to have an insight.

Literature survey involves the collection of literature in a selected area of research in which the researcher has limited knowledge and critically examine and compare them to have a better understanding. This update of data will be useful for related researches in the future. Examples of this are the review price index and periodic stock index reports.

Experience survey involves the survey of experience of experts or specialists in the selected field which will form the database for future research. The study of insight stimulating study is a special kind of study orientation to have insight into research topics i.e case study.

Conclusive Research

The conclusive research starts whenever the exploratory researches end. This is to say that a conclusive research seeks to test the hypothesis of research problem formulated by the exploratory research and draw definite conclusion.

The conclusive research can be classified as descriptive research and experimental research.

Descriptive Research

Descriptive research is aimed at describing the characteristics of respondents in relation to a particular product or practice of importance.

Experimental Research

Experimental research is aimed at establishing the effects of a set of factors on response variable of a system of study. This research is usually conducted in a controlled environment. ANOVAs can be used to analyze such research.

Modeling Research

Real life problem of present volumes are often reduced to a model since a model is an abstraction from reality. A particular problem of interest to the society or business organization can be formulated as a model. A model can be symbolized model, mathematical model and simulation model.

Symbolic Model

Symbolic model is the representation of the performance measure of the system of interest or value of a system in terms of its variables. That is the attribute of a system are represented by an equation. Regression equation is

an example of a symbolic model.

Mathematical Model

Mathematical models are basically operations research models that seek to solve complex real-life problems arising from the direction and management of large system of men, machines, materials and money in an industry, business and government (panneerselvan 2010). Some examples of mathematical models are the linear programming models, transport models, assignment models etc.

Simulation Model

Simulation model is an imitation of reality which is used to yield a motion of reality. It is an experiment conducted in a real-life stochastic system in a scale time-frame to extract as many average operational statistics as possible to formulate respective decision guidelines at all levels of management in industry, business and government. Simulation model can be discrete or continuous.

3.3 Research Methods

These are various steps (along with rationale for each step) to be adopted by a researcher in studying a problem. Although the sequence of steps or procedures (methodology) is dependent on the problem to be studied. The following steps provide a useful procedural guideline for meaningful research.

1) Problem definition (tentative selection of problem);
2) Initial survey of literature:
3) Selecting and stating the research problem.
4) Extensive literature survey
5) Sources of data
6) Formulation of hypothesis
7) Research design
8) Sampling design
9) Collection of data
10) Execution of the project
11) Analysis of data
12) Testing of hypothesis
13) Arriving at generalization and

14) Preparation of report.

Definition of Problem (Tentative Selection of Problem)

At the beginning of research, the researcher must be able to define the problem he wants to study. In which case, the researcher must single out the problem in consultation with specialists or careful consideration of the situation at hand. The selection of appropriate problem is the basis of success in research.

Initial Survey of Literature

Initially, the researcher is expected to organize or gather what is called "frame of references" for the research. If this frame of reference is successfully gather or developed, then the researcher is able to satisfy his own research problem in meaningful context.

Selection of Research Problem

Selection of research problem and or statement of the problem is closely relted to the initial survey of literature in conjunction with the research guide or specialist which enables the researcher to pick up a topic which he may undertake. The topic should not be too broad or too narrow. This step of defining the research problem is the most important step in research methodology.

The problem to be investigated must be defined unambiguously so as to differentiate between relevant and irrelevant data.

The definition of the problem and appreciating the statement of problem determine the data to be collected, the characteristics of data which are relevant, the relationships to be explored, the choice of techniques to be use and the form of the final report. A researcher writing a thesis for Ph.D is expected to write synopsis of topic and submit it to the research board for approval. Hence, it is important that the researcher should write down a brief summary once the problem is selected.

Extensive Literature Survey

Once the problem is selected and the synopsis topic is written down, the next thing that a researcher should do is to undertake extensive literature survey by abstracting and indexing journals and both published and unpublished bibliographies that are connected to the work at hand. The conceptual literature concerning the concept and the theories may be

reviewed by the researcher. However in doing this the researcher should note that one source will lead him to another. therefore these literature should be carefully studied.

Sources of (Data) Information

The researcher at this juncture should be able to specify the type of information he intends to gather to enable him conduct the research, and the sources of these information. Data can generally be classified as primary data and secondary data.

Primary data refers to the information whose origin is directly as a result of the problem under investigation and therefore, the information is usually collected by the research during the research process.

While the secondary data are information that are collected by another organization or government agencies and already exists.this can be used for investigation. The sources of secondary data are so many and can be doughtful at times, and the data itself are usually eratic. Therefore, care must be taken in the used of these types of data.

Scholars have come up with several tests to take care of some for these problems, such as the unit root test of stationary of time series data, the co integration test, the causality test etc.

Formulation of Research Hypothesis

At this stage the research is required to state the research hypothesis. a research hypothesis is a statement of assumption that is made in order to lend to test the logical and empirical reality of the study. Hypothesis is very important such that. it should be specific to the research at hand. Hypothesis can be classified into the Null hypothesis and the Alternative hypothesis.

The Null hypothesis (usually denoted by Ho) is set in such a way that small evidence can nullify it. For instance we are comparing the performance of level ane students and level two students in terms of attitude to leaving and we state that their attitude is the same then we have set a null hypothesis since small evidence in terms of different can nullify this our assumption.

An Alternative (assumption) hypothesis is set according to the theoretical foundation of the work at hand, for instance theory specified that quantity demand is a negative function of price hence, the higher the price the lower the quantity demanded. While a null hypothesis will states that price does not have any impact on quantity demand, however, an alternative hypothesis will states that price has a negative impact on quantity demand.

The role of hypothesis is to guide the researcher by delimiting the area of search and keeping him on the right track as well as shapens his thinking so as to focus his attention on the important facets of the problem at hand, hypothesis is also important because it indicate the type of data required as well as the approach and the analytical techniques to be used.

The Research Design

The research required to prepare a research design which is the conceptual structure within which the entire research revolves. It is the arrangement of how data are collected and analysed such that the research purpose will be achieved and the research cost minimized. It involves (i) the means of obtaining information (ii) the availability and skill of the research staff (iii) the methodology i.e detail explanation of the way in which selected means of obtaining information such as observation studies, questionnaires, etc. Will be organized and the reasoning leading to selection; (iv) the time and cost of research.

3.4 *Methodology of Economic Research*

The methodology of economics research is refers to the steps in economic research. It is the paths through which economists follow in their effort of investigate economic theory. We can classify the methodology of economic research into several stages. Therefore economists proceed in their analysis of economic problems in the following way.

(1) Define the problem by stating the theory or hypothesis

(2) Specify the mathematical model of the theory

(3) Specify the econometric model

(4) Obtain and confront the model with data

(5) Estimate the parameters of the econometric model.

(6) Evaluate the estimate or verify the model by testing hypothesis and other tests.

(7) Forecasting and prediction

(8) Using the model for control or policy purpose

Definition of the research Problems

A research is associated with problem; therefore economist is always out to provide solutions to certain economic problem. Hence the first thing he does is to define the problem by stating a hypothesis or bringing forth a theoretical statement which is expected to be verified.

For instance if we intend to study the consumption patterns of student of Nasarawa State University Keffi we may bring fort the theoretical statement. We all know from theory that consumption is a function of income i.e. $C = f(y)$. We also know that consumption can be influenced by other factors such as taste, weather etc.

We may decide to state that, as the income of a consumer increase he tends to consume less than proportionate to the change in income. In this way, we are trying state clearly the research problem in such a way that available evidence will assert or disprove it and this will make our research work much easier.

Specification of the Mathematical Form of the Model

After defining the research problem, the next thing to do is to specify the mathematical model of the theory or statement of hypothesis earlier stated. This we can specify the statement of the theory in mathematical form, in such a way that could be better understood.

For instance, scholars postulates that consumption has a positive relationship with income, they did not specify it in mathematical form. Hence, mathematical economists suggests that $C = f(y)$ *and* $C = \beta_0 + \beta_1 y$ *where* $0 < \beta 1 < 1$

Where C = *consumption expenditure,*

y = *income,*

β_0 *and* β_1 = *are parameter of the model.*

A model is a set of mathematical equation with dependent variable at the left hand side and independent variable at the right hand side. Consumption is influenced not only by income alone but by other factors like taste, weather, etc. and this can be specify thus

$$C = f(Y, T, W)$$

$$C = \beta_0 + \beta_1 Y + \beta_2 T + \beta_3 W$$

Where C = *consumption expenditure*

Y = *income*

T = *taste*

W = *weather*

B_0, β_1, β_2, *and* β_3 *are parameters.*

The variables at the right hand side are the independent variables while the only one at the left hand side, (i.e. cons... is a dependent variable which depends on those on the right. It should be noted that the number of explanatory or independent variables to be included in the model depends on the nature of the problem under study and the availability of data in respect of such variables.

Specification of the econometric form of the model

In economic research, only important variables are included in a study, especially those that are easily measurable. The less important and immeasurable variables are taken care of or captured by the introduction of a random variable, usually denoted by U in the model. The value of the random variable cannot be observed like the ones already in the model.
E.g. $C = \beta 0 + \beta_1 Y + \beta_2 T + \beta_3 W + U$.

The introduction of the random error term make the model an econometric model because mathematical economist specify a mathematical form of economic theory without regards to other factors that are less relevant but have influence on the dependent variable or without regards for variables that are not easily measurable but could have influence on the dependent variable.

B_0, β_1, β_2, β_3, otherwise called parameters tells us the relative influence of their respective variables on the dependent variable. Hence, in specifying our model, we should expect that the estimates of these parameters should conform to theoretical or a priori relationship between their respective

variables and the dependent variable. That is to say that we expect them to be positive the respective variables are positively related to the dependent variable.

Data gathering and confronting the model with data

After specifying a model, the next stage is to gather data required and confronting the model with the data required for estimation. Three types of data are commonly used in an empirical research, the time series, the cross sectional and the pooled data.

Time series data is the data that is collected over time; such data may be collected daily, weekly, monthly, quarterly or yearly. It may be quantitative, such as output, income, price, or qualitative such as religion, sex etc. The cross sectional data are data that are collected at a given point in time, such as opinion pool, population census etc.

Both time series and the cross sectional data can be brought together depending on the requirement of the research. Such data is called pooled data. These types of data are usually designed or classify by their sources.

Two sources of data are prominent in empirical research. The primary source of data, which is collected through the use of questionnaire or personal interview and the secondary data, is collected by a source of data which are not collected directly by the researcher himself. Such data are collected by certain government agencies of private organizations.

Estimation of the parameters in the model.

Estimation involves confronting the model with the data gathered so as to obtain numerical values for the estimates of the parameters. The processes of obtaining this value are as follows

(1) Examination of the identification condition of the function in which we are interested, so as to know whether the estimated parameters have a unique value for a given data.

(2) Examination of the problem of aggregation of variable included in the model

(3) Checking the degree of correlation between the explanatory variables, that is the degree of multicollinearity between the explanatory variables

(4) Choosing appropriate econometric technique for the estimation of the function and the critical examination of the underlying assumptions of the chosen technique.

This stage of econometric research requires the knowledge of various econometric methods, their assumptions and the economic implication for the estimates of the parameter. For now, the statistical technique of regression analysis is the main tool use to obtain estimates of the parameters. These estimates give the empirical content to the model in question.

Testing of hypothesis or verification of the model

The process of verification of a model is concerned with the determination of the statistical significance of the estimates of parameters and the entire model.

Assuming that a particular model is a reasonable approximation of reality, we have to develop suitable criteria to find out whether the estimates obtained in the model have conform to the expectation of theory that is been tested.

Before we accept that a finding is a confirmation of the theory, we must enquire whether this estimate is sufficiently below unity to convince us that this has not occurred by chance or peculiar to the particular data we use. Such confirmation of refutation of economic theory on the bases sample evidence is base on a branch of statistical inference (hypothesis testing).

There are three major criteria by which we can verify our model. The a priori criteria, base on economic theory, the statistical criteria base on statistical theory, and the econometric criteria base on econometric method.

3.5 Criterias for verification of a model

The a priori Criteria

Economic theory prescribes, or defines the signs and magnitude of the coefficients of variables. For instance economic theory prescribe that consumption is a positive function of disposable income

$C = f(Y_d)$ ..... *Economic theory*

$C = \beta_0 + \beta_1 Y_d$ ..... *Mathematical economics*

$C = \beta_0 + \beta_1 Y_d + U_i$ .....*Econometrics model*

On a priori bases, we expect that $\beta_0 > 0$ and $\beta_1 > 0$ but less than 1 (i.e. $0 < \beta > 1$) since economic theory prescribed a positive relationship between consumption and disposable income. Unless otherwise stated, these restrictions must always hold.

If the signs and magnitude do not conform to our hypothesis, then we should be able to state the reason for the abnormality. Therefore the a priori criteria are based on theoretical expectation.

Statistical Criteria

The statistical criteria are referring to as the first order tests. The most widely used statistical criteria are the covariance, correlation and standard deviation or standard error. The covariance shows how two or more variables vary together. If two variables vary together, assuming they are dependent, the covariance is the extended value of the product of the variables from the means. Correlation measures the degree of linear relationship between two or more variables.

It may be positive, negative or zero. The coefficient of correlation lies between *-1 to +1*. If two variables do not correlate, *correlation is zero*. If two variables are perfectly positively correlated *the value is 1*. The square of correlation coefficient is known as R^2. Thus, the standard deviation or the standard error measures the dispersion of the variable from their means. Higher standard error implies inefficient estimates while low standard error implies efficient estimates.

Econometric Criteria

The econometric criteria are known as the second-order test and are aimed at the investigating whether or not the assumptions of the econometric method are satisfied. These criteria assist us to determine whether the estimates of our parameter are unbiased, efficient, consistent and sufficient. The criteria help us to detect the violation of the ordinary least square, such as zero mean, constant variance, absence of multicollinearity, serial correlation and the normality of the disturbance term. Econometric criteria make use of *F-test* in testing the overall significance of our model, the stability of coefficient etc. the use of Durbin-Watson test for serial correlation, the *Von-Newmanratio test* have proved valuable in econometric research.

Forecasting

The ability to use the estimated model for the purpose of forecasting is one of the main objectives of econometric technique. If the model does not refute the hypothesis or theory, we may use it to predict the future value of the dependent variable on the basis of know or expected future value(s) of the

explanatory variables.

A researcher can attempt to forecast the future by using the estimate of the model for period not included in our sample. It should be noted that the parameter estimate may be economically meaningful and statistically and economically correct for a sample period under consideration, yet may not be well suitable for forecasting due, for example, to rapid change in the structural parameters of the relationship in the real world.

One way of using the forecasting power of a model is to apply the estimates of the model to a period not included in the sample. The resultant value is usually compared with the actual value and difference between the actual value and the forecast value is subjected to significance test. If the difference is statistically significant on examination, we conclude that the forecasting power of the model is poor.

Another way to establishing the stability of the estimates and the performance of the model in terms of its forecasting power is to estimate the function with an expanded sample. That is the sample includes additional observations. The original estimates will normally differ from the new estimates. The difference is tested for statistical significance, with appropriate technique. This will inform the decision whether to accept or reject the model for lack of forecasting ability.

Consider another measuring the forecasting ability of a model in which the government is contemplating on increasing the level of investment in order to increasing the level of investment in order to increase income and consumption. Assume the marginal propensity to consume *(MPC) is 0.75*, which is *75%* of the increase income goes to consumption. The multiplier is *1/1-0.75 = 4*. In other word, N1million increase in investment will increase the level of income by N4million.

3.6 *The desirable properties of good model*

1. The model must be relevant to the problem at hand. That is how relevant is the model to the problem at hand.
2. The theoretical justification of the model. That is how the model can be justified theoretically.
3. The model must be simple to understand. That is how simple is the model. However, simplicity must not be at the expense of reality.
4. The model must conform to real life situation and the data must be consistent.
5. The model must be stable and consistent.

6. The model must have a good forecasting ability. That is, how powerful is the forecasting ability of the model.

3.7 *Reasons for poor performance of a model*

(1) The value of the explanatory variable used in the forecasting may be inaccurate.

(2) The estimates of the coefficient may be poor due to deficiencies in the sample data.

(3) Sporadic changes in the economic environment may render the estimates unnecessarily fit as basis for forecasting into the future that has different economic outlook from that one in which the initial estimates were obtained.

3.8 *The use of the estimated model for control or policy purpose*

The model of an economy is important to a firm in term of the ability of the firm to forecast the future of its product, but government uses the model to control the economy and to actually make policy decision. The primary objective of any model is controls. Suppose that an economist estimates the consumption function as $C = 10 + 0.8Yd$, where C is consumption and Yd = the disposable income. Assume that government wishes to reduce unemployment rate to about 6% by increasing expenditure to *4,00billion* Naira. What level of disposable income will achieve the desired consumption expenditure? This gives

$$4.00 = 10 + 0.8Yd$$
$$4.00 - 10 = 0.8Yd$$
$$3.90 = 0.8Yd$$
$$Yd = 390/0.8$$
$$= 487.50billion$$

That is an income level of *487.50billion* will produce a consumption expenditure of 400billion. Thus an estimated model can be used for control and policy decision making.

CHAPTER FOUR

SOURCES AND METHOD OF DATA COLLECTION

4.1 Sourses of data

Data can generally be classified as primary data and secondary data.

Primary data refers to the information whose origin is directly as a result of the problem under investigation and therefore, the information is usually collected by the research during the research process.

While the secondary data are information that are collected by another organization or government agencies and already exists, this can be used for investigation. The sources of secondary data are so many and can be doughtful at times. The data itself is usually eratic, therefore, care must be taken in the used of these types of data.

As stated earlier, Scholars have come up with several tests to take care of some for these problems, such as the unit root test of stationary of time series data, the co integration test, the causality test etc.

4.2 Methods of Data Collection

Primary data can be collected through the following methods

1. Observation method
2. Questionnaire method
3. Interview method

4.2 Observation Method

This method is categorised into participant observation and non participant observation. The participant observationis the one in which the observer become part of object or the subject of observation. This means that the researcher become part of the environment which he is studying. For instance, if a researcher is studying certain behaviour of prison inmate, then he must be part of the prison society or a prisoner such that he will be able to study them closely.

A non participant observation is the one in which the observer study or observe from outside of the environment in which he is studying.

Observation method is the most commonly used method of collecting data. The advantage of this method is that it eliminates subjective bias and the information actually relate to what is currently happening. However it is expensive and it gives limited information.

Scientific and non scientific observation

Observation that is adequately planned and carried out in a systematic manner, with the aim of obtaining information that can be used to achieve set objectives which can be trusted and controlled is referred to as a scientific observation, while non scientific observation is the usual watching of events, individuals or process by passersby who have little or no interest in the information so gained by sight.

Structured and Unstructured Observation

Scientific observations are usually structure; hence a structured observation is that which involves a formal but organized procedure with well defined objectives, while an unstructured observation could be formal but have loosely organised procedure.

Laboratory Observation and Non Laboratory Observation

The observation of certain research elements are better done in a controlled (laboratory) environment, in which case a reasonable level of control as well as the application of certain substance or healthy is required, while some issue or element of research in which case, a reasonable level of freedom is required to allow these elements of research to exhibit the characteristic required for observation.

Observation as a systematic technique used to obtain information in order to accomplish research objective. This method of collecting data enjoyed the following advantages.

1. It provides a first-hand information
2. It saves time
3. The information collected are more realistic since the observer observed what is actually happening.
4. It provides wider range of information (i.e deeper information) can be obtain as the observer go closer to the event or individual observed
5. It is relatively cheaper in cost
6. It is a reliable method of collecting data.

Disadvantages

1. It is applicable to relatively small group
2. There is also the problem of observation bias
3. It could be risky especially participant observer in hostage environment

4. It does not study respondents view

4.4 Guidelines for Conducting an Observation

1. The condition of observation should not in any way be tempered with. That is the elements should be observed in their usual condition.
2. The element of observation should not know that they are been studied.
3. The use of rating scale or checklist is important for proper recording of observation.
4. Give enough time before recording the behaviour of an element of study. Do no rush in recording initial behaviour.
5. Be objective in writing both the behaviour and the summary of what you observed.
6. Audio visual device are important in gathering information

4.5 Questionnaire Method

In this type of method, a carefully structured questions are distributed to respondents who answers them and returns same to the researcher who mange and use such information. This is the most used method of collecting data because of the following advantages,

i. It attract low cost no matter the spread and the size of the population
ii. Respondents have the freedom and the time to give a well articulated answer.
iii. The information is free from the bias of the research team.
iv. It enhances convenience in reaching out to difficult respondents
v. It pave way for the collection of large samples
vi. Wider coverage of population
vii. Better accuracy of data
viii. It saves time of data collection

It however has the following disadvantages

i. This method is used only when respondents are literate
ii. There may be problem in returning questionnaire (i.e low rate of returning)
iii. It is not flexible like interview method since when questionnaire are distributed questions can not be changed
iv. The identity of the respondent is not know by they interviewers

v. It may be filled by an assistant and not the desired respondent
vi. The problem of ambiguous responses.

This type of survey is essentially a mailing survey, however, depending on the spread and the size of the population some researcher have resort to the direct distribution of questionnaires, hence they go far and wide to distributes the questionnaires directly to the respondents thereby identifying the respondents and their responses are collected the same way in a later data decided by the respondents.

Both methods of distributing questionnaire requires the following

1. **The determining the mode of distribution**
2. **Designing covering letter**
3. **Deciding questionnaires length**, contents, layout and format
4. Notification and following details
5. Selection of the type of envelop to be used
6. The type of incentive, if any, to be given to potential respondent.

In this type of method a pilot survey is important to test the questionnaires. If is the rehearsal of the main survey, such that it will bring to light the weakness of any, of the questionnaire and the survey technique in relation to the study in question.

4.6 The Essentials of a Good Questionnaire

1. It must be short and simple to understand by the respondent minimum size of a questionnaire has the advantage of not discouraging the respondent.
2. Questions must proceed in logical sequence such that it will move from easy to more difficult questions
3. Technical terms, vague expressions with different interpretations must be avoided in the questionnaire.
4. To a larger extent the number of open ended questions should be minimized since they are often difficult to answer and may discourage the respondent from responding to other question.
5. Questions affecting the sentiments of the respondent must be avoided.
6. Adequate space for the answers should be provided in the questionnaire to aid editing

7. The questionnaire should also have some detection on how to answer questions.

4.7 *Questionnaire Design*

The questionnaire is most important component of this type of data collection method. since the success of the method lie hereby on the capacity of the questionnaire to acquire the required responses. Hence. the need for well designed and articulate questionnaire. depending on the research. the designing of the questionnaire may have the following steps.

i. Identification of research issues
ii. Formulation of question
iii. Deciding questions wording
iv. Deciding appropriate sequence of questions
v. Pre-testing of the questionnaire
vi. Review of the questionnaire for improvement

Identification of research issues: research issues and the objectives define not only the direction of the research itself. but things to be measured in the research. That is to say that it is based on these issue that measurement about the research are made. Therefore the questionnaire will suffer from ambiguous questions of the researcher is no clear about the kind of measurement to be made. Hence. the research issues about which measurements are made should be clearly identified and defined. (i.e the researcher should define the different hypothesis of the research measurements).

4.8 *Formulation of Questions*

If the research issues are clearly identified and defined. it also mean that the type of measurement required are clearly known. this have the researcher with the task of formulating questions that make this measurement possible as well as the question format. To minimize the distortion of the research focus the questions should be such that the respondent will comprehend or understand the intuition of the question.

The format of the question may be close or opend ended although the close ended questions are mostly preferred by most respondent because it provide model answers. the close ended question are sometime used to get the opinion about the components of the system of study.

Deciding Question Wording

In drawing out question the researcher or the interviewer should be able to use familiar language that the respondents can understand and should be able to convey the heart and the feelings of the research. otherwise the respondents may reluctantly give answers that may negatively affect the outcome of the research or the respondents may refuse to answer the questions since he/she does not understand the question.

The Researcher may Consider the Following Guidelines

i. It is important to use a familiar language to the respondents and the question should be simple and direct to the point.
ii. The use of vague language should be avoided
iii. Each question should focus on a particular item or part of the research issues.
iv. Leading questions should be avoided i.e questions that may tell the researcher the supposed answer in the heart of the researcher to avoid bias.
v. Question should be such that may bring out the respondents attitude and feelings in their responses.
vi. It should have clear instructions.
vii. It should be applicable to all respondents in the sample frame
viii. Minimum number of word in a question is also important

Appropriate Sequence Questions

Question should be set such that it would be in a simple form, graduating and adaptive in fashion e.g

1. The first few questions should relate to the personal details of the respondents
2. Introductory questions to establish report with respondents
3. Simple and direct question to build an opinion that the questionnaire is simple answer.
4. Focused questions relating to the objectives and issues of the research
5. Main body of the questionnaire should contains to obtain needed information.

Pre-Testing of Question

A pilot survey should be undertaken in order to test the questionnaire

of as to ascertain the degree to which the respondent understand the questions and to check the relevance of the questions to the objective and issue of the research.

Review of Questionnaire for Improvement

After the pilot survey there is the need to review or prove the questions especially where the questionnaire have not addressed i.e area of inadequate of the questionnaire or whether the questions have actually address the issues and objectives the research.

4.9 Personal Interview

In this method of gathering (data) information. the researcher, the interviewer and the interviewee and the interview environment form the components parts of the interview. It is a survey method of data collection in which the investigator elicits oral information from the respondent in a personal interaction. In this case the interviewer or the investigator uses a structure questionnaire with a properly laid down procedure asking questions therein to the respondent (i.e the interviewee under a defined interview environment).

Closed-Ended and Open-Ended Interviews

Close ended interview is usually a structured interview whereby the investigators abide strictly to the procedure laid down which any open ended interview is usually an unstructured interview during which certain informal distortions are made. In this case the investigator may go beyond the main objective of the research in order to obtain certain information that may help couople the research report.

Advantage of personal interview

1. The information obtain is usually highly reliable such that valid conclusion can be obtain.
2. The investigator may understand the interview more closely and obtain both personal data and characteristics as well as his attributers within his environment
3. The personal contact enables the investigator to used visual aid. This may help the researcher especially if the checklist the responses written down are not adequate.
4. In this type of method the researcher obtains greater returns in terms

of responsesd

5. It provides the researcher the opportunity for return visit to the respondents in case of error inadequately in the information obtained.

Disadvantages

1. It very costly and time consuming
2. The bias element may be high the interviewer may be subjective and emotional
3. The method is limited by the language barrier and the level of education of the respondents.
4. Application to relatively few groups.

4.10 The essential properties of a good interviewer

1. The interview environment must be pleasant to the investigator and the respondents.
2. The investigator or the interviewer must wear a worm and friendly looks
3. There should be an atmosphere of the respondent to provide the information freely.
4. The interviewer must be modest in dressing, vocabulary, fluent and articulate.
5. There must be the absence of complex especially on the part of the respondent.
6. The interviewer must be pleasant, creative and hardworking
7. Both the interviewer and the interviewee must avoid vague languages.
8. The comportment of the interviewer must be excellent etc.

Telephone Interview

This type of the data collection method involves the reaching out to respondent through phone. This will require an efficient telephone system. In some cases it is considered the most cost effective data collection method.

Advantages

1. It is cost effective
2. It is much quicker since, the travel time of the interviewer is eliminated.
3. It pares way for reaching respondents in a distant area

4. It save time and energy
5. It is less risky compared with a participant observer
6. The rate of responses is like to be higher than other methods
7. It is easy to recall telephone interview.

Disadvantages

1. It discovers the use of virtual aids
2. It is restricted to respondents who have access to telephone only
3. The respondent may decide to off his/her phone and discontinue the responses

CHAPTER FIVE

TECHNIQUES OF DATA PRESENTATION

5.1 Data Presentation

Data presentation is the meaningful arrangement of the data collected so as to make them useful as to create better insight into the research problem for which the data is collected. This can be done through the use of the following tools

1. Frequency distribution
2. Cumulative frequency distribution
3. Relative frequency distribution
4. Charts, graphs and tables

5.2 Frequency Distribution

Frequency distribution is the distribution of the number of occurrences of a given raw data. Initially data come in an unorganized manner as follow.

62	83	72	95	37	44	56	63	82	95
95	72	63	95	44	37	63	72	82	95
72	63	37	72	82	44	63	72	82	95
44	56	63	82	72	56	63	72	82	
56	72	56	63	72	56	63	72	82	

Table 5.1 discrete data

Given the above data it is important to organize this information in a frequency distribution table i.e we arrange data according to their number of occurrences. This distribution can be classified into discrete frequency distribution (i.e for grouped data).

Frequency Distribution Table

A frequency distribution table is a table that depicts the number of occurrences of different values of random variables. It shows the different value of a random variable on one side and their corresponding numbers of

occurrences on the other side. To construct a frequency distribution table, a researcher must first arrange the value or group of values of the random variables in a particular order (i.e ascending or descending order), then place a tally of occurrences against each value according to the number of occurrence. Hence the tally marks for each value is equal to the number of occurrences and therefore if counted become its frequency as shown in table 5.2 and 5.3

Monthly Wage	*Tally*	*Frequency*
37	|||	*3*
44	||||	*4*
56	𝍸 ||	*6*
63	𝍸 |||	10
72	𝍸 𝍸 ||	*11*
82	𝍸 ||	8
95	𝍸	6 / 48

Table 5.2 discrete (ungrouped) frequency

The above table shows the frequency of discrete or ungrouped data. In the case of a continuous data that cannot be counted because it has a continuous dismal points or if the data so gathered is too large as to warrant the construction of ungrouped frequency distribution table whereby values of the random variable can be grouped before taking the number of occurrences of each group as shown in table 5.3 below, an appropriate class interval is therefore selected and the data is placed within this interval as shown below.

Class Interval	*Tally*	*Frequency*
34-40	|||	*3*
41-50	||||	*4*
51-60	𝍸 ||	*6*
61-70	𝍸 |||	10
71-80	𝍸 𝍸 ||	*11*
81-90	𝍸 ||	8
91-100	𝍸	6 / 48

Table 5.3 continuous (grouped) frequency distribution table

To construct the above table, the researcher must first decide the class weight of the class interval and grouped the data accordingly and use the same procedure as obtained in the discrete ungrouped frequency distribution table.

From table 5.2 and 5.3 above it is easy to determine the value of the random variables that occur most frequently, i.e the mode of the distribution and that group becomes the modal group, hence, the wage corresponding to the group is the modal wage. A closer look at table 5.3 above reveal the presence of classes and class intervals, class limits and of course class-boundaries, respectively for the given distribution.

The class in this respect represents the range within which the members of a given classes fall. It is a group containing elements or values of a random variable (data) that are closely related in terms of their characteristics. The class of *31-40* in table 5.3 above is a good example of class. Hence a class specifies the range within which a given data fall. The class is determined by the researcher based on the type and volume of data obtained.

The class limits are the end values of the given class. They constitute the terminal values of a given class. That is the lower limit and the upper limit of the class. The lower limit of the first class in table 5.3 is 31, while the upper

limit is 40. Hence, the intervals of a class are the lower and the upper end values of the range within which a given data falls.

The class boundaries are the values which differentiate the different classes. They constitute the points of departure between different classes. The class interval determines whether a data falls in the lower class or in the upper class. In a class of 31-40 for instance 30.5 is the lower class boundary and 40.5 is the upper class boundary.

Therefore the class boundary is the bounder point that separates one class from another. The class mark in the midpoint of each class. It is det rmined by adding up the upper and the lower boundaries together and dividing them by two. It is point exactly half way between the boundaries of a given class. It is important (useful) in the determination of the mean of grouped data. for instance if the lower boundary of a class is *30.5* and the upper boundary is *40.5 (then 30.5 + 40.5 = 71/2 = 35.5) is* the class mark or midpoint of that particular class.

The class width or interval on the other hand is the difference between two successive class boundaries, that is, the lower and the upper boundaries.

Class frequency is the number of appearances, occurrences of values (data) of random variable, falling in a particular class. It is important in the measurements of data that fall in particular class, as well as in the measurements of the mean, the mode as well as the median of grouped data.

5.3 Cumulative Frequency Distribution

This is the addition of the successive frequencies from one particular observation or group to another such that it formed a modified frequency, starting from the first observation or group of observations to the last. Cumulative frequency increase as successive frequency is added

Monthly Wages	Tally	Frequency	Cumulative Frequency
37	III	3	3
44	IIII	4	7
56	IIII I	6	13
63	IIII IIII	10	23
72	IIII IIII I	11	34
82	IIII II	8	42
95	IIII	6	48

Table 5.4 Cumulative frequency table (ungrouped data)

5.4 Relative Frequency

This is the ratio between the frequency of an observation or group and the totalOtal frequency. It is a modifying frequency distribution which is determined by dividing each frequency by the total frequency as shown below.

Class Interval	*Frequency*	*Relative Frequency*
31-40	3	3/48=0.0625
41-50	4	4/48=0.0833
51-60	6	6/48=0.125
61-70	10	9/48=0.208
71-80	11	11/48=0.229
81-90	8	8/48=0.167
91-100	6	6/48=0.125

Table 5.5 relative frequency distribution table

5.5 Charts

These are diagrams of different shapes and sizes with which data can be presented for easier comprehension and analysis. Charts are classified into pie charts, bar charts, multiple or skewed charts, component bar charts, histogram, frequency polygon etc.

5.6 Pie Chart

This is a circle with which data can be categorized within the area of the circle. The chart is divided into different section representing different categories of the data in a form of sectoral division. In this case the angle of each sector is proportional to the frequency (size) of the corresponding category or component. This means that the larger the angle, the bigger the sectoral division or component of the chart, the large the size of the data.

Given the following responses of students of faculty of social science Nassarawa State university Keffi who agreed that the consumption pattern of students change as they move from one level to the other depending on their department, we can construct a pie chart thus

Department	Response of students
Economics	45
Sociology	33
Psychology	46
Political science	96
Geography	24
Total	246

Table 5.6 response of students by department

To construct the pie chart, we consider that the revolution of a circle is 360^0 and the total responses of students is 246 then we can find the angle for each component as follows;

Department	Response of students
Economics	$45/246 \times 360^0 = 55.65^0$
Sociology	$43/246 \times 360^0 = 28.23^0$
Psychology	$46/246 \times 360^0 = 67.32^0$
Political science	$98/246 \times 360^0 = 143.41^0$
Geography	$24/246 \times 360^0 = 35.12^0$

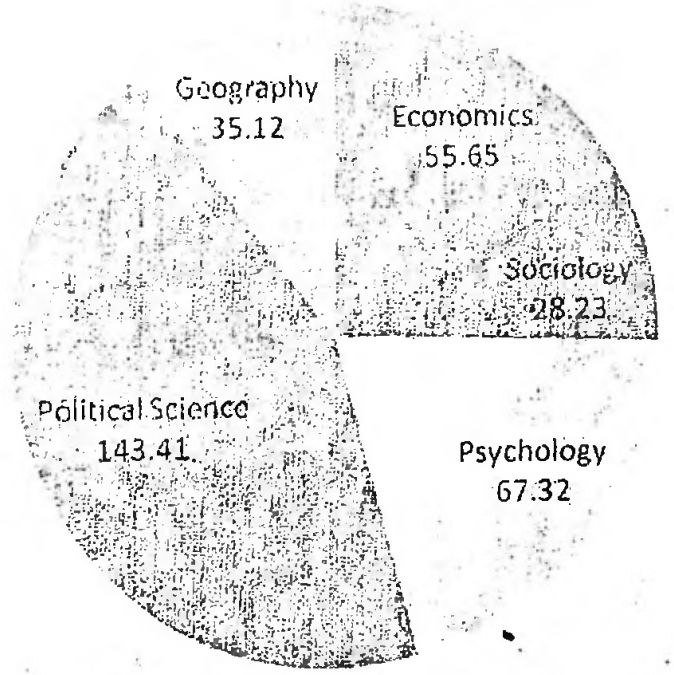


Fig 5.1 Pie charter

The pie chart above clearly indicates the degree covered by responses by each department. It usually depicts the information obtained.

5.7 Bar chart

This is a pictorial representation of data where the data is represented in a form of bars to give a graphical view of the data in questions in such a way that the frequency or size of each category is represented by a vertical bar against other categories in the proportion of its size, relative to that of others. Bar charts are classified into simple bar charts, multiple charts, and stacked bar charts, respectively.

A simple bar chart is a graph that contains separate bars, each representing the frequency or size of each category of the data in question. To construct a simple bar chart, (i) plot a positive graph with its horizontal and vertical axes; (ii) represent the size or frequency of the categories on the vertical axes (in the case of table 2.7, the response on the vertical axes) and the categories (in the case of table 2.7, the departments) on the horinzontai axes; (iii) draw a chart against the category according to their size (i.e the height of the chart should correspond to the size or frequency of that category).

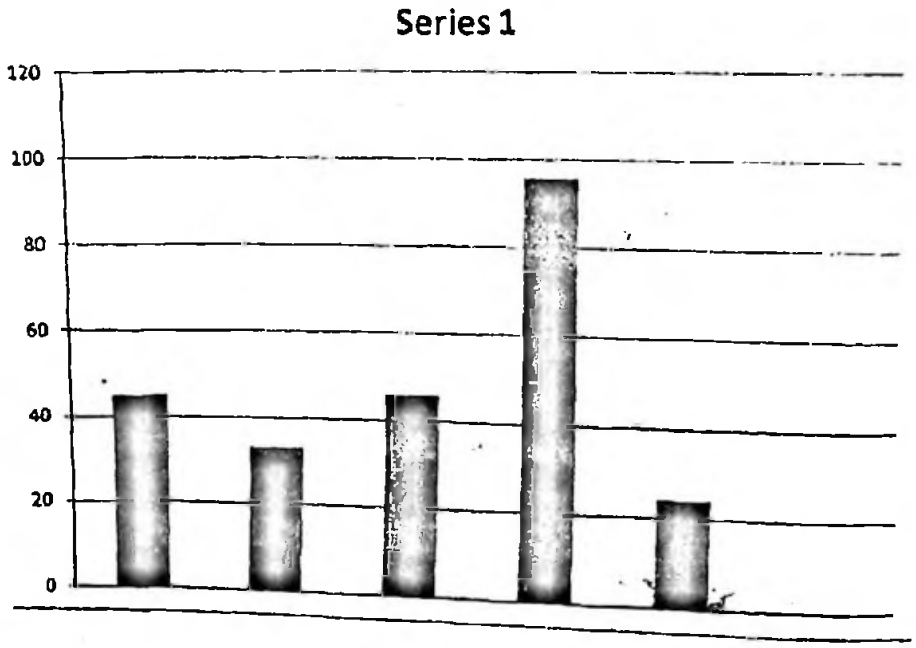


Fig 5.2 a simple bar chart

The responses of students according to their department can be seen at a glance using the bars. Political science has the highest respondents.

Component bar chart

This is a graphical view of a given data represented in one bar chart such that each component of this data is represent only by its size measured within a bar, as can be seen in the fig 2.3 below

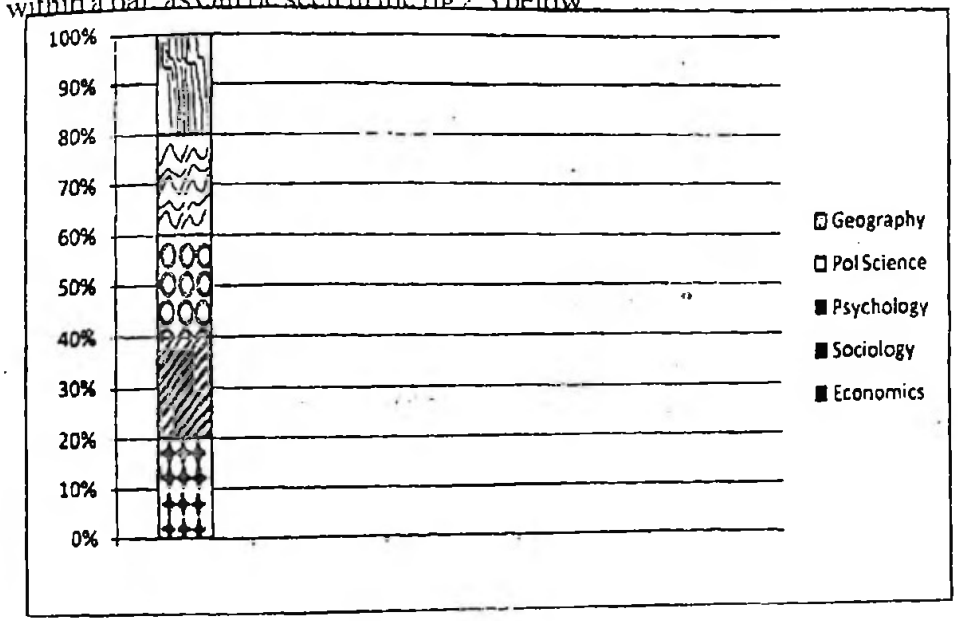


Fig 5.3 component bar chart

Stacked or Multiple Bar Chart:

This is a pictorial representation of multiple data in such a way that the size or height of each category of the multiple data is shown as an adjacent vertical bar against other categories in the proportion of their sizes.

The responses of students of faculty of social science on change in their consumption orientation as they move to higher level according to their departments can be used to demonstrate a stacked or multiple bar charts.

Department	Yes	No	Undecided
Economics	45	40	15
Sociology	33	49	18
Psychology	46	33	21
Political science	98	1	1
Geography	24	52	25
Total	246	179	79

Table 5.7 Reponses on consumption orientation

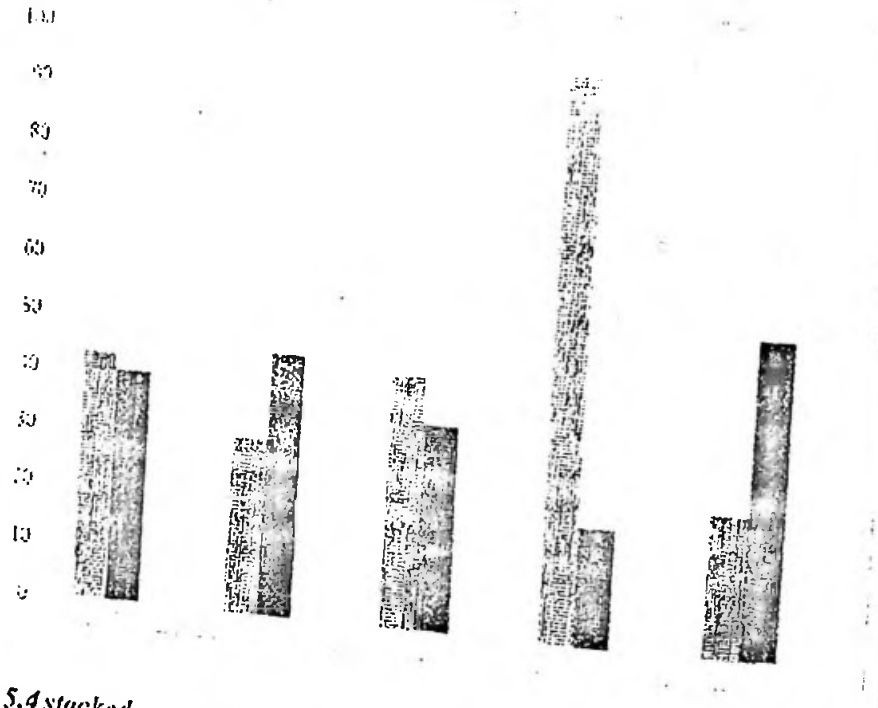


Fig 5.4 stacked or multiple bar chart

The three sets of information obtained from each department can be

seen clearly represented by three bars put together

5.8 *Histogram*

This is the graphical display of a group of data (i.e the entire data) such that the bars are stacked together. It represents the distribution of the pare of both the vertical and the horizontal axes. Histograms are constructed using the class mark or the class boundary and the frequency distribution.

Class Interval	Class Boundaries	Frequency
21-30	20.5-30.5	4
31-40	30.5-40.5	9
41-50	40.5-50.5	12
51-60	50.5-60.5	8
61-70	60.5-70.5	5

Table 5.8 frequency of income level of primary school teachers in keffi local government

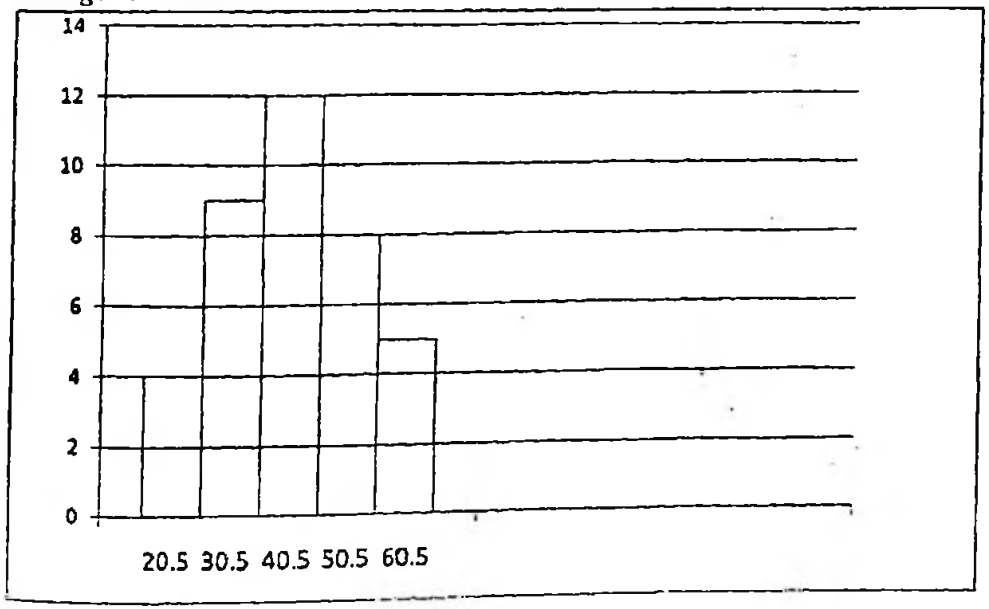


Fig 5.5 histogram

5.9 Frequency Polygon

This is a line that connects the midpoints at the top of the different vertical bars of the histogram

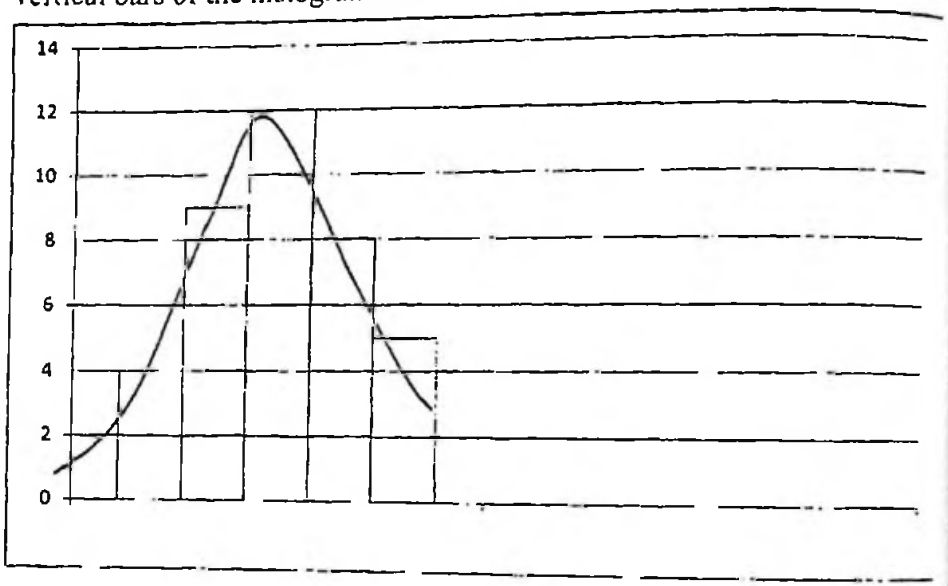


Fig 5.6 frequency polygon

5.10 Cumulative Frequency Curve (Orgive Curve)

This is a cumulative frequency polygon which is important in the presentation of data since it enables the researcher to calculate the mean. the upper quartile and the lower quartile at a glance. It is classified into less that orgive and the more than orgive curves. In the case of the less than orgive curve. the upper limit of each class is ploted on the X-axis against their cumulative frequencies on the Y-axis, while in the case of more than the orgive curve. the lower limits of each class is taken on the X-axis against their cumulative frequencies on Y-axis.

Consider the following data

Class Interval	*Frequency*	*Cumulative Frequency*
Less than 21-30	*4*	*4*
Less than 31-40	*9*	*13*
Less than 41-50	*12*	*25*
Less than 51-60	*8*	*33*
Less than 61-70	*5*	*38*

Table 5.9 less than or give cumulative table

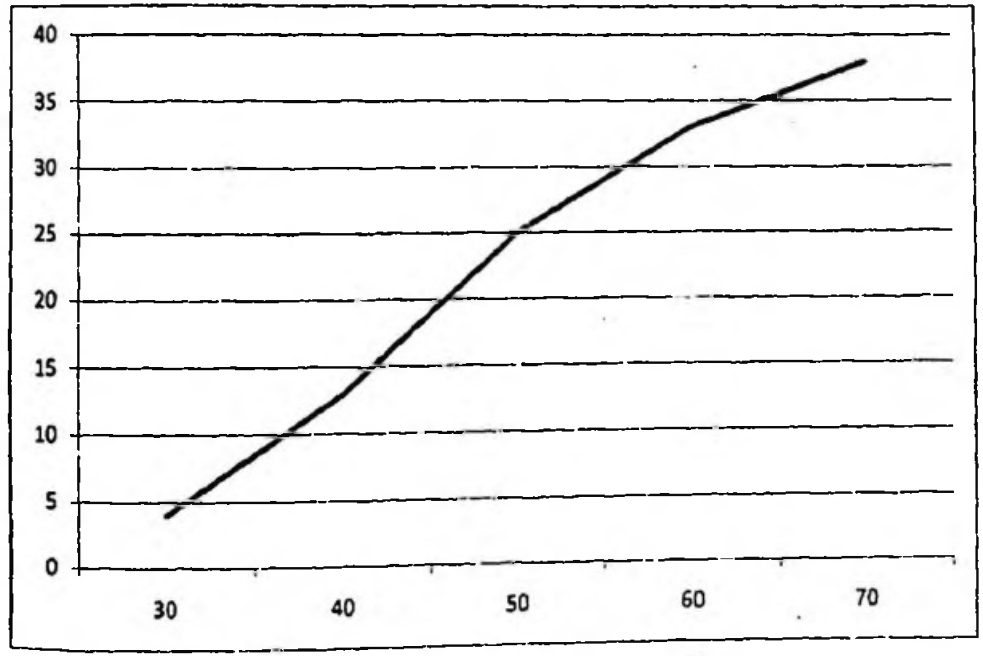


Fig 5.6 less than or give curve

It can be seen that the less than or give is positively slopped indicating that the upper limits are used to plot the graph.

Class Interval	Frequency	Cumulative Frequency
More than 21- 30	4	38
More than 31 - 40	9	34
More than 41 - 50	12	25
More than 51 - 60	8	13
More then 61 - 70	5	5

Class Interval Frequency Cumulative Frequency

Table 5.10 more than or give cumulative table

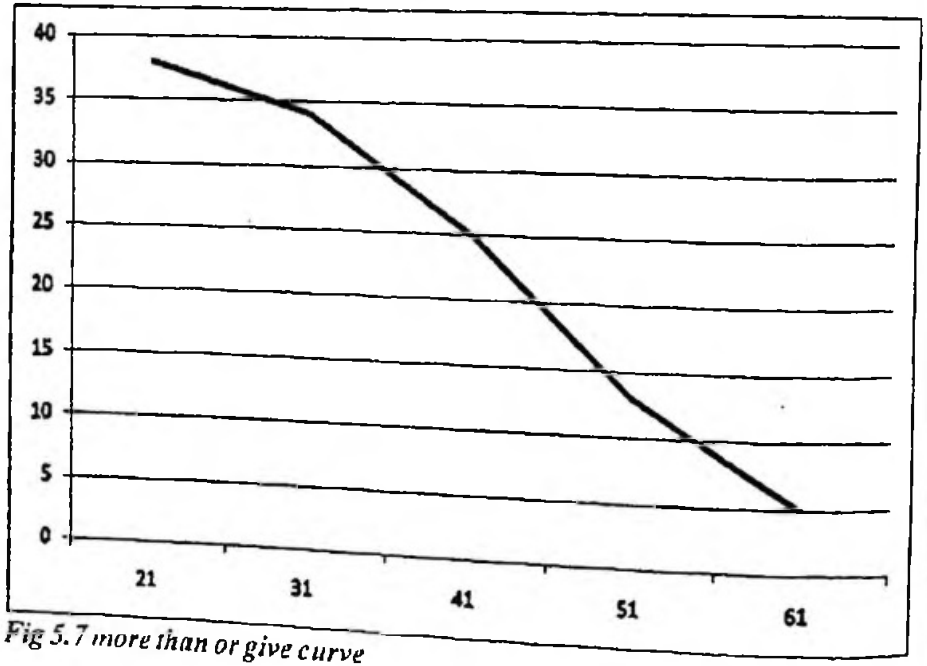


Fig 5.7 more than or give curve

The more than orgive is negatively sloped indicating that the lower limits are used or plotted against the cumulative frequency

CHAPTER SIX

MEASRURES OF CENTRAL TENDENCY AND DISPERSIONS

Data usually tend around the central value(s), hence, data is understood or managed from the view point of these values. The dispersion of any data is also around its central values. Therefore data can be compared through the use of these measures. Statistical measures are classified into (a) measures of central tendency (b) measures of dispersion or variation (c) measures of skewness.

6.1 Measures of central tendency

This is central value within the data upon which all other values in the data tend to.

6.2 Arithmetic mean

This is the most important and generally used descriptive measure. It is the ratio between the sum (total) of the observation (data) and the number of observations. It is classified into that of the mean of ungrouped data and the mean of a grouped data. The mean of ungrouped data is given by:-

$$\overline{X} = \frac{\sum x}{n}$$

Where $\overline{X}$ = *arithmetic mean*, $\sum$ = *summation sign*
x = the observation n = number of observation or sample size

Example 6.1: Given the following scores of student in a statistical class.

No of students	1	2	3	4	5	6	7	8
score	65	35	51	45	40	63	48	56

Table 6.1 scores of students in statistics

$$\bar{X} = \frac{\sum x}{n} = \frac{35+40+45+46+48+51+56+63+65}{6}$$

$$= \frac{449}{9} = 49.89$$

Arithmetic means of a grouped data

This can be measure by the formula

$$\bar{X} = \frac{\sum fx}{\sum f}$$

Example 6.2 consider the following age of students admitted into the department of economics Nasarawa State University Keffi in 2011.

Age Interval	*F*	*x MIDPOINTS*	*Frequency FX*
15-20	*78*	*17.5*	*1365*
21-25	*91*	*23*	*2093*
26-30	*62*	*28*	*1736*
31-35	*10*	*33*	*330*
36-40	*09*	*38*	*342*
	250		*5866*

Table 6.2 age of students

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{5866}{250} = 23.46$$

Then, the average age of students admitted is approximately 24 years

The moving average or weighted mean is given by

$$\bar{X} = \frac{\sum fwxi}{\sum w}$$

Where X = weighted arithmetic means, Xi = the observation wi = weighted of the observations and n = the total number of observation.

Example 6.3 There revenue generated in Nasarawa state for the year *2009, 2010 and 2011 are 200, 250 and 300* million respectively. The corresponding weight are 1,2, and 3. Using the (moving average), the revenue that can be generated in 2012.

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{200 \times 1 + 250 \times 2 + 300 \times 3}{1 + 2 + 3}$$

$$= \frac{200 + 500 + 900}{6}$$

$$= \frac{1600}{6} = 266.67$$

Advantages of Mean as a measure of Location

1) It provides a clear picture of the population under study since it enables the researcher to get his bearing.
2) It forms the basis for comparison in terms of the characteristics of the population.
3) It gives a reliable starting point for the measurement of the distribution of the population.
4) The mean of a population can also be used to make inferences about a population.

6.3 Median

This is the middle value when the data is arranged in ascending or descending order. It corresponds to the division of the entire data into half.

Median of underground data

If the number of scores is odd, then the media is the number that is located in the middle of the list. If the number of scores is even, the media can be found by computing the media of the two middle numbers.

Example 6.4, find the median score of the following students in statistics course

35. 40, 45 46. 51. 63

The two middle numbers are *45 and 46*

Therefore,

Median $= \frac{45 + 46}{2} = \frac{91}{2} = 45.5$

Median of grouped data

$$Median = \frac{L + (N/2)\ F}{f}\ C$$

Where L = lower limit of the median class. N = total number of the (frequency) observation f = frequency of the median class, F = the cumulative frequency of the previous class with reference to the media

Example *6.5*: considering table *6.2* determines the cumulative frequency and age of students admitted into economics department in *2011*.

Age Interval	*Frequency*	*Cumulative Frequency*
15-20	*79*	*78*
21-25	*91*	*169*
25-30	*62*	*231*
31-35	*10*	*241*
36-40	*09*	*250*

Table *6.3* cumulative frequency of the age of students admitted into economics department Nasarawa State University Keffi *2011*.

To find the media $= \frac{L + (N/2)\ F}{f}\ C$

median class $= N/2 = \frac{250}{2} = 125$, *which is within the class 21 25*

When checked both from top and bottom of the frequency

Therefore $L = 21$

$N = 250$

$F = 78$

$F = 91$

$C = 5$

$$Median = L + \frac{(N/2)\ F}{f}\ C$$

$$= 21 + \frac{125\ 78.5}{91}$$

$$= 21 + \frac{47}{91}.5$$

$$= 21 + (0.52)\ (5)$$

$$= 21 + 2.58$$

$$= 23.58$$

6.4 Mode

This is the value in a set of data that occur most frequency

Mode of ungrouped data

This is the value of the observation which has the highest frequency in a set of ungrouped data.

Example 6.8: consider the following scores of students in statistics continue assessment.

26 10 15 29 15

The mode is 15 since it is the frequent occurrence or since it appears twice in distribution.

Mode of ungrouped data,

The mode of ungrouped data is given by:

$$Mode = L + \frac{f_1}{f_1 + f_2} . C$$

Where L = lower limit of the modal class

f_1 = the difference between the frequency of the modal and the frequency before the modal class.

f_2 = the difference between frequency of the modal class and the frequency of the class after the modal class

C = the width of the class interval

Example *6.9* using table 6.3, find the mode

Solution: From table 4.9, the class with highest frequency is *21 25*

Therefore,

L – 21 and the frequency is 91

$F_1 = 91\ 78 = 13$

$F_2 = 91\ 62 = 29$

$C = 5$

Therefore, mode $= L + \frac{f_1}{f_1 + f_2} C$

$$= 21 + \frac{13}{13 + 29} \times 5$$

$$= 21 + \frac{13}{42} \times 5$$

$$= 21 + 0.31 \times 5$$

$$= 21 + 1.547$$

$$= 22.547 = 22.55$$

6 5 *Measure of dispersion (variation)*

Variation refers to the extent to which the (observation) values of random variables vary from each other. It is the spread of observations around the mean.

6.6 Range

The range of a set of observation is the difference between the highest observation and the lowest observation. Therefore,

$R = H - L$

However, the coefficient of range is the ratio between the difference of the values and the sum of the values in question.

$$Coefficient\ of\ range = \frac{H - L}{H \times L}$$

This is also used as a measure of variation

6.7 Quartile

The quartile of a set of values can be regarded as those three points of division of a data set into four equal groups, each representing asset of the items being sampled. It is a sub-population of items defined by classifying individual item according to whether the values of the population concerned falls into one of the four ranges defined by the three values. The mean; the first quartile deviation and the third quartile.

The first quartile is the value of a random variable or item of interest (e.g x) corresponding to the first twenty-five (25%) percent of the total frequency of the grouped data. It is given by

$$Q_1 = L_1 + \frac{(N/4)\ F_1}{f_1} C$$

Where

L = lower limit of the first quartile class

F_1 = the cumulative frequency of the class before the quartile class

f_1 = the frequency of the first quartile class

The third quartile is the value of the random variable or item of interest (e.g.x) corresponding to the last twenty-five *(25%)* percent of the total frequency of the grouped data. It is given by

$$Q_3 = L_3 + \frac{(3N/4)\ F_3}{f_3} C$$

Where L_3 = lower limit of the third quartile class;

F_3 = the frequency of the class before the third quartile class

f_3 = the frequency of the third quartile class.

While $Q_1 = Q_3 - Q_1$

6.8 Quartile deviation

This is half of the difference between the third quartile and the first quartile. It is given by

$$QD = \frac{Q_3 - Q_1}{2}$$

The coefficient of quartile diviation isgiven by

$$\frac{Q_3 - Q_1}{Q_3 \times Q_1}$$

Example 6.10: given the following data on the distribution of *2010/11* GST grades in Economics department Nassarwa State University keffi

Score interval	*no of student(f)*	*cumulative frequency*
10-19	*5*	*5*
20-29	*9*	*14*
30-39	*12*	*26*
40-49	*62*	*88 1st*
50-59	*95*	*183*
60-69	*53*	*236 3rd*
70-79	*29*	*265*
80-89	*2*	*267*

Table 6.4

Determine (a) the first quartile (b) The third quartile (c) the quartile deviation and (d) the coefficient of quartile deviation.

Solution

$N = ?f = 267$

$C = 10$

$N/4 = 267/4 = 66.75$ hence the class interval for which the cumulative frequency just exceed *66.75 is 40 49* therefore, $L_1 = 40, f_1 = 62$. *and* $F_1 = 26$ *then*

$$Q_1 = L_1 + \frac{(N/4)\,F}{f_1} C$$

$$= 40 + \frac{66.75\ 26}{62} .10$$

$$= 40 + \frac{40.75}{62} .10$$

$$= 40 + (0.66)(10)$$

$$= 40 + 6.6$$

$$= 46.6$$

3 (N/4) = 3(267/4) = 3(66.75) = 200.25. Hence, the class interval for which the cumulative frequency just exceed 200.25 is 60 69, therefore, $L_3 = 60, f_3 = 53$, *and* $F_3 = 183$ *then*

$$Q_1 = L_1 + \frac{(N/4)\ F}{f_3} .C$$

$$= 60 + \frac{3(267/4)\ 183}{53} .10$$

$$= 60 + \frac{200.25 - 183}{53} .10$$

$$= 60 + \frac{17.25}{53} .10$$

$$= 60 + (0.325).10$$

$$= 60 + (3.25)$$

$$= 63.25$$

a. $$QD = \frac{Q_3 - Q_1}{2}$$

$$= \frac{63.25 - 46.6}{2}$$

$$= \frac{16.65}{2}$$

$$= 8.33$$

b. $$\text{Coefficient of } QD = \frac{Q_3 \times Q_1}{Q_3\ Q_1} /2$$

$$= \frac{63.25 - 46.60}{2}$$

$$= \frac{16.65}{109.85}$$

$$= 0.15$$

6.9 Mean Deviation or Average Deviation

This is the mean of the arithmetic difference between each observation of a set of data from its mean. That is to say that the difference

between the individual observation and its mean is sumed up and divided by the number of observation. The mean deviation of ungrouped data is given by

$$MD = \frac{\sum (X - \overline{X})}{N}$$

Where X = individual observation

$\overline{X}$ = arithmetic mean of the set of data

N = number of observation

The mean deviation of a grouped data is given by

$$MD = \frac{\sum f_i (X - \overline{X})}{\sum f}$$

Where F = frequency of the individual observation

X = the individual observation

$\overline{X}$ = the arithmetic mean of the set of data

$\sum f$ = total frequency of the set of data

The coefficient of MD is given by

$$Coefficient\ of\ MD = \frac{MD}{Mean}$$

6.10 Variance

This the sum of squares of the difference between the individual observation and its mean divided by the total number of observation. It is determined by squaring the individual deviation from the mean, adding up and dividing the sum by the total number of observations. The variance of ungrouped data is given by

$$S^2 = \frac{\sum (X - X)^2}{N}$$

The variance of grouped data is given by

$$S^2 = \frac{\sum f(X - X)^2}{\sum f}$$

Where $\sum f = N$ = total number of observation

X = observation

X = arithmetic mean

Standard deviation or standard error

This is the square root of variance. It is the most used measure of variability that is used to test hypothesis. It is the standard measure of variance between the observations and their mean. Once the variance is determined, the standard deviation is not far fetched since, it is simply the square root of variance and it is given by

$$\theta = \frac{\sqrt{\sum (x - \bar{x})^2}}{\sum f}$$ *in the case of grouped data while*

$$\theta = \sqrt{\frac{\Sigma (x - \bar{x})^2}{N}}$$ *in the case of ungrouped data*

Example 6.11: Using the frequency distribution in table 6.4: determine the (a) mean deviation (b) variance and (c) the standard deviation.

Class Interval	*Frequency*	*x Midpoints*	*fx*	*(X-X)*
10-19	*5*	*14.5*	*72.5*	*-39.4*
20-29	*9*	*24.5*	*220.5*	*-29.4*
30-39	*12*	*34.5*	*414*	*-19.4*
40-49	*62*	*44.5*	*2759*	*-9.4*
50-59	*95*	*54.5*	*5177.5*	*0.6*
60-69	*53*	*64.5*	*3418.5*	*10.6*
70-79	*29*	*74.4*	*2160.5*	*20.6*
80-89	*2*	*84.5*	*169*	*30.6*
Total	*267*	*396*	*14391.5*	*-35.2*

Table 6.5 midpoints and mean deviations

$F(X-X)$	$(X-X)^2$	$f(X-X)^2$
-197	1552.36	7761.8
-264.6	964.36	7779.24
-232.8	376.36	4516.32
-582.8	88.36	5478.32
57	0.36	34.2
561.8	112.36	5955.08
597.4	424.36	12306.44
61.2	936.36	1852.72
0.2	4354.88	45684.12

Table 6.6 mean deviation

$$X = \frac{\sum fx}{\sum f} = \frac{14391.5}{267} = 53.90$$

$$\sum f(X-X) = 0.2$$

$$\overline{F(X-X)^2} = 45684.12$$

Mean deviation is given by

$$(a)\ MD = \frac{\sum f(X-\overline{X})}{\sum f} = \frac{0.2}{267} = 0.0007$$

Variance is given by

$$\theta^2 = \frac{f(X-\overline{X})^2}{\sum f} = \frac{45684.1}{267} = 171.10$$

$$\theta = \sqrt{\frac{f(X-\overline{X})}{\sum f}} = \sqrt{\frac{45684.12}{267}} = \sqrt{171.10}$$

Example 6.12: Given the following scores of students in (statistics) Continuous Assessment 26, 28, 15, 20, 18, determine the variance and standard deviation

$$Variance = \frac{\sum(X-\overline{X})}{N}$$

In this case of ungrouped data, we first find the mean the same as that of the grouped data

$\sum X = 107$

$N = 05$

$$Therefore\ \overline{X} = \frac{\sum x}{N} = \frac{107}{5.} = 21.4$$

X	$(X-X)$	$(X-X)^2$
26	(26 21.4) = 4.46	21.16
28	(28 21.4) = 6.6	43.56
15	(15 21.4) = 6.4	40.96
20	(20 21.4) = 0.4	0.16
18	(18 21.4) 3.4	11.56
Total		117.4

Table 6.7

Table 6.7 diviation of scores of students in statistics

$$Variance\ \theta^2 = \frac{\sum(X-X)^2}{N} = \frac{117.4}{} =$$

Standard Deviation

Standard deviation is the square root of variance, it is obtained by taking the square root of the value of variance. it is given by

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum(X - \overline{X})^2} = \sqrt{23.48} = 4.85$$

CHAPTER SEVEN

PROBABILITY THEORY

7.1 Review of set theory

The set theory is a mathematical theory that is important in the understanding of probability theory such that it provides a better insight and or platform for the understanding of probability theory as well as probability distributions.

Set concepts and notations

Set concepts and notation not only help us talk about certain mathematical ideas with greater clarity and precision, but are indispensable to a clear understanding of probability.

We can think of a set of any collection of objects specified in such a way that we can tell whether any given object is or is not in the collection. Capital letters, such as *A, B, and C,* are often used to designate particular sets. Each object in a set is called a member or element of the set. Symbolically.

$a \in A$ *means* *" a is an element of set A"*

$a \notin A$ *means* *" a is not an element of set A"*

A set without any elements is called empty or null set. For example, the set of all people over 10 feet tall is an empty set. Symbolically

$\ominus$ *represents "the empty or null set"*

A set is usually described either by listing all its elements between braces { }, or by enclosing a rule within braces that determines the elements of the set. Thus, if $P(x)$ is a statement about x, then

$S = \{x\ P(x)\}$ *means "S is the set of all x such that P (x) is true"*

Recall that the vertical bar in the symbolic form is read "such that". The following example illustrates the rule and listing methods of representing sets.

Example 7.1

rule		*listing*
$\{x\ x$ *is a weekend day*$\}$	=	*{Saturday, Sunday}*
$\{x\ x^2 = 4\}$	=	$\{1,3,5, \ldots\}$

The three dots ... in the last set in Example 7.1 indicate that the pattern established by the first three entries continue indefinitely. The first sets in Example 7:1 are finite sets (intuitively, the elements can be counted the last set is an infinite set (if we try to count the elements, there is end). When listing the elements in a set we do not list an element more than once.

Problem 7.1 Let G be the set of all numbers such that $x^2 = 9$.

(A) Denote G by the rule method.

(B) Denote G by the listing method.

(C) Indicate whether the following are true or false: $3 \in G$, $9 \in G$

Answers *(A)* $\{x\ x^2 = 9\}$ *(B)* $\{-3,3\}$ *(C)* *True, True*

If each element of a set A is also an element of set B, we say that A subset of B. For example, the set of all female students in a class subset of the whole class. Note that the definition allows a set to be a sub-set of itself. If set A and set B have exactly the same elements then, the sets are said to be equal. Symbolically.

$A \subset B$ *means "A is a subset of B"*

$A = B$ *means "A and B have exactly the same elements"*

It can be proved that

? is a subset of every set

Example 7.2 *List all the subsets of the set {a, b, c).*

Solution *{a,b,c}, {a, b}, {a,c}{a}, {b}, {c}, ?*

Problem 7.2 *List all the subsets of the set {1,2),*

Set Operation The union of sets *A and B*, denoted by $A \cup B$, is the set of all elements formed by combining all the elements of *A* and all the elements of *B* into one set. Symbolically.

Union

$A \cup B = \{x\ x \in A \quad or\ x \in B\}$

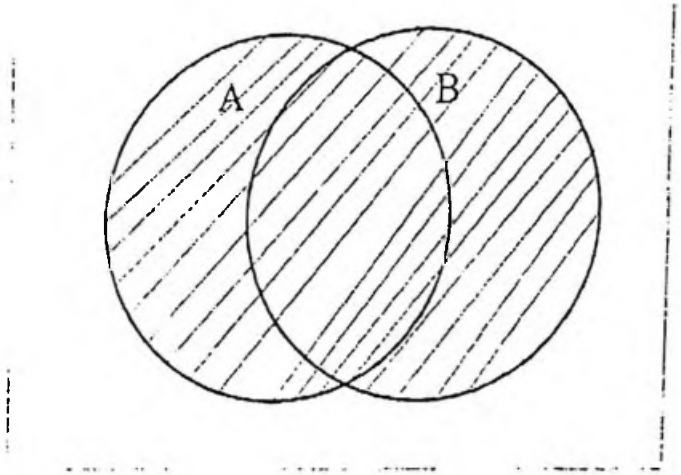


Figure 7.1: A U B is the shaded region

Here we use the word "or" in the way it is always used in mathematics: that is, x may be an element of set A or set B or both. Venn diagrams are useful aids in visualizing set relationships. The union of two sets can be illustrated as shown in Figure 1. The intersection of sets *A and B*, denoted by *A B*, is set of elements in set A that are also in set *B*. Symbolically,

Intersection

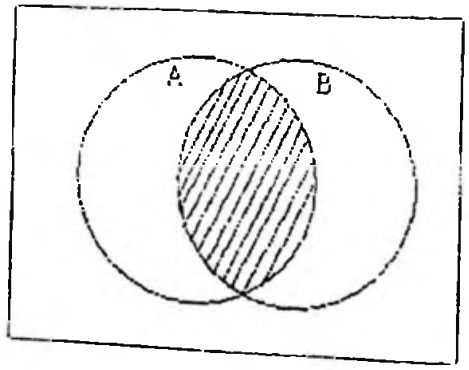


Figure 7.2 = A n B

$A \cap B = \{X \mid X \in A \text{ and } X \in B\}$

This relationship is easily visualized in the Venn diagram shown in Figure 2. If $A \cap B = \emptyset$, then the sets *A and B* are said to be disjoint; this is illustrated in Figure 3.

The set of all elements under consideration is called the universal set *U*. Once the universal set is determined for a particular discussion all other sets in that discussion, must be subsets of U. We now define one more operation on sets called the complement. The complement of A (relative to U) denoted by A', is the set of elements in U that are not in A (see Fig.4). *Symbolically*.

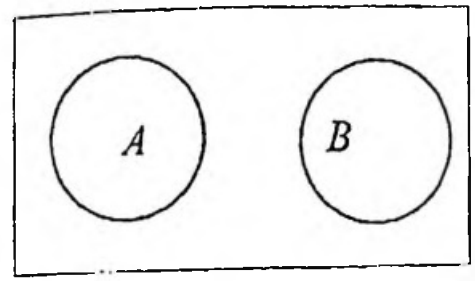
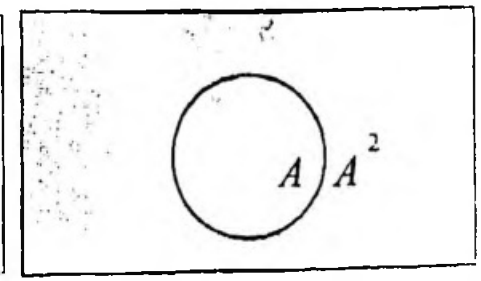


Figure 7.3 A Ù B = ⊖



The complement of A is A'

Complement

$A' = \{x \in U \mid x \notin A\}$

Example 7.3 *If A = {3,6,9}, B = {3,4,5,6,7}, and C = {4,5,7}, then*

A U B = {3,4,5,6,7,9}

A B = {3,6}

A C = ⊖ A and C are disjoint

C' relative to B is {3,6}

Problem 7.3 *If R = {1,2,3,4}, S = {1,3,5,7}, and T = {2,4}, find:*

(A) {1,2,3,4,5,7} (B) {1,3} (C) ⊖ (D) {1,3}

Example 7.4 From a survey involving 100 college students, a marketing research company found that seventy-five students owned stereos, forty-five owned cars, and thirty-five owned cars and stereos.

(A) How many students owned either a car or a stereo?

(B) How many students did not own wither a car or a stereo?

Solutions Venn diagrams are very useful for this type of problem. If we let

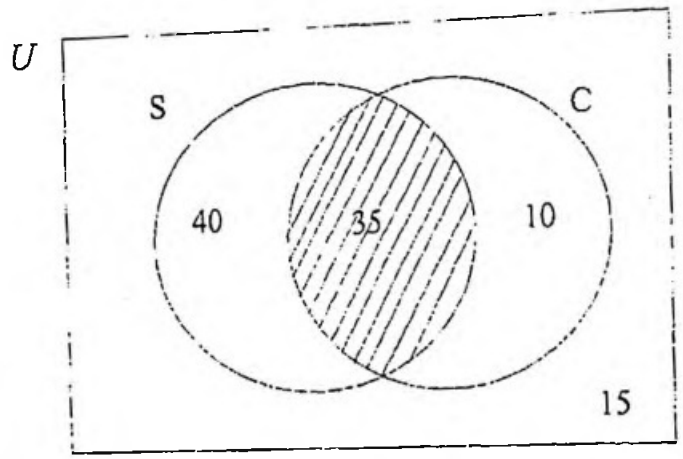
U = Set of students in sample (100)

S = Set of students who own stereos (75)

C = Set of students who own cars (45)

S C = *Set of students who own cars and stereo (35)*

Then



Place the number in the intersection first, then work outward

The number of students who own either a car or a stereo is the number of students in the set $S \cup C$. You might be tempted to say that this is just the number of students in S plus the number of students in C, $75 + 45 = 120$, but this sum is larger than the sample we started with! What is wrong? We have actually counted the number in the intersection *(35)* twice. The correct answer as seen in the Venn diagram, is $40 + 35 + 10 = 85$

(A) The numbers of students who do not own either a car or a stereo is the number of students in the set $(S \cup C)'$: that is, *15*.

Problem 7.4 Referring to Example 4:

(A) How many students owned a car but not a stereo?

(B) How many students did not own both a car and a stereo?

Answers (A) 10 [the number in $S' C$] (b) 65 [the number in $(S \ C)'$]

Note in Example 4 and Problem 4 that the word "and" is associated with "intersection" and the word "or" is associated with "union"

7.2 *Probability*

In most cases, business operations, decision making and of course

inferential making are based on the probabilities or likelihood of events because the occurrence of an event cannot be predicted accurately way before hand due to uncertainty involved in the occurrence of other igniting events. Therefore, probability is an important analytical tool in the society, such that the society should attempt to measure uncertainty. For instance if entrepreneur decides to open another outlet of his business, he must weigh the probability effect of the outlet on his already established headquarter plus the probability of meeting other challenges ahead of such branch in terms of demand and the quality of his product.

7.3 Fundamentals of Probability.

To understand the concept of probability or to understand probability problem, the following terminologies are important and they constitute the basic foundation of probability.

(i) Experiment

(ii) Event

(iii) Sample space

An experiment is a process that produces outcome. The tossing of a coin is an experiment because it can result to certain outcome.

An experiment has the following attributes:-

(1) There will be two or more outcomes

(2) The list of all outcomes can be specified in advance.

(3) Actual occurrence of the outcomes can be predicted in advance.

Sample space is the list of all possible outcomes of an experiment. It consists of all possible simple events that is to say that it consist of all outcomes that cannot be broken down any further. For example in a toss of a coin the possible outcomes are head *(H)* or tail *(T)*. therefore, the sample space of this experiment consist of head (H) and tail *(T)* and cannot be broken any further because they are two simple events. In the case of an experiment with a die, the sample space are *1,2,3,4,5,6.*

Event is considered as the sub-set of a sample space. It is a set of possible outcome such that it could be the occurrence of a particular outcome or a combination of outcomes. For example, the tossing of two similar coins produce the following outcomes *HH, HT, TH, HH*, which constitutes the sample space, but the events of this experiments are *HH, HT, and TT*, which is a simple event in an outcome that cannot be broken down any further. For example the rolling of a single die is an experiment and the result of *2* is an event, therefore the outcome of *2* is a simple event because it cannot be

broken down any further and the sample space consists of these simple events. *1.2.3.4.5.6.*

7.4 Approaches or Types of Probability

1. Classical or apriori approach to probability
2. Relative frequency or empirical approach to probability
3. Objective approach to probability
4. Subjective approach to probability

Classical or apriori probability is based on the assumption that each sample points is equally likely to occur which is known by reason alone such that if an event can occur in 6 different ways out of a total number of n possible ways all of which are equally likely to occur. then, the probability of the event is . Therefore, the probability is given by

P(A) =

And P(B)=

For example if two coins are tossed. the sample space will be *HH. HT. TH. TT*: and the events are *HH. HT, and TT*. Hence, their frequencies are 1.2.1. respectively. The probability of one head and a tail (*HT*) occurring is given by

$$P(HT) = \frac{\text{number of favourable out come}}{\text{Total number of out come}} = 2/4 = 0.5$$

Relative frequency or Empirical probability is based on the records of actual experiences over a period of time such that if after n repeated experiment, and n is very large, an event is observed to occur in c number of times, then the probability of the event is c/n. If an experiment is conducted in large number of times = *n* and an event *A* is said to occur in a number of times, then the probability of event *A (i.e P(A))* is estimated as follows

$$P(A) = \frac{\text{number of time of occurrence of } A}{\text{Total number of trails}}$$

$$= A/n$$

For example if a coin is thrown *100* times and the tail came 60 times. then the probability of tail coming up would be estimated as 60/100 which the relative frequency of occurrence of the event.

Objective probability is based on objective evidence or some cogent Subjective probability is based on personal experiences. This type of probability is constantly needed for business decision purposes. in view of the fact that reliable objective evidence may not be constantly available. This type of probability is estimated as follows

Probability of an event = $\frac{\text{number of successes}}{\text{Total number of trails}}$

7.5 The Axioms of Probability

Notation of probability

P denote a probability

A, B and C denote specific events

P (A) denotes the probability of an event 'A' occurring.

Therefore, if we have a sample space denoted by *S*. hence, *S* is discrete (i.e. countable). then all subject correspond to event. But if S is non discrete (i.e not countable). then only special subject (called measurable) correspond to event.

To each event A in the class of C events we associate a real number *P (A)*. Then P is called probability function as stated above while *P (A)* is the probability of event *A*, if the following axioms are satisfied.

Axiom 1. For every event *A* in a class of *C P(A) = 0*

This means that probability cannot be less than zero. That is to say that probability is positive. Therefore, probability lies between zero and 1.

Axiom 2. For sure or certain event *S* in the class of *C* $P(S) = 1$

This means that for any event that is certain to occur the probability is equal to 1.

Axiom 3. For any number of mutually exclusive event $A_1, A_2, \ldots\ldots$ In the class of *C*

$P(A_1 \cup A_2 \cup \ldots\ldots) = P(A) + P(B) + \ldots\ldots$

In particular for two mutually exclusive events $P(A \cup B) = P(A) + P(B)$.

7.6 Theorems on probability (based on the above Axioms)

Theorem 7-1 if $A_1 (A_2, P(A) = P(A_2)$ *and* $P(A_2, A_1) = P(A_2) - P(A_1)$

Theorem 7-2 for every event A

$0 = P(A) = 1$.

This means that probability lies between 0 and 1.

Theorem 7-3 $P \leq = 0$

This means that an impossible event has zero probability.

Theorem 7-4 if A is a complement of A then

$P(A') = 1 \; P(A)$

Theorem 7-5 *if* $A = A_1 \cup A_2 \cup \ldots\ldots A_n$ *where* $A_1, A_2 \ldots\ldots A_n$ are mutually

exclusive events, then
$P(A) = P(A_1) + P(A_2) + \ldots\ldots P(A_n)$
In particular if A = the sample space S, then
$P(A_1) + P(A_2) \ldots\ldots\ldots P(A_n) = 1$

Theorem 7-6 if A and B are any two events
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
More generally, if A_1, A_2, A_3 are any three events, the $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$.

Theorem 7-7 for any events A and B
$P(A) = P(A \cap B) + P(A \cap B')$

Theorem 7-8 if event A must result in the occurrence of one of mutually exclusive events $A_1, A_2 \ldots\ldots\ldots A_n$ *then* $P(A) = P(A \cap A_1) + P(A \cap A_2) + \ldots\ldots\ldots P(A \cap A_n)$

7.7 *Assignment of Probability*

If a sample space S consist of finite number of outcomes $a_1, a_2, \ldots\ldots a_n$ then according to theorem 7-5 $P(A_1) + P(A_2) + \ldots\ldots\ldots P(A_n) = 1$
Where $A_1, A_2 \ldots\ldots\ldots A_n$ are elementary events given by $A_i = \{a_i\}$. This means that we have assumed equally probability for all simple events, hence
$P(A_i) = \frac{h}{n}$ = ***the classical approach.***

Note that we can use the other approaches to assign probability.

The assignment of probability provides a mathematical model or way of solving probability problems and the success of which must be tested by experiment like other sciences. This had given us the coefficient of estimating probability.

7.8 *Calculations of Probability*

It is very important to consider the nature of the probability in question before estimating it. For instance the way one may estimate a simple event is different from that of a compound event.

Simple or single event

If an event can happen in a ways and tails to happen in 'b' ways, and all of

them are equally likely to occur such that only one of them occur.

then the probability of the event happening is $\frac{a}{(a+b)}$ and the probability of the event not happening is $\frac{b}{(a+b)}$. The two probabilities must sum up to one (i.e happening and not happening). Hence, if one of them is obtained the other can be worked out easily as $[1 - \frac{a}{a+b}]$ and it will be equal to $\frac{b}{a+b}$. Hence the probability of happening can be estimated thus

Probability of an event happening =

$$\frac{Numbers\ of\ favourable\ ways}{Total\ number\ of\ ways\ in\ which\ an\ event\ can\ happen}$$

Example

If a card is drawn at random from an ordinary pack, what is the probability that the card will be a court card?

In this case we have 12 court cards which equal to favourable ways the card will be drawn among *52* cards which is the total number of cards available. Therefore, number of favourable ways = *12*

Total number of ways the card will be drawn = *52*

Hence, $P(court\ card) = \frac{12}{52} = \frac{3}{13}$ = required probability.

Mutually Exclusive Events

The head or tails of a coin are mutually exclusive events. That is to say that if one happened, the other will not happen on any trial. Therefore, the probabilities of the two events can be added up to obtain the probability that at least one of them will occur. This is consistent with the addition rule of probability.

That is to say that if A and B are two events, then the $P(A \cup B) = P(A) + P(B)$ since *A and B* are mutually exclusive.

But if A and B are not mutually exclusive, or can occur consecutively, then $P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$.

See theorem 7-6.

Example ;

In a thrown of a die what is the probability that a number greater than 3 will appear.

The numbers in a die are mutually exclusive (i.e. *1,2,3,4,5,6*) and the number greater than 3 are *4,5,6* which can not appear simultaneously.

If we assume that the die is fair, then $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

Therefore, $P(4 \cup 5 \cup 6) = P(4) + P(5) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

Compound Events

An event may constitute a compound event where two or more events happened simultaneously, in other words the events occur together. Therefore if two independent events happen together, the probability of one may not affect the probability of another. However, if they are dependent events then the probability of one will affect the probability of another. For instance, in picking a black ball in a box containing 6 black balls and 5 red balls,if one pick the black ball and replaces it then the probability of the first black ball will not affect the probability of the other balls. But if the ball was picked without replacement (dependent events) then the probability of the other balls is affected by the probability of the first ball that is picked.

two events are independent (i.e. equal chances) then

$A \cap B)$ or $P(AB) = P(A) \cdot P(B)$

That is the probability of '*A*' multiplied by the probability of *B*. (i.e. *multiplication rule*).

Statistically, independence is expressed as $P(B/A)$ that is, the probability of *B* given that *A* has occurred. This is consistent with the conditional probability of event B given that event *A* has occurred. That is the occurrence of event B is a direct consequence of event *B*.
For independent events $P(B/A) = P(B)$
And $P(A/B) = P(A)$.

Hence, the occurrence of event *B* does not affect the occurrence of event *A*.
In the case of dependent event, the probability that event *A* and event *B*, happened given as

$P(AB) = P(A).\ P(A/B)$ *for* $P(A)$

This indicates that the joint probability of *A and B* is equal to the conditional probability of *B* given that *A* has happened, times the probability of *A*.
Likewise $P(AB) = P(B).\ P(A/B)$ *for* $P(B)$.

7.9 Unconditional Probability or Marginal Probability

This is the probability of occurrence of an event that is not conditional on the occurrence of another event. Hence the $P(A)$ *and* $P(B)$ are cases in example of Marginal probability which has been prepared.

7.9 Relationship between Conditional Marginal and Joint Probability.

$$P(B/A) = \frac{P(AB)}{P(A)}$$

Where $P(B/A)$ = the conditional probability of event *B* given that *A* has occurred.

$P(AB)$ = the joint probability of event *A* and event *B* happening together.

$P(A)$ = Marginal probability of event A happening.

Example

What is the probability of obtaining two heads in two throws of a single coin?

The probability of the first thrown *is* $\frac{1}{2}$ while

The probability of the second thrown *is* $\frac{1}{2}$ and

Both are independent of each other therefore, required probability *is* $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Example

From a box containing six red and 4 black balls, two balls are drawn, one at a time without replacement. Prepare a joint probability table for this experiment and compute the following probabilities.

7.10 Joint Probability

Second Draw / First Draw	Red ball	Black ball	Marginal Probability of Outcome of First Draw
Red ball	$P(R_1R_2) = 0.33$	$P(RB) = 0.27$	0.60
Black ball	$P(B_1B_2) = 0.27$	$P(B_1B_2) = 0.13$	0.40
Marginal probability of second draw	= 0.60	= 0.40	1.00

In the above table the intersection of rows and columns are joint probabilities and they are (Kotharil,( )

$$P(R_1 R_2) = P(R_1) \cdot P(R_2/R_1) = \frac{6}{10} \times \frac{5}{9} = 0.33$$

$$P(R_1 B_2) = P(R_1) \cdot P(B_2/R_1) = \frac{6}{10} \times \frac{4}{9} = 0.27$$

$$P(B_1 R_2) = P(B_1) \cdot P(R_2/B_1) = \frac{4}{10} \times \frac{6}{9} = 0.27$$

$$P(B_1 B_2) = P(B_1) \cdot P(B_2/B_1) = \frac{4}{10} \times \frac{3}{9} = 0.33$$

Because the first ball drawn is not replaced and thus is the case o dependent event.

The column on the right shows the marginal probabilities of th outcomes of the first draw as follows

$$P(R_2) = \frac{6}{10} = 0.6$$

$$P(B_1) = \frac{4}{10} = 0.4$$

The bottom row gives the marginal probabilities of outcomes of th second draws as follows

$$P(R_2) = P(R_1 R_2) + P(B_1 R_2) = 0.33 + 0.27 = 0.6$$

$$P(B_2) = P(R_1 R_2) + P(B_1 B_2) = 0.27 + 0.13 = 0.4$$

Other required probabilities can be worked out and under:

$$P(B_1/B_2) = \frac{P(B2B1)}{P(B)} = \frac{0.13}{0.40} = 0.33$$

$$P(R_2/B_1) = \frac{P(R2B1)}{P(B1)} = \frac{0.27}{0.40} = 0.67$$

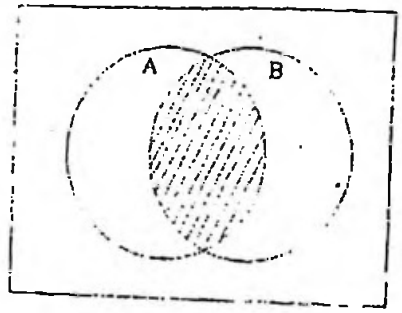
$$P(R_2/R_2) = \frac{P(R2R1)}{P(R1)} = \frac{0.33}{0.60} = 0.55$$ *(Kothari, 1978)*

7.11 Calculation of Conditional Probability

As explained earlier conditional probability is the probability of occurrence of an event given that another event has happened. For instance let A and B are two events. The probability of occurrence of A given that B has occurred is

$$P(A/B) = \frac{P(RnB)}{P(A)} = \frac{P(AB)}{P(A)} .$$

$$P(A \cap B) = P(A) \cdot P(A/B)$$



A∩B

Example

Find the probability that a simple toss of die will result in a number less than *4*

(a) If no information is given and

(b) If it is given that the toss results in an odd number.

a. Let *B* denote event less than *4*. Since *B* is the union of the events *1, 2, or 3* turning up. *Theorem 7-5* state that

$$P(B) = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Assuming equal probability for the sample points,

b. Let *A* be the event *(odd number)* we have *1 and 3* therefore,

$$P(1) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Hence, $P(B/A) = \frac{P(AnB)}{P(A)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$ *given that* $P(A) = \frac{1}{2}$ *while* $P(A \cap B) = \frac{1}{3}$

Theorem of Conditional Probability

Theorem 7-9

For any three events A_1, A_2, A_3 we have $P(A_1 \cap A_2 \cap A_3) = P(A)\, P(A_2/A_1)\, P(A_3/A_1 \cap A_2)$ that is to say that the probability that A_1, A_2 and A_3 all occurred is equal to the probability that A_1 occurred times the probability that A_2 occurred given that A_1 has occurred times the probability A_3 occurred given that both A_1 *and* A_2 have occurred.

Theorem *7-10*

It an event A must result in one of the mutually exclusive events A_1, A_2 ...An then

$$P(A) = P(A_1)\, P(A/A_1) + P(A_2)\, P(A/B_2) + \ldots\ldots\ldots P(A_n)\, P(A/A_n).$$

The case of an independent events that is in a situation where the probability of B occurring is not affected by the occurrence or non-occurrence of event *A (i.e. if* $P(B/A) = P(B)$*.* then $P(A \cap B) = P(A)\, P(B)$.

7.12 Bayers' Theorem or Rule

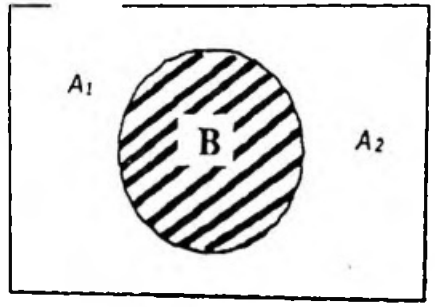
This is otherwise called the probability of causes is a unique method for calculating revised probability revised in the form of investigating the probability already obtained so as to take care of certain new information that would have affected the probability so obtained. Suppose that $A_1, A_2, \ldots\ldots A_n$ are mutually exclusive events. union is the sample space S. i.e., one of the events must occur if A is any event then the following theorem may hold.

Theorem 7-11 (Bayer's Rule)

$$P(A_k/A) = \frac{P(Ak)P\left(\frac{A}{Ak}\right)}{\sum P(Aj)\,P\left(\frac{A}{Aj}\right)}$$

This enable the researcher to find the probability of the various events $A_1, A_2, \ldots\ldots A_n$ that can cause event A to occur.

It is important in the determination of the probability of an event (say a machine not working correctly) given certain information *(say the probability of defective article)*. It has important applications in evaluating the worth of additional information especially with the context of decision analysis. For instance, *let* A_1 *and* A_2 be the sets of events which are mutually exclusive and exhaustive such that $P(A_1) + P(A_2) = 1$, and let B be a simple events which intersect each of A events.



Event B intersecting A events, that is to say that event B is a part of the two side like "*A, and B*" *and* "*A, and B*"

Therefore, the probability of event A, *given event* B is given by P

$$(A/B) = \frac{P(A1 \text{ and } B)}{P(B)}$$

The probability of event A_2 given event B is given by $P(A_2/B) =$

$$\frac{P(A1 \text{ and } B)}{P(B)}$$

Where $P(B) = P(A_1 \text{and } B) + P(A_2 \text{and } B)$

$\text{and } P(A_1 \text{and } B) = P(A_1) X P(B/A_1)$

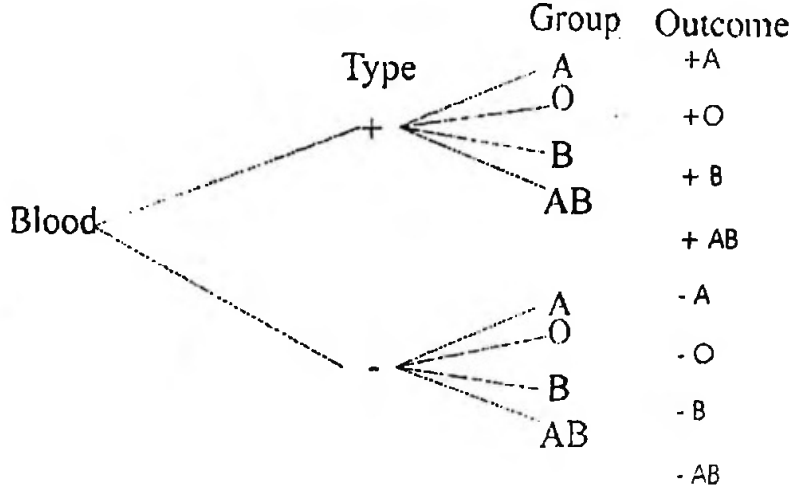
$P(A_2 \text{and } B) = P(A_2) X P(B/A_2)$.

1.13 Counting or Sophisticated Ways of Counting

Sample points in a sample space are not large in most cases, such that direct counting of these sample points to obtain the desired probabilities is not difficult. However, in a situation where these sample point become so large that direct counting is not possible, then certain sophisticated ways of counting are necessary.

Counting Rule: For a sequence of two events in which the first event can occur m ways and the second event can occur n ways, the event together can occur a total of '$m \times n$' ways.

For instance, if a physician wants to select at random one of the two blood types and one of the 4 blood group, the total number possibilities is $2.4 = 8$



: in an organizations how many versions of questioning are open to the interviewer if all possible arrangements are made.

Solution: the interviewer has 5 possible choices of asking his first question, 4 possible choices of asking his second question, 3 choices for the third question, 2 choices for the fourth question and 1 choice of asking the fifth question of the questions is $5.4.3.2.1 = 120$

This brings us to the issue of factorial since the solution above can be generalized or solved using factorial rule.

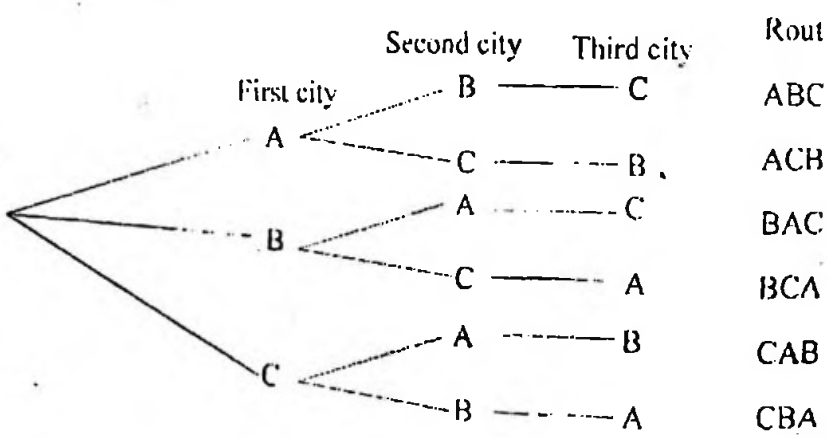
7.13 Factorial Rule

A collection of n different items can be in order of n! different ways (That is the first item may be selected in n different ways, the second item may be selected in *n -1* ways, and so on.

For instance, if a researcher intends to conduct a survey in *3* cities *A Band C* and he is looking at the shortest route to these cities. How many shortest routes can he take?

Solution:

Using factorial rule, the three different cities can be average in *3!* = 6 different shortest routes. There are *3* choices of routes for the first city to be visited. When in the first city there are only *2* different options of routes to reach the second city, and there is only an option of route left when the researcher left to reach the third city as demonstrated below



From the above diagram 3! = 6 different possible routes available for the researcher to reach the three cities.

7.14 Permutation

The six different options of routes leading through the three cities above also reflect six different arrangements of the cities in a kind of simple and adaptative sequence available to the research, and such arrangement is called permutation.

Permutation refers to the sequential arrangements which can be made by taking some or all of the numbers of things involve. Supposed that there are three things *A B and C*, they can be arranged in 6 different ways. *ABC, AC'B, BAC, BCA, CAB, CBA*, as obtained in the above diagram.

This sequential arrangement can be done in such that, if there are *n* objects, then there are also *n* ways of choosing the first item for the arrangements, then there *n 1* ways of choosing the second item and so on, until all the arrangement is done. Hence permutation is given by

$${}^nP_r = n\,(n-1)(n-2)\ldots 1 = n!$$

where ${}^nP_r$ = *number of permutations*

r = factors or products

n = number

In this case $r = n$ therefore,

$${}^nP_r = n(n-1)(n-2)\ldots 1 = n!$$

n! = n factorial. Hence in terms of factorial we write

$${}^nP_r = \frac{n!}{(n-r)!}$$

Permutation rule (*when items are all different)* if the items are all different, then the number of permutation or sequence of r items selected from n available items.

Items (not allowing repetition) are

$${}^nP_r = \frac{n!}{(n-r)!}$$

n! = n factorial

n = number of different items

r = number of different items taking at a time.

This permutation rule requires the following conditions

1. There must be a total of n different items available *(i.e this rul does not apply if some of the items are identical to others).*

2. r of n items must be selected without repetition

3. the arrangements of the same item must be considered to be different ordering of the same items are counted separately.

Example

Given the letters *A B C D E*, determine the permutation consisting of *3* letters each that can be formed.

Since $^nP_r = \frac{n!}{(n-r)!}$

Where $n = 5$

$r = 3$

$= \frac{5!}{(5-3)!}$

$= \frac{5.4.3.2.1}{2.1} = 5.4.3 = 60$

Permutation rule (when items are identical to others). If there are items *with* n_1 *alike* ... n_2*alike* ... n_k *alike*, then the number o permutation of n items is

$$^nP_r = \frac{n!}{n_1!\, n_2!\, n_k!}$$

Example: determine the different permutation of the ten letter word statistics.

The word statistics consist of $1a\ 1c\ 2i_s\ 3S_3\ 3t_s$ therefore

$$^nP_r = \frac{10!}{1!.1!.2!.3!.3!}$$

$$= \frac{10.9.8.7.6.5.4.3.2.1}{1.1.2.1.3.2.1.3.2.1}$$

$= 50,400$

7.15 Combination Rule

When we intend to select r items from n different items and we do not take order into account, then we are really concerned with possible combinations rather than permutation, such that one combination such as *ABC ACB BAC BCA CAB CBA* which constitute one combination but they are six permutation, therefore, the number of permutation is always greater than the number of combination.

The application of combination rule requires the following conditions

1. there must be a total of n different items available

2. r of n items must be selected without repetations

3. re arrangement of the same items in the same grouping must be considered. Therefore, the number of combinations of *n* different items taken *r* at a time is given by

$$^nC_r = \frac{n!}{(n-r)!\, r!}$$

Where *n = number of different items given*

r = number of different items taken at a time out of the different items given.

Example

Given that the governing council of Nassarawa State University Keffi consist of *9 members*. Each year they elect a *3* person committee to oversee appointments and promotion of staff. Each year they also elect 9 chairman, vice chairman and secretary.

a. When the council elect the appointment and promotion committee. how many different committees are possible?

a. When the council elect 3 officials (*the chairman, vice chairman and secretary*). how many different slates of candidates are possible?

Solution

In the case of committee the order is not relevant as in the case of electing the officials therefore

a. $n = 9$

$r = 3$

$$9(3) = \frac{n!}{(n-r)!\,r!} - \frac{9!}{(9-3)!\,3!}$$

$$= \frac{9.8.7.6.5.4.3.2.1}{(6.5.4.3.2.1)3.2.1}$$

$$= \frac{362.880}{4320}$$

$$= 84$$

b. The number of sequence or permutation $r = 3$ people elected from $n = 9$ available members.

$$^9P_3 = \frac{n!}{(n-r)!} = \frac{9!}{(9-3)!}$$

$$= \frac{9.8.7.6.5.4.3.2.1}{6.5.4.3.2.1}$$

$$= 9.7. \text{ or } \frac{362.88}{720}$$

$$= 504$$

In this regard. there are 84 different committees of 3 council members but there are *504* possible slate of candidate. This gives a clear picture of the difference between permutation and combination such that combination is less than permutation because of their difference in the denominators

CHAPTER EIGHT

PROBABILITY DISTRIBUTIONS

8.1 Random Variable

A random variable is a variable *(usually represented by X)* that has a single numerical value that determined by chance for each outcome of an experiment. It has a specified physical, geometrical and other significance and it takes on different values as a result of the outcome of a random experiment. It is a function which assigns numerical value to each elements of the set of events that may occur. For instance, a random variable can be the number of students' absence from lecturers in a particular week. It could be the height of students admitted into Economics department e.t.c.

E Descrete random variable is a variable that has a finite or countable number of values.

A continuous random variable is the one that has an infinite or uncountable or many values and those values can be associated with measurement on a continuous scale in such a way that there are no gap or interruptions.

For instance if a coin is tossed twice. then sample space will be *(HH HT TH TT)*

Let X = the number of heads that can come up n the sample point *HH = 2 heads, HT = 1 head, TH = 1 and TT = 0*

Sample space	*HH*	*HT*	*TH*	*TT*
X	*2*	*1*	*1*	*0*

The random variable identified above can be used to identify the probability distribution for each of the random variable. Therefore when all the values of the random variable along with their corresponding probabilities are determined, then we have a probability distribution.

8.2 Probability Distribution

Probability distribution is the listing of the probabilities of all possible

outcomes of an experiment. In other words probability distribution gives the probability for each value of random variables.

Suppose that the possible values of random variables given by X_1, X_2 ..... X_n are arranged in some order. Supposed also that these values of random variable are assumed with probability given.by

$$P(X = X_k) = f(X_k)$$

Where $k = 1,2,.....$

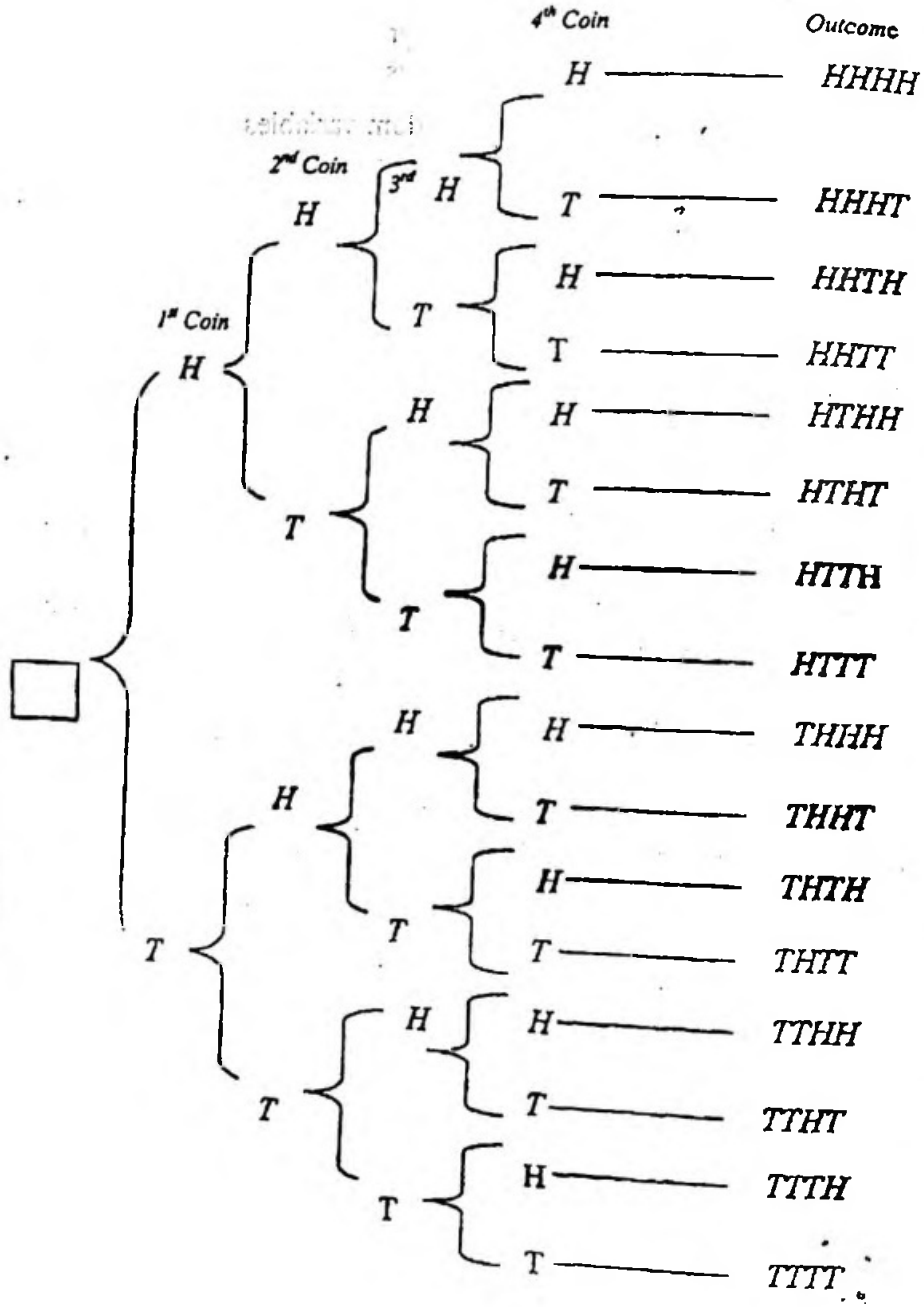
Or more conveniently given by probability function or probability distribution

$$P(X = x) = f(x)$$

Where $f(x) = 0$ *and*

$$\Sigma f(x) = 1$$

For instance, consider the tossing of four coins simultaneously. The outcome the experiment is 16 and the events are *HHHH, HHHT, HHTT, HTTT, and TTTT*. The frequencies of these events are *1,4,6,4 and 1* respectively.



The outcomes of tossing four coins

Base on the above outcomes of tossing four coins, the frequency and the probability distribution of the experiment are as follows.

Event	*Number of outcomes of event*	*Probability of occurrence of the event*
HHHH	*1*	$\frac{1}{16} = 0.0625$
HHHT	*4*	$\frac{4}{16} = 0.25$
HHTT	*6*	$\frac{6}{16} = 0.375$
HTTT	*4*	$\frac{4}{16} = 0.25$
TTTT	*1*	$\frac{1}{16} = 0.0625$
Total	*16*	*=1.00*

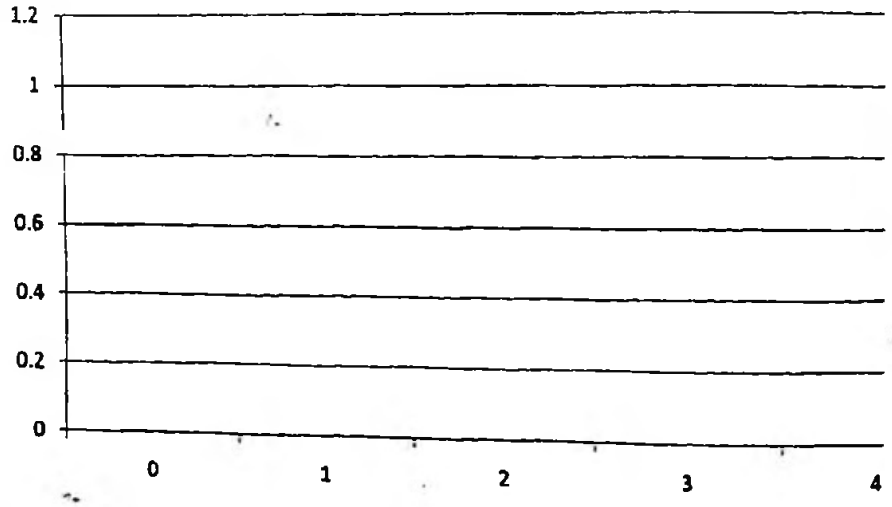
The discrete probability of tossing four coins

Let X (i.e descrete random variable) be the number of head. That occur in each of the events HHHH, HHHT, HHTT, HTTT, TTTT, we can see that:

X = 0 when the event is TTTT, (i.e the number of head is 0)
X = 1 when the event is HTTT, (i.e the number of head is 1)
X = 2 when the event is HHTT, (i.e the number of head is 2)
X = 3 when the event is HHHT, (i.e the number of head is 3)
X = 4 when the events is HHHH, (i.e the number of head is 4)

Based on the above therefore, the probability distribution is

$$P(X) = \begin{cases} \frac{1}{16} & if\ x = 0 \\ \frac{4}{16} & if\ x = 1 \\ \frac{6}{16} & if\ x = 2 \\ \frac{1}{16} & if\ x - 4 \end{cases}$$



Number of head occurred (x)

The mean, variance and standard deviation of probability distribution

(1) The mean for probability distribution is given by $\mu = \sum x \,.\, P(x)$

(2) The variance for probability distribution is given by $\theta^2 = \sum[(x - \mu)^2 \,.\, P(x)]$ and therefore, $= [\sum x^2 \,.\, P(x)] - \mu^2$

(3) the standard deviation for a probability distribution is given by

$$\theta\sqrt{[\Sigma x^2 \,.\, P(x)] - \mu^2}$$

Calculating $\mu\theta^2$ and σ for probability distribution

X	P(x)	x . P(x)x	x^2 . P(x)
0	$\frac{1}{16} = 0.0625$		
1	$\frac{4}{16} = 0.250$		
2	$\frac{6}{16} = 0.375$		
3	$\frac{4}{16} = 0.250$		
4	$\frac{1}{16} = 0.0625$		
Total	$\sum P(x)$	$\sum x \,.\, P(x)$	$\sum x^2 \,.\, P(x).$

The standard deviation is the square root of variation and hence,

$$\sigma = \sqrt{\sum x^2 \,.\, P(x) - \mu^2}$$

Where $\sum x^2 \,.\, P(x) =$

$\mu = \sum x \,.\, P(x) =$

$\mu^2 = \sum x \,.\, P(x) =$

Hence $\sum x^2 \,.\, P(x) - \mu =$

Therefore $\sigma = \sqrt{\sum x^2 \,.\, P(x) - \mu}$

8.3 *Binomial Probability Distribution*

In every experiment, there is the probability of success or failure, hence if a researcher construct as many trails of such experiment there is the probability of obtaining either more of success or less of it. For instance a manufacture may want to know the probability of having 3defects out 10 output produced or a researcher may want to know the probability of *12* success in *20* trials of an experiment. Therefore, the most appropriate probability that may be applied in these case is the Binomial Probability.

Binomial probability distribution is a descrete probability distribution which is based on Bernoulli process. It is concerned with

descrete random variables such that it describes descrete data resulting from Bernolli process,

Binomial experiment have an element of 'twoness' because it is associated with any two mutually exclusive events. For instance, consider the experiment of tossing a coin, the outcome may be either a head or a tail and hence, the sample space is *(Head, Tail)*. Therefore, the number of event is equal to *two(2)*. The occurrence of head in this experiment will prevent the occurrence of tail since they are two mutually exclusive and collectively exhaustive events. In this case the probability of occurrence of head in a trail is *0.5* while the probability of occurrence of tail is *0.5*.

In a trial tossing of a a coin between two people, the probability of success (P) to a person is 0.5 and the probability of failure *(q = 1 -P)* to the same person is *0.5*, if n repeated trials is performed then the objective of the game is to find the probability of having x success for the person and hence the experiment is termed as *Bernolli process or Binomial experiment* that n repeated trials are performed. However, this experiment must meet the following requirements:

1) The experiment must have a fixed number of trials

2) Each trial must have all outcomes classified into two categories. That is, there must be only two mutually exclusive and collectively exhaustive events.

3) In all trials the observations are independent of one another. That is, the trials must be independent; and the outcome of any trail doesn't affect the probabilities in the other trials.

4) The probabilities must remain constant for each trial.

If the requirements are achieved then we may apply the Binomial probability function give by

$$P(x \text{ success in a trails given } p) = {}^nC_x P^x q^{n-x}$$

$$\text{That is } P(x=x) = {}^nC_x P^x q^{n-x}$$

$$= \frac{n!}{(n-x)!x!} \cdot P^x \cdot q^{n-x} \quad (x=0,1,2,3)$$

where $${}^nC_x = \frac{n!}{(n-x)!x!}$$

n = number of trails

p= probability of success in a single trail

q=(1-p)=probability of failure in a single trail

X = random variable.

Binomial probability depends on n and p which are the parameters, such that the knowledge of p defines the probability of x since n is already known by the definition of the problem. Hence the value of *p* determines the shape of the binomial distribution. Therefore when *p(= 0.1)* is small the binomial distribution is skewed to the right. When *p (= 0.5)*, the binomial distribution is symmetrical. When *p* is larger than 0.5, the distribution is skewed to the left, but if *p* remain up and become much numerous to form a *bell shape*.

Determination of Binomial Experiment

To identify or determine whether an experiment is a binomial experiment or not depend on whether the experiment satisfy the condition for Bernolli process. For instance if *10%* of the students in level one are left handed, suppose we want to find the probability of getting exactly *3* left-handed students, given that there are *15* students in the class.

Here, we have to check whether this experiment satisfy the layed down conditions for application of binomial distribution formular.

(1) The number of trials *(15)* is fixed

(2) The trials are independent because left handedness and right handedness of any one student does not affect the probability of any other student being off handed.

(3) There are two categories of outcomes in each trial. That is the student is either left hand or right-handed.

(4) The probability of *0.10* remains constant for different student.

Therefore the experiment is binomial because it satisfies all condition for binomial experiment.

Now let us proceed to identify the values of *n, x, p, and q*

(1) Since there are 15 students in the class, then $n = 15$

(2) We want to find the probability of exactly 3 left-handed students. The $x = 3$

(3) The probability of a left-handed student (success) is 0.1 (i.e 10%), hence $p = 0.1$

(4) The probability of failure *(not left handed) q is equal to 1 p = 1 0.1.* Therefore $q = 1\ 0.1 = 0.9$.

To apply the binomial formula, since we have

n = 15 = number of trials

x = 3 = number of success among n trials

p = 0.1 = probability of success in any trials

q = 0.9 = probability of failure in any one trial

$$\text{Therefore, } P(x) = \frac{n!}{(n-x)!\,x!} \cdot p^x \cdot q^{n-x}$$

$$= \frac{15!}{(15-3)!\,3!} \cdot 0.1^{(3)} \cdot q^{(15-3)}$$

$$= \frac{15!}{(12)!\,(3)!} = 0.001 \cdot 02824536$$

$$= 455\ (0.001)\ (0.282429536)$$

$$= 0.129.$$

Hence, the probability that exactly *3 of the 15* students will be left handed is *0.129*.

Mean, variance, standard deviation and other measurement of binomial distribution.

The mean $\mu = n \cdot p$

The variance $\sigma^2 = n \cdot p \cdot q$ *or* $n.p(1-p)$ *where* $q = (1-p)$

The standard *variation* $= \sigma = \sqrt{variance} = \sqrt{n}.p.q$

$$Kurtosis = 3 + \frac{6p + 6p2}{n.p.q}$$

$$Skewness = \frac{1-2p}{\sqrt{n}.p.q} \text{ or } = \frac{1-2p}{\theta}$$

8.4 Poison Distribution

Poison distribution is a descrete probability distribution that is used to represent the number of occurrence of an event in one unit of time. It is a distribution that implies the occurrences of some events over a specified interval such that the random variable x is the number of occurrence of the event in an interval. This interval can be time, distance, area, volume e.t.c. This distribution is based experience, that is to say that it is based on empirical results of past experiments relating to the problem under study. Poison distribution is applied when the probability of occurrence of an event is very small that is p or 1 p is almost equal to unity while n is very large. The number of students entering the library within a minute is an example of a random variable in a poison distribution. Therefore the random variable of interest in poison distribution is the number of occurrences of a given event during a given interval.

The probability of the event occurring x time over an interval is given by

$$P(x) = \frac{\lambda x \,.\, e - \lambda}{x!} \qquad X = 0,1,2,3....$$

Where e = 2.71828 is the of the natural logarithm.

λ = *average number of occurrence per specified interval. (i.e the mean of the distribution)*

X = number of occurrences of a given event.

The poison distribution or process has the following requirement.

1. The random variable x is the number of occurrence of an event over some trial. The mean relating to a given interval can be estimated on the bases of past data concerning the variable under study.

2. The occurrence of event must be random

3. The occurrence must be independent of each other

4. The occurrence must be uniformly distributed over the interval being used.

Properties of the poison distribution

The mean $\mu = \lambda = (n.p$ *where* $p = \lambda/n$ *hence means* $\lambda/n.\ n = \lambda)$

The variance $\sigma^2 = \lambda$

Standard deviation $\sigma = ?\ \lambda$

Coefficient of skewness $\alpha_3 = 1/?\ \lambda$

Coefficient of kurtosis $\alpha_4 = 3 + (1/\lambda)$

Moment generating function $M(E) = e\,\lambda\,(e-1)$

Characteristic function $\sigma\,(w) = e\,\lambda\,(e-1)$.

Poison distribution differs from binomial distribution in the following ways

(1) Binomial distribution is affected by the sample size n, and the probability p. But poison distribution is effected only by the mean x.

(2) In a binomial, the possible value of the random variable x *are* $0, 1, \ldots$ n, but a poison distribution has possible x *value of* $0, 1, 2, \ldots$ With no upper limit.

Poison distribution is in most cases used when the value of n is not known or cannot be estimated. For instance the number of people that crosses a high traffic road may be established but it cannot be stated how many people that was not able to cross. The goals scored by a team in a football match is given, it cannot be stated how many goals could not be scored.

Example

The arrival rate of newly admitted students arriving at the school gate follows the poison distribution with a mean arrival rate of 4 per 10 minutes interval. Find the probability that

1. Exactly 0 students will arrive in 10 minutes interval
2. Exactly two students will arrive in 10 minutes interval
3. At most two students will arrive in 10 minutes interval
4. At least three students arrive in 10 minutes interval

Solution

The arrival rate = *10 minutes*

The mean arrival rate = $\lambda = 4$

(a) $P(x = 0, 4) = \dfrac{\lambda^x\, e^{-x}}{x!} = \dfrac{4^0.e^{-4}}{0!} = 0.018$

(b) $P(x = 2, 4) = \dfrac{\lambda^x\, e^{-x}}{x!} = \dfrac{4^2.e^{-4}}{2!} = 0.146$

(c) $P(x = \leq 2, 4) = \dfrac{\lambda^x\, e^{-x}}{\leq x!} = \dfrac{4^0.e^{-4}}{0!} = + \dfrac{4^1.e^{-4}}{1!} + \dfrac{4^2.e^{-4}}{2!} = 0.238$

d) $P(x = \leq 3, 4) = 1 - \sum \dfrac{\lambda^x.e^{-\lambda}}{} = 1 - \{\dfrac{4^0.e^{-4}}{0!} + \dfrac{4^1.e^{-4}}{1!} + \dfrac{4^2.e^{-4}}{2!}\}$ = 1-0.238

= 0.762

Relationship between poison distribution and binomial distribution

When n is large *(i.e. $n \geq 20$)* approaching infinity and p is small *(i.e. $P \leq 0.5$)* approaching zero poison distribution can be approximated for binomial distribution. Hence, when these two conditions are fulfilled, we can use the mean of binomial distribution in place of that of poison distribution as follows

$$P(x = \lambda) = \frac{(n.p)^x\, .\, e^{-(n.p)}}{x!}$$

Example

Given that there are 20 cars in a driving school and the probability of using a car in a day is 0.02, what is the probability that exactly 3 of the cars will be used in the same day. Use both poison and binomial distribution and compare the result.

In this case, $n \geq 20$ *and* $P \geq 0.5$, thereby satisfying the condition for approximation. Therefore the mean of binomial distribution can be used in place of that of poison distribution.

$$\text{Hence } P(X = x) = \frac{(n.p)^x \cdot e^{-(n.p)}}{x!}$$

$$\text{Therefore } P(X = 3) = \frac{(20 \times 0.2)^3 \cdot e^{-(20 \times 0.2)}}{3!}$$

$$= \frac{(0.4)3.\ (0.67032)}{3 \times 2 \times 1}$$

$$= \frac{(0.064).\ (0.67032)}{6}$$

$$= 0.00715$$

Binomial probability function

$$F(x = r) = {}^{n}C_r\ p^r\ q^{n-r}$$

Since $n = 20, r = 3, p = 0.02$ *and* $q = 1 - 0.02$

Therefore, $f(x = 3) = {}^{20}C_3 (0.02)^3 \cdot (0.98)^{17}$

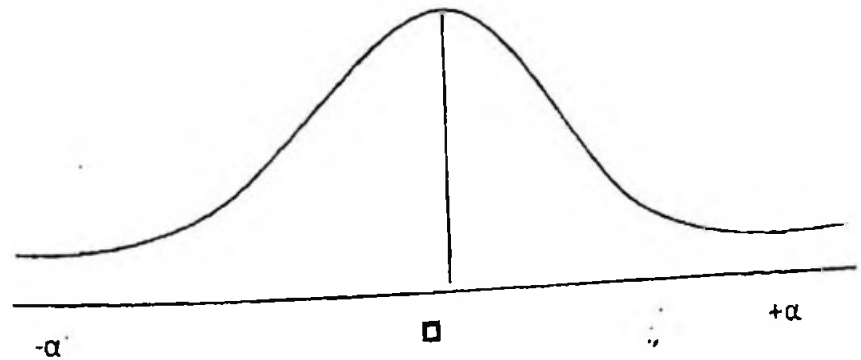
$$= 0.00650$$

Given the two results above, it is clear that the difference is not much hence; poison distribution tends to be a good approximation of binomial distribution

8.5 *Normal Distribution*

A normal distribution is a continuous probability distribution. such that a continuous random variable has a normal distribution if that distribution is symmetrical and bell-shaped. This type of probability distribution describes the probability link between statistic and a parameter that is between sample values and the population value from which the sample is drawn, and hence, it is important in the inferential aspect of the statistics. Therefore, it found application in many of the real-life situations, which are usually modeled as normal distribution. The exam scores of students in statistic examination, the salaries of lecturers of NSU Keffi follows normal distribution and therefore are cases in example of normal distribution.

Normal distribution is bell shaped such that distribution is symmetrical about its mean unlike the binomial distribution function, the normal distribution function is a density probability function.



Normal Distribution

The distribution function is given by

$P(x) = 1/\sigma ? 2\pi \exp [(x - \mu)^2]$ *where* $- \alpha ? x ? \alpha$

Where μ = *mean of the distribution*

σ^2 = *variance of the distribution*

σ= *standard deviation of the distribution*

σ= *standard deviation of the distribution*

X = *normal random variable*

Or

$P(x) = 1/\sigma? 2\pi \cdot e^{-\frac{1}{2}(x-\mu}$

The formular for cumulative density distribution of the normal distribution in the range of the random variable from $-\alpha + x$, or the corresponding density probability function is given as

$$f(x) = \int_{-\alpha}^{x} \frac{1}{\sigma\sqrt{2\pi}} exp\left[\left(\frac{x-\mu}{2\sigma^2}\right)^2\right] dx$$

This is the same thing as

$$f(x) = P(x ? x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\alpha}^{x} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The most important parameter of interest in normal distribution are the μ (mean) and σ*(standard deviation)* and it is impossible to provide separate table of values of the normal distribution for different combinations of μ and σ because the number of such combinations is not finite. Hence, there is the need for a standard (special) distribution that can (facilitate) help to use the normal distribution. Therefore the standard normal distribution which has zero mean and a constant variance *(i.e. mean – 0 and variance = 1)*, is arrived at to serve as a form of standardization of the normal distribution consistent with the observation of real – life situation.

If the observation of real – life situation follows normal distribution with mean μ and variance σ^2 then its random variables can be converted into a standard normal random variable using the following transformation

$z = x - \mu$

Where Z = standard normal variable. Hence, the corresponding distribution is called the standard normal distribution which is given as

$P(Z) \frac{1}{\sqrt{2\pi}} exp[-1/2z^2]$ *where* $-\alpha < z < \alpha$ *with mean* $= 0$ *and*

variance $= 1$

Properties of Normal Distribution

Mean $= \mu$

Variance $= \sigma^2$

Standard Deviation $= \sigma$

Coefficient of Skewness $= \alpha_3 = 0$

Coefficient of Kurtosis $= \alpha_4 = 3$

Given a sample of *200* students, their statistics test scores follow a normal distribution with mean $= 10$ and standard deviation $\sigma = 2$. Using *0.05* level of significance:

(a) Determine the probability of scoring less than *11* marks in the test and find the number of students that score less than *11* marks.

(b) Determine the probability of scoring more than *12* marks and find the number of students with more than *12* marks.

(c) Determine the probability of scoring between *7 and 11.2* marks in the test and find the number of those that score marks between *7 to 11.2* marks.

Solution

(a) P (x? 11) = $p\left[\frac{x-\mu}{\sigma} \leq \frac{11-10}{2}\right]$

$= P(z ? 0.5)$

$= 0.6915$

Hence the probability of scoring less than 11 marks is 0.6915 and the number of students having less than *11* marks is $2 X 0.6915 = 138$.

(b) P (x? 12) = $p\left[\frac{x-\mu}{\sigma} \cdot \frac{12-10}{2}\right]$

$= P (z ? 1)$

$= 1 - P(z ? 1)$

$= 1 - 0.8413$

$= 0.1587$

Hence, the probability of scoring more than *12* marks is *0.1587* and the number of students that score more than *12* marks is $2 X 0.1587 = 31.74 ? 32$ students.

(c)$P (7 ? x ? 11.2) = P(x ? 11.2) - P(x ? 7)$

$=p\left[\left(\frac{x-\mu}{\sigma} \leq \frac{11.2-10}{2}\right) - p\left(\frac{x-\mu}{\sigma} \leq \frac{7-10}{2}\right)\right]$

$= P (z ? 0.6) - P(z ? -1.5)$

$= 0.7257 - 1 + 0.9332$

$=0.659$

The probability of scoring between 7 marks and *11.2* marks is *0.659* and the number of students that score between 7 and 11.2 marks are $2 x 0.659 = 131.8 ? 132$ students locating the z scores when given the probability.

The z standard normal probability table contains both the z scores and the area (probabilities). Therefore one can locate the z values when the area is given. However, it is easier to locate the area using the z values.

If the area or probability is given, that is to locate the z score one should be mindful to avoid confusion between z value and area or probabilities. It is important to note that the z values lies at the extreme vertical left and the horizontal top most line. That is the first line. That is the first line at the extreme left of the table and the extreme top of the table apart from which, the entire body of the table and the areas or the probabilities. Therefore, if the probability is given, to locate the z values we may use the following steps.

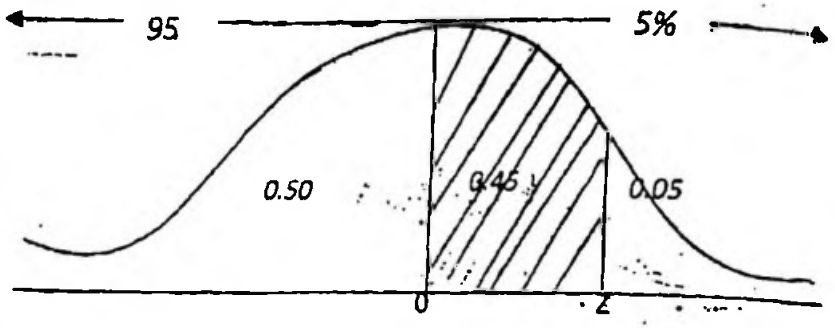
1. Identify the probability representing an area bounded by the center line; this may be the original probability, or it may be found by using the original probability.

2. Using the probability representing the area bounded by the center line, locate the closest probability on the z table and identify the corresponding z values.

3. If z values are positioned to the left of the center line, the probability is negative.

Example

Given the temperature reading that is normally distributed with mean $\mu = 0^0c$ and standard deviation $\sigma = 1^0c$. find the temperature separating the bottom *95%* from the top *5%* i.e. the area below



Standard normal distribution

Where $\mu = 0$

(i.e. the 95 percentile separating the top 5% from the bottom 95%).

Locating probability when given z value

To locate probability when given z value, there are some important things we have to keep in mind.

(1) Avoid confusion between z values and areas or probability. Remember that the z values lies on extreme left and across the top of the z table, and they are distances along the horizontal scale. Whereas the number on the body of the table are the areas or probabilities..

(2) z values to the left of the center line of the normal curve are always negative.

To locate the z value, the following steps are important

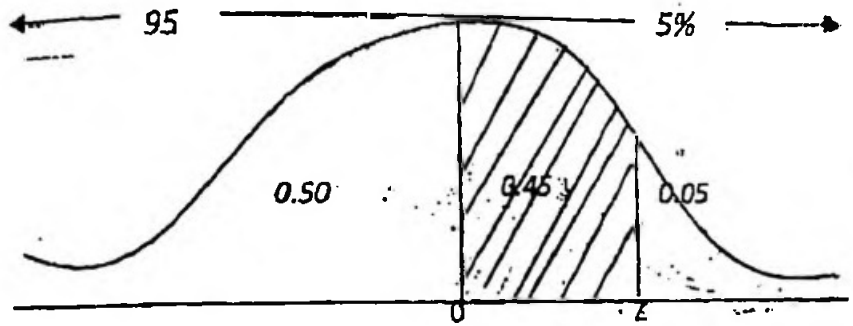
1. Identify the probability representing an area bounded by the center line; this may be the original probability or it may be found by using the original probability.
2. Locate the closest probability on the z table using the probability representing the area bounded by the center line and identify the corresponding z value.

1. Mark the z value negative if it lies on the left hand of the center line on the normal curve.

Example

Given the thermometer with temperature readings that are normally distributed with a mean $\mu = 0^{\circ}c$ and standard deviation $\square = 1^{\circ}c$. find the temperature corresponding to (P_{95}) 95 percentile, that is, find the temperature reading separating the bottom 95% from the top *5%. (Triola 1977).*

Solution



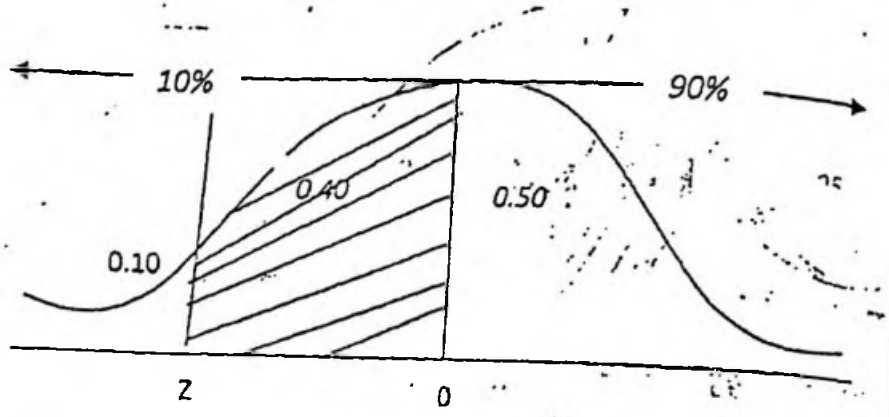
Sing the region bounded by the center line where $\mu = 0$. on one side, such as shaded region of *0.45*. We first locate the area 0.45 on the body of the z table and find the corresponding z value.

On the z table the area *0.45* lies between *0.449* and *0.4505*. Now if we sum the two values together and divide by two it is equal to *0.4500* and this area correspond to *1.645* z value and therefore, the z value is equal to *1.645*.. Hence *95* percentile is the temperature reading o

f $1.645^{\circ}c$. When tested at freezing, *95%* of the reading will be less than or equal to $1.645^{\circ}c$ and *5%* of them will be greater than or equal to $1.645^{\circ}c$.

Example

Using the same thermometer, find the temperature reading separating the bottom *10%* of all temperature from the top *90% (i.e. find the* 10_{th}, 10^{th} *percentile*).



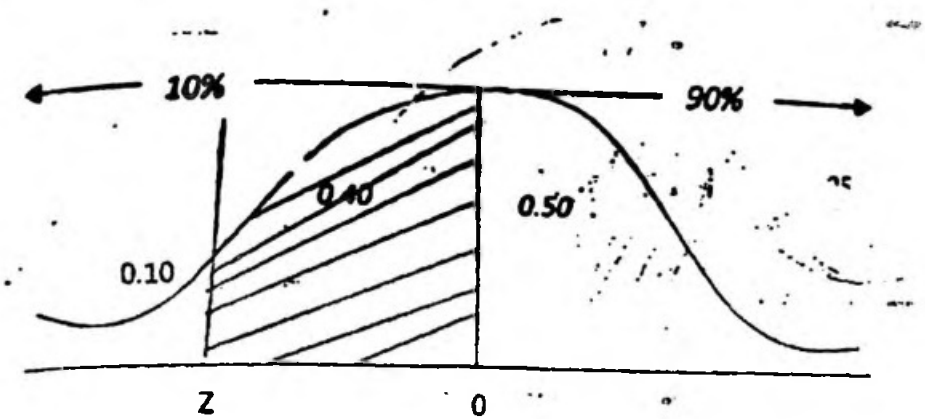
Since the z table is designed for areas bounded by the center line, the shaded area *0.40* (corresponding to *50% - 10%)*, on the body of the table, the closest value is *0.3997*. therefore the corresponding z value is *= 1.28.*

Since the z value lies below the mean of 0 then, it must be negative. Therefore, the *10th* percentile is *-1.28°c*. When tested at freezing *10%* of the thermometer reading will be equal to or less than *-1.28°c*.

Using the center line, *μ = 0* the shaded region is *0.45* since from the *μ = 0* center line to the right is equal to *0.50*. Hence, taking away *5%* it remains *0.45*. Therefore on the z table, the area of *0.45* lies the value of *0.4495* and *0.4505* putting them together divided by *2*, we have *0.4500* corresponding to z value of *1.645*, therefore, the *95* percentile is the temperature reading of *1.645°c*.

Example

Find the temperature reading separating the bottom *10%* of all temperatures from top *90%*



On the z table the area *0.40* correspond to the closest value of *0.3997* and it correspond to *z = 1.28*. Therefore, the temperature reading will be equal to or less than *-1.28°c*.

Locating z value when given the probability (non standard normal distribution)

There are three important things to keep in mind:-

1. Do not confuse z values and areas or probability. Remember that z values are distances along the horizontal scale. In the z table the z values are in the left column and across the top row but areas are found in the body of the table.
2. Choose the correct *(right/left)* side of the graph. A z value separating the top *10%* from the other will be located on the right side of the graph. but a z value separating the bottom 10% will be located on the left side of the graph.
3. A z value must be negative whenever it is located to the left of the center line of 0.

To locate the z value the following steps are important.

1. Draw a bell shaped normal distribution curve and enter the given probability (percentage) in the appropriate region of the graph and identify the x values being sought

2.. Find the z value (on the z table) corresponding to the region bounded by x and the center line of 0.

One can do this by (1) finding the closest area and then identify the corresponding z value. (2) Making z value negative if it is located to the left of the center line

2. Use the z formula by entering the values of μ, , and the z value found in step 2. then solve for x. as follows

3. Use the z formula by entering the values of μ, σ , and the z value found in step 2, then solve for x. as follows

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

Then $x = \mu + (z.\sigma)$

4 Verify *(on the bases of the sketch of the normal curve)* that the solution make sense in the context of the graph and in the context of the problem.

8.6 *Non staudard Distribution*

In theory, we assumed that a normal distribution or a normal population has zero mean and (1) constant variance, however most normally distributed population have a non zero mean and non constant variance (i.e. a standardization different from one).

Finding probabilities for non standard distribution, the nonstandard normal distribution can be standardized using the formula.

$$Z = \frac{x - \mu}{\sigma}$$

Once –score, the researcher can use the standard z table to locate the probabilities. The following procedures are important when finding probabilities.

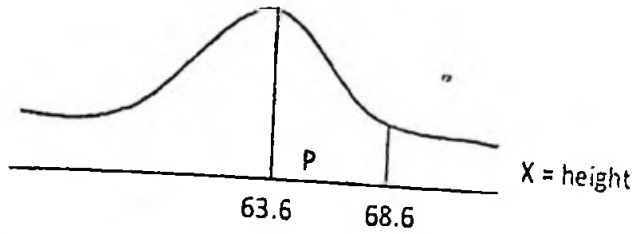
1. Draw a well labeled normal curve with the scores and then shade the regions representing the desired probability.
2. For each relevant score x that is bounced by the shaded region use the z-formula above to find the equivalent z-score.
3. Then find the area of the shaded region on the standard z table and

therefore the area become or is the desired probability.

Example

Given a normally distributed height of students with a mean of *63.6* inches, and a standard deviation of *2.5* inches, find the probability that a randomly selected students height is between *63.6* inches and 68.6 inches.

Step (1) given $z = \frac{x - \mu}{\sigma}$



Step (2) $z = \frac{x - \mu}{\sigma} = \frac{63.6 - 63.6}{2.5}$

$= 2.00$

Step (3) find that $z = 2.00$ *correspond to an area of 0.4772 on the z-table.*

The probability can be express thus $P\ (63.6 < x < 68.6) = P(0 < z < 2.00) = 0.4772$

That is the probability of selecting a student with the height between *63.6* inches and *68.6 inch is 0.4772.*

In other words, the researcher may conclude that *47.72* percent of students have heights between *63.* and *68.6* inches.

Finding z-score for nonstandard distribution.

To find the z score the researcher must keep the following in mind.

(1.) Z-score and area must be clearly defined and understood, note that z-scores are distances along the horizontal scale areas represent regions under the normal curve and are found on the body of the normal curve.

(1.) The appropriate side of the curve must be chosen, note that score at the top *10%* can be located at the right while bottom 10% is located at the left hand side of the curve (i.e. from zero).

(2.) Note also that from zero to the left of the curve is negative and vice versa.

The following steps are important in finding z-score.

1. A sketch of a bell shaped normal curve is important indicating the probabilities (percentage) in the appropriate region of the graph, and indicating the x value being sought.

2. Find the z-score corresponding to the region bounded by x and the center line of 0.

3. Using the z formula, enter the values of μ , σ and the z-score to solve for x as follows:

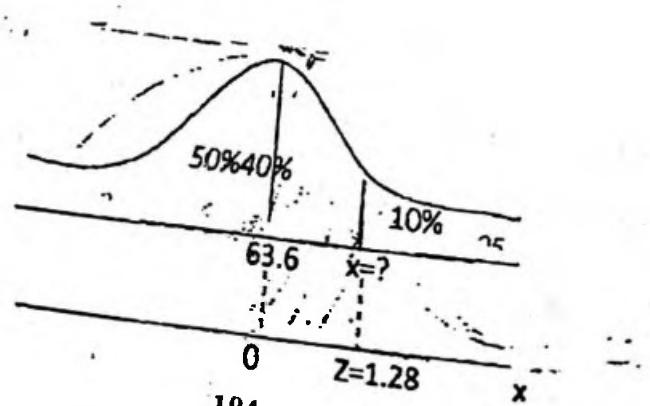
Since $z = \frac{x-\mu}{\sigma}$

Therefore, $x = \mu + (z.\sigma)$

4 Then check whether the solution makes sence in the context of the graph and the problem as well

Example, Given a normally distributed height of students with mean = *63.6* inches and standard deviation = 2.5 inches, find the height separating the bottom *90%* from the top *10%*

Step (1)



The area from the bottom to the top 10% is 90% therefore, *90% - 50% = 40% = 0.4000.* Hence we look out for the area *0.4000* in the body of the table. And if is closer to *0.3997* on the table. This area corresponds to z-score of *1.28.*

Step (2) $z = 1.28$

$\mu = 63.6$

$\sigma = 2.5$

$x = ??$

And $z = \frac{x-\mu}{\sigma} = 1.28 = \frac{x-63.6}{2.5}$

To solve for x therefore,

$(1.28).(2.5) = x - 63.6$

$(1.28.2.5) + 63.6 = x$

$X = 66.8$

Step *(3)* = the solution is reasonable because it is greater than *63.6.*

8.7 Central Limit theorem

The central limit theorem states that as the sample size n increase, the sampling distribution of the sample means approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. Suppose we take a sample from any population and find the mean $\bar{x}$ with particular characteristics, we draw another sample from the same population and find the mean x_2. If we draw many of such samples from that same population and find their means to be $\bar{x_1}, \bar{x_2}, \bar{x_3}, \ldots\ldots \bar{x_n}$. The central limit theorem is important in the following respects

1. It enables the researcher to use a population that is not normally distributed so long as the sample size is large. That is, it enables the researcher to approximate the distribution of the sample mean by normal distribution.

2. It enables the students' distribution as the sample size gets larger, such that it gives the t-test its robustness. That is to say that lack of normality of data does not affect the character of t-test.

Therefore, given that the random variable 'x' has a distribution which is normal or not normal, with mean μ and standard deviation σ, the sample size u are randomly selected from this population, then

1. The distribution of the sample mean $\bar{x}$ will, as the sample size increase approach normal distribution.

2. The mean of the sample means will be the population mean μ, While

3. The standard deviation of the sample means will be $\sigma/\sqrt{n}$.

Hence, for sample size n larger than 30, the distribution of the sample mean can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size n becomes larger.

If the original population is itself normally distributed, then the sample means will be normally distributed for any sample size n.

The mean of the sample means is denoted by μ_x μ while the standard deviation of the means is denoted by $\sigma_{\bar{x}}$ σ/?n standard error of the means.

Example, Given that the population of married women has a normally distributed weight, with the mean of *173 1b* and a standard deviation of 30 1b, find the probability that

a. If one woman is randomly selected, her weight is greater than *180 1b*

b. If 36 different women are randomly selected, their mean weight is greater than *180 1b*.

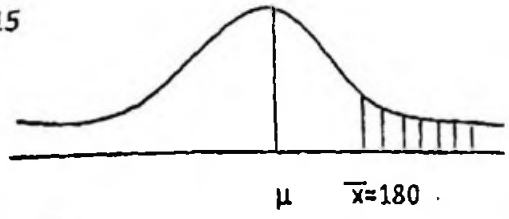
Solution (a)

The area A on the z table is equal to *0.0910*. Since area *A* and area *B* = *0.5 then* $0.5 - A = B = 0.5 - 0.0910 = 0.4090$. Hence the probability of one of the married woman weighing more than *180 is 0.4090*.

(b) Since we are dealing with the means of group of 36 women, we are dealing with the distribution of sample means then it is important to use μ_x *and* σ_x instead of μ *and* σ. And hence applying the central limit theorem

$\mu_{\bar{x}} = \mu = 173$

$\sigma_{\bar{x}} = \sigma/?n = 30/?36 = 15$



$\mu = 30$

$\sigma_x = \sigma/?n = 5$

Therefore to determine the shaded area

$$Z = \frac{\bar{x} - \mu}{\sigma_x} = \frac{180-173}{30/36} = 7/5 = 1.40$$

Z = 1.40 corresponds to an area of 0.4192 on the z-table.

The shaded region = *0.5 0.4192 = 0.0808* , therefore the probability that 36 married women have a mean weight greater than *180 1b is 0.0808* . Note that the probability of one of the women to weight more than *180 1b is 0.4090*, but the probability of 36 women to weigh more than *180 1b is 0.0808*. This indicates that it is easier for an individual sample to deviate from the mean than it is for a group. Hence as the sample gets larger the distribution of the mean becomes approximately normal.

Central limit theorem is important in the following respects:

1. It enables the researcher to use a population that is not normally distributed so long as the sample size is large to enable the researcher to approximate the distribution of the sample mean by normal distribution.

2. It enables the student t distribution to be approximated to normal distribution that it gives the t-test its robustness. That is to say tha the lack of normality of data does not affect the character of the t test.

Therefore, given that the random variable x has a distribution which is
And the sample size n are randomly selected from this population, ther (1) the distribution of the sample mean x will, as the sample size increase approach normal distribution. (2) the mean of th e sample

of the sample mean will be /n.

Hence, for sample size n larger than 30, the distribution of the sample mean can be approximated reasonably well by a normal

distribution. The approximation gets better as the sample size n becomes larger.

If the original population is itself normally distributed, then, the sample means will be normally distributed for any sample size n.

The mean of the sample mean is denoted by $\mu_x = \mu$ while the standard deviation of the mean is denoted by $\sigma_x = \sigma/?n$ = standard error of the mean.

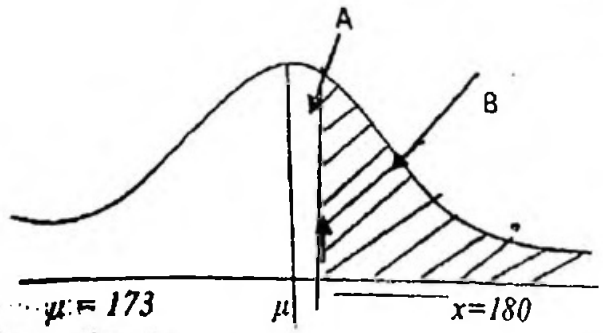
Example:

Given that the population of married women has a normally distributed weights with mean of 173lb and a standard deviation of 30lb from the probability that

(a) If one woman is randomly selected, her weight is greater than 180lb

(b) If 36 different women are randomly selected, their mean weight is greater than 180lb

Solution



$\mu = 173$

$\sigma = 30$

$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{180 - 173}{30} = 0.23$

The area A on the *Z-table* is equal to *0.0910*. Since area A and area B = *0.5 then 0.5* $A = B = 0.05$ *0.0910 = 0.4090*. Hence, the probability of one of the married women weighing more than *180lb is 0.4090*.

CHAPTER NINE

SAMPLING THEORY

9.1 Population, Sample Size and Sample Design

If the data or information requirement of a particular research work is to be sourced by the researcher, then the researcher must consider the universe or the population size. A population is the number of elements under study. In practice the researcher cannot enumerate all the items in the population since the study bias will be larger, time will be wasted, and it will involve a lot of money.

In this light therefore, the research may resort to a sample of such population in question which must necessarily be a representative of the population. A sample is a subset of a population. Algebraically if the population size is N and the sample size is n which is < N of the population. Samples can be classified into probability random sample and non probability or purposive samples.

Probability random samples are those samples which are selected in such a way that every element from the population has an equal chance of being selected and the probability of error is known while purposive samples or non probability sample are those samples which are based on the choice of the researcher.

9.2 Rationale for sampling

1) **Convenience:** - In most cases, the population becomes too large for the researcher to study, therefore the researcher may have to reduce it to a sizeable number that he can handle efficiently. This will give certain level of convenience to the researcher and spur him to put all energies and materials together to achieve greater research result.

2) **Representation:** - The idea of sample is that each segment of the population is adequately represented such that the result of the research will reflect the attribute views and the aspiration of the entire population. Apart from this, the information can be obtained from each segment and hence the research work will cover greater scope and the result will give sufficient information about the population under study.

3) **Less Cost:** - The cost of studying the entire population may be beyond the researcher's capability, however, if a sample is drawn, the cost of carrying out the study is greatly reduced and hence the stress involved in sourcing enormous fund for the study is reduced.

4) **Time:** - Time is also an important aspect of a research work. In as much as sample ensures convenience, it also ensures speed and hence saves time taken in the collection and management of information and of course, this enables the researcher to obtain quicker and reasonable result.

5) **Accuracy:** - The fewer the data collected the more accurate the management of such information. Therefore sample ensures accuracy in the collection and analysis of research information.

9.3 Sampling Design

The researcher should be able to prepare a sampling design which is a plan on how a sample should be drawn or selected. It specifies the procedure to be used in collecting samples. A good sampling design should contain the following:

1. **The type of universe or population size**
   The researcher must be able to look at the nature of population from which the sample can be drawn, whether it is finite or infinite. city workers or village settlement.

2. **Sampling Unit**
   The researcher must select his sampling units which will suite the purpose of his research work. Sampling units take the form of (a) Geographical unit such as state, district, town etc (b) constructing unit such as house, flit, etc (c) social units such as the family, club, school, etc (d) the individual.

3. **Sampling frame or source list**
   The next step in sampling design is to develop the name or a list of all items of the entire population or universe from which the sample is to be selected. It is important that such sample frame or source list should be as representative of the population as possible.

4. **Size of the sample**

The researcher is expected to select an optimum sample size. such that is not too small and is not too large to produce some kinds of bias in the research work. An optimum sample is the one which fulfill the requirement efficiency. representativeness. reliability and flexibility. The researcher should note that a small randomly selected sample is superior to a large but badly selected sample. Hence a sample selection technique is very important in this regard it play an important part in the determination of the size of sample and of course optimum sample size.

a. Homogeneity of the population or universe: if the population of study is homogeneous then a small sample size can serve the purpose of representativeness of the population and therefore became the optimum sample size. The following can be considered in the selection of sample size.

B. Number of classes proposed: if many class groups are formed then the larger will be the sample required since smaller sample might not be able to give a reasonable number of items in each class group.

9.4 The Determination of a Sample Size

The determination of the sample size is as important as the research itself. It is in this light that scholars were so emphatic about what determines the sample size; some argued that the size of the population is the bases of the determination, but some opined that the sampling technique can do the magic, since it can help in the determination of an optimum sample size.

Two approaches are prominent in this regard, the first approach is to specify the precision of estimation desired and the second approach uses the Bayesian statistic "to weigh the cost of additional information against the expected value of additional information" (in Cothari, 2009).

Sample size by the first approach which is commonly used in practice, by % suppose a sample random sample is to be drawn from a finite population of size N in order to estimate the percentage "P" in the population, we may state on the bases of our experience that we want the estimate of "P" i.e "P" to be within ? of the true value of "P"

with probability i.e pr (|p-p|<) < where P2 is the estimate of "p", then the sample size is given by the following formula.

$$n = \frac{Z\alpha_2 N? (1 - ?)}{(N-1)\ \theta^2 + Z\alpha_2\ p(1-?)}$$

Where $Z\alpha^2$ represents that value such that the probability of standard normal variable exceeding it is $(1-\alpha)/2$. The value for a chosen a level can be obtained directly from the table given Z value for the standard normal distribution. But what value to be assigned to *"p"* in the above formula is a formidable problem for *"p"* is the unknown value we are actually trying to estimate. One method may be to take value of *"p"* = 0.5 conservatively in which case "n" will be maximum and the sample will yield at least the desired precision. The other method may be to make an initial estimate of *"p"* which may either be based on personal judgment or may be the result of a pilot study. (It has been observed by many that a pilot study of 225 or more items may result in a reasonable approximation of *"p"* value for the purposes. *(Kothari 2009).*

In the case of infinite population the above formula is reduced to

$$n = \frac{Z\alpha^2\ "p"\ (1-?)}{\sigma^2}$$

Example 9.1

What should be the size of the sample if a simple random sample from a population of *200* items is to be drawn to estimate the percent effective within 2% of the true value with 95% probability? What would be the size of sample if the population happens to be infinite in the given case?

Solution

In this question we have been given

$N = 2000$

$\alpha = 2\% = ?$

$\alpha = 95\%$ corresponding to which z_α value from the table value of standard normal variable is *1.96*

Let the value of *"p" = 20% i.e. 0.02*

This may be on the bases of experience or the basis of pilot study. The sample size for a population of 2000 item in the given case would then be worked out as follows

$$n = \frac{Z\alpha^2 N? (1-?)}{(N-1)\sigma^{?} \div 2\sigma^{?} p (1 - ?)}$$

$$= \frac{(1.96)^2 (200). (0.02) (1 - 0.02)}{(200-1) (0.02)^2 + (1.96)^2 (0.02) (1-0.02)}$$

$$= \frac{(3.8416) (2000) (0.02) (0.98)}{(1999) (0.0004) + (3.8426) (0.02) (0.98)}$$

$$= \frac{(3.8416) (39.2)}{(0.7996) (0.753)} = \frac{150.59}{0.8749}$$

$$= 172$$

But if the population happens to be infinite then the sample size will be as follows

$$n = \frac{^{2} p(1- \ )}{^{2}}$$

$$= \frac{(1.96)^2 (0.02) (0.98)}{(0.02)^2}$$

$$= \frac{(3.8416) (0.02) (0.98)}{(0.0004)}$$

$$= \frac{(0.0753)}{(0.0004)}$$

$$= 188$$

Sample Size by Average

If we desire to estimate average μ, the value lies within 8^2 of μ with probability α (i.epr $(1x - \mu 1 < \sigma)$, the sample size n is given by the formula

$$n = \frac{Z\alpha^2 N \sigma^2}{(N-1)\sigma^2 + Z\alpha^2 \sigma^2}$$

If the population happens to be infinite the formula for determining the sample size is reduced to

$$n = Z\alpha^2$$

Determine the size of the simple random sample for estimating the true weight of the cereal container for the population with N = 5000 on the bases of the following information.

1. The variable of weight = 4 ounces on the bases of pas records;
2. Estimate should be within 0.4 ounces of the true average weight will 99% probability.

What would be the size of the sample if infinite population is assumed in the given case *(Cathori, 2009).*

Solution

The information given are

$N = 5000$

$\sigma^2 = 4$ *ounces*

$\sigma^2 = 0.4$ *ounces*

However, a reasonable size of sample can minimize the error in making decision. Hence it is advisable to determine the sample size for a desired accuracy base on population variance.

Sample Size Determination Using Sample Mean

If populations mean μ is unknown and variance σ^2 is known, if a

Sample Size Determination Using Sample Mean

If populations mean μ is unknown and variance σ^2 is known, if a sample size a ($n > 30$) is selected from the population, then the confident interval is estimated using the sample mean at the desired significance level (α) then

$$X - n\left(\mu \frac{\sigma^2}{n}\right)$$

The confidence interval

$$\mu = x \pm Z_{\alpha/2} \frac{\sigma^2}{\sqrt{n}}$$

If the required accuracy about the mean is ± D, then the confident interval is obtained

Confident interval of $\mu = x \pm D$

Therefore $Z\alpha/2 \frac{\sigma}{\sqrt{n}} = D$

From the above expression, the formula for sample size n via this methods

$$n = \frac{[Z_{\alpha/2}\ \sigma^2]}{D}$$

Panneerselvam 2010.

Example 9.3

The axial length of pistons followed a normal distribution with the variance of 64mm. Determine the sample size such that the mean length of pistons is within plus or minus 2mm with a confidence level of 95%. (Panneerselvam, 2010).

Solution

$\sigma^2 = 64$

$\sigma = 8$

Confidence level $(1 - \alpha) = 0.95$ **therefore**

$\alpha = 0.05$

$\alpha/2 = 0.025$ *then.*

$Z_{\alpha/2} = 1.96$ and the desired accuracy about the mean

$D = 2$

Hence $n = \frac{[Z_{\alpha/2}\ \sigma]^2}{D} = \frac{[1.96 \times 8]^2}{2} = 61.5 = 62$

Sample size determine by sample proportion.

Given that the mean = p and

$$\sigma^2 = \sqrt{\frac{\rho\ (1 - p)}{n}} = \sqrt{\frac{pq}{n}}$$

then, $p \sim \frac{(p.\ pq)}{n}$

And the confidence interval $= p \pm Z_{\alpha/2} \sqrt{\frac{pq}{n}}$

If the required accuracy about the proportion is ± D then the confident interval of proportion $= p \pm D$. Therefore,

$$Z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

From the above expression, the formula for the sample size n in

$$n = \frac{Z_{\alpha}^2{}_{/2} pq}{D}$$

Example 9.4

The manager of a bank feels that 35% of branches will have entranced yearly collection of deposit after the introduction of like in interest rate. Determine the sample size such that the mean proportion is within plus or minus 0.06 at confidence level of *90%.*

Solution

Given $p = 0.35$

$q = (1 - p) = 0.65$

and the confidence level $(1 - \alpha) = 090$

Therefore $\alpha = 0.10$

and $\alpha/2 = 0.05$

then, $Z_{\alpha/2} = 1.64$

for the desired accuracy $D = 0.06$

hence $n = \frac{Z_{\alpha}^2 pq}{D^2} = \frac{(1.64)^2 (0.35)(0.65)}{0.06^2}$

$= 170$

9.5 Sampling Techniques

The researcher should be able to choose an appropriate sampling technique or method that is compatible with the problem at hand. Sampling techniques can be classified into probability sampling and non probability sampling. By probability sampling we mean that each in the population have equal chance of being selected. This type of sampling is based on the judgment and the convenience of the researcher.

Probability Sampling Techniques

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multistage sampling

Simple Random Sampling:- this type of sampling can be classify into simple random sampling with replacement and simple random without replacement.

Simple Random Sampling with Replacement:- this type of sampling

provide room for equal probability for each unit of the population such that each unit has the same equal chance of be selected. and the probability of selecting each unit of the population is given by 1/N, this shows that each of the remaining units required for the sample is selected in the same way as the first unit is selected

Steps in the Selection of a Simple Random Sample with Replacement

If a random number containing four (4) digits is placed in a rectangular format with rows and columns, the steps of simple random sampling with replacement are as follows

1. If N is three-digit number say *400* representing the size of the population, then the units of the population are numbered as *1,2,3,4,...., 400* (i.e. the entire sample frame.
2. After that fix the initial position in the random number table.
3. Then read the first three-digit of that number and select the unit from the population with that serial number as the first unit of the sample. Later replace that unit in the population.
4. Then decide the direction of the movement for selecting the next number either to the right or below with reference to the current position (initial position) in the random number table.
5. Accordingly select the first three-digit (as you have done for the first item selected) of the next number in the random number table.
6. Then select the unit from the population with the corresponding serial number as the second unit of the sample. Later replace that unit in the population.
7. Repeat step 4-6 until the required number of units of the sample are selected from the population. (Panneersalvam. 2010).

Note that if the number at the extreme right of a row or at the bottom most of a column is replaced, then switch over to the left of next row or top of the next column.

Simple random sampling without replacements. here. each unit of the population has different (probability) chance of being selected as a unit of the sample.

Therefore the probability of selecting the first unit $= 1/N$

The probability of selecting the second unit $= \frac{1}{N-1}$

The probability of selecting the third unit $= \frac{1}{N \quad (n-1) \text{ etc}}$

In this type of sampling technique . a sample is selected from the population using monte-carlo simulation.

The step of which are as follows

1. Find the probability 1/N of 10 for instance, then $1/N = 0.1$
2. Find the cumulative probability of the N interval by summing the successive 1/N the detail of the monte-carlo simulation are shown below

Unit of Population	*Probability*	*Cumulative Probability*	*Cumulative Interval*
1	0.1	0.1	0.0-09
2	0.1	0.2	0.10-0.19
3	0.1	0.3	0.20-0.29
4	0.1	0.4	0.30-0.39
5	0.1	0.5	0.40-0.49
6.	0.1	0.6	0.50-0.59
7	0.1	0.7	0.60-0.59
8	0.1	0.8	0.70-0.69
9	0.1	0.9	0.80-0.79
10	0.1	1.0	0.90-0.99

(Paneerselvam)

3. Find the two digit random number from the random number table and check the range in which it lies in the above table.
4. Select the corresponding unit of the population and include it as the first unit of the sample.
5. Find the probability $1/(N-1)$ and repeat steps 2-4 to select the second unit of the sample from the population.
6. In the same manner, repeat step 2-4 until all the remaining units starting from the third unit till the unit of the sample are selected from the population using the probabilities $1/(N-2)$, ..., $1/(N-(n-1))$ *respectively.*

Systematic Sampling

In this type of sampling. the first unit of the sample is selected using the simple random sampling and the remaining units of the sample are selected at fixed interval n. Where

N= sample size

For instance, if the size of the population $N = 600$ and the sample size $n = 30$. The units of sample frame are divided into n number of intervals based on the ratio N/n as follows

Sampling interval width $1 = \frac{N}{n} = \frac{600}{300} = 20$

the sampling frame which consist of units of serial numbers from 1 - 600 is no w divided into 30 intervals *e.g. 1-20,... 58-60* and hence, the total number of intervals will equal to the desired sample size *30.*

The first unit of the desired sample will be selected randomly from the first range i.e *1-20* and then the rest are selected b y adding the interval width to the unit selected.

For instance if the 9th unit is selected within the range *1-20* then the next unit of the sample will be the 29th in the second range and hence 20 is then added to 29 to get the 3rd unit of the sample and so on, until the desired sample size n = 30 is realised.

Therefore if the first unit of the sample is selected, it is easier to select

the remaining units of the desired sample size.

Stratified Sampling

In this type of sampling the entire universe or population is divided into a specific set of strata such that each of the strata have a unique characteristic e.g the city. town and village. Each of which has a special characteristics. This type of sampling can be proportional if the same units of sample are allocated to each stratum which is called disproportional stratified sampling if the stratums are allocated different units of samples to be selected.

Proportional stratified sampling. If there is no much difference between the members of stratums. it is good to use proportional stratified sampling

If N = population size and n = total sample size. then Ni = size stratum i

And, ni = sample size of stratum i.

Then, ni is divided based on the following relationship.

$$\frac{N_1}{n_1} = \frac{n_2}{n_2} = \frac{n_3}{n_3} = \ldots = \frac{nk}{nk} = \frac{n}{n}$$

$$And\ N_1 + N_2 + N_3 + \ldots Nk = N$$

$$n_1 + n_2 + n_3 + \ldots nk = n$$

Example 9.5

A survey is conducted to analyze the status of employment of the recent batch of students of a premier university. To conduct stratified sampling for this study. the population which consist of different colleges is divided into three strata, viz., government colleges, aided colleges and self financing colleges since the regulation of ALCTE and the respective universities ensure. uniform standard of infrastructural and educational standard, proportional stratified sampling is used for this study. The total number of engineering colleges in the university is 200. The number of engineering colleges in the three categories viz., government, aided, and self financed are 20, 50 and 130, respectively. If the sample size is 20, determine the number of colleges to be sampled from each category. (Panneerselvam, 2010).

The number of engineering colleges in the university N =200

Number of government engineering colleges $N_1 = 20$

Number of aided engineering colleges $N_2 = 50$

Number of self-financing engineering college $N_3 = 130$

Total sample size n = 20

The relationship of proportional stratified sampling is as follows

$$\frac{N_1}{N_1} = \frac{N_2}{N_2} = \frac{N_3}{N_3} = \frac{n}{N}$$

Where

n_1 = number of colleges to be sampled from government category

n_2 = number of colleges to be sampled from aided category

n_3 = number of colleges to be sampled from self-financed category

Then, $\frac{n}{N} = \frac{20}{200} = 0.1$

Therefore, $n_1 = \frac{n}{N}(N_1) = 0.1\ (20) = 2$

$$N_2 = \frac{n}{N}(N_2) = 0.1\ (50) = 5$$

$$N_3 = \frac{n}{N}(N_3) = 0.1(130) = 13$$

Disproportional stratified sampling. If there is a great variance between the strata either in term of certain advantage than the other (e.g city and village). then it is better to use the disproportional stratified sampling. Therefore, the stratum that has more advantage in terms of number items in the population can have proportionally more sampling units than the other. This type of sampling can be computed thus

$$N = \frac{\sum n}{\sum q_i} \text{ where } i = 1, 2, 3 \ldots k.$$

Where n_i = number of sampling units of stratum
σ_i = *standard deviation of stratum*
n = *total sample size*
$q = N1/N$; *the ratio of stratum with that of the population*
n_i = *size of stratum*
n = *size of the population*
And K = *the number of strata*

Example 9.6

The education department of a state went the study, the standard of education in scholars. The analyst of the study divided the entire population of the schools into schools in the rural locations, semi-urban locations and urban locations. He found more variation between the schools in rural locations because of varying infrastructure than the variation between schools in semi urban locations. Same is the caser between the semi urban locations and the urban locations. The analyst prefers to use disproportionate stratified sampling in which the categories of location represent strata. The total number of schools in the state is 1200. The number of schools in the rural, semi-urban and the urban locations are 500, and 300, respectively. The variance of educational standard of the schools in these locations on 1-10 scale is 49, 16, and 4. If the disproportionate stratified sampling is to be sued with the sampling size of 90 determine the number of units for each category of locations.

Solution

Stratum 1: location of schools in rural areas
Stratum 2: Location of schools in semi-urban areas
Stratum 3: Location of schools in urban areas
We also have

Size of the population $N = 1200$
Size of stratum 1, $N1 = 500$
Size of stratum 2, $N_2 = 400$
Size of stratum 3, $N3 = 300$

Variance of stratum 1, $\sigma^2_1 = 49$

Variance of stratum 2, $\sigma^2_2 = 16$

Variance of stratum 3, $\sigma^3_3 = 4$

Sample size n = 90

Table 2 detail of calculation of sampling units

Stratum Number	*Size of Stratum*	$q = \frac{Ni}{N}$	σ	$\frac{qi\,\sigma\,in}{N}$	$qi\,\sigma\,i$	$n = \frac{q_1\,\sigma}{in\sum q_1\theta_1}$
1	500	0.416667	7	262.5	2.91669	52
2	400	0.333333	4	120.0	1.333332	25
3	300	0.2500000	2	45.0	0.500000	10
Total					4.750000	90

The researcher should note that (i) this technique is only applicable if the sample to be drawn does not constitute a homogenous group (ii) a random sampling is used to select the units of samples, hence, this type of sampling is an example of a unit sampling, (a situation whereby two sampling technique are used).

Cluster Sampling

In this type of sampling technique, the primary sampling units represent groups of elements rather than individual elements of the population. Then all the elements clustered in that primary sampling unit can be investigated. This type of sampling technique is very economical. The cluster population must be dissimilar (heterogeneous) in terms of attributes while elements within a cluster should be similar in terms of attributes. Hence any one of the clusters is randomly selected and all the units of that cluster are selected to arrive at inference about the population.

Multi-Stage Sampling

The researcher may employ more than one stage to sample the item of the population in a situation where the population is too large and too costly to handle. It is more practicable to select larger primary sample units such as state then reduce them to districts within the states, and then towns down to the family within the towns. This will make it easier to be administered such that the sampling frame can be developed in partial units. It also reduced cost

of administration and the time.

9.6 *Non-Probability Sampling*

Probability sampling has certain limitations despite the fact that it gives better accuracy in terms of confidence level inference. Some of these difficulties may be in terms of (i) a situation where the sampling units may be sparsely distributed such that it may be difficult to use probability sampling method. (ii) illiteracy of respondents such that the researcher may have to choose area where the respondents are literate and can understand his language (iii) refusal to fill the questionnaire (iv) negligence on the part of the interviewer to draw out sampling plan etc.

Hence the researcher may resort to convenience sampling to take care of the problem encountered.

Judgment Sampling

This is non-probability sampling such that the sampling unit are selected based on the value judgment and or intuition of the researcher or certain experts on the field of research. In this type of sampling the expert who is experienced in the sampling frame may guide the researcher in selecting sample units or the researcher may use his intuitive judgment and experience in selecting the sampling units from the sampling frame, this sampling technique open up widows for personal biases in the result except for seriousness on the part of the research.

Quota Sampling

The population from which sample is to be drawn can be categorised into groups based on their peculiarities like age difference educational background etc, then certain percentage of the sample size to be drawn can be allocated to these segmentation of the population representing the number of sample to be drown from each of the categorization of this population. This type of sampling is called the quota sampling

1. Purpose of sampling
2. Important feature of good sample
3. Benefits of sampling over census

Sampling distribution

The selection of unit of sample depends on the choice of the researcher. A particular units of sample can be selected so many times. in

which case, that unit of sample is replaced and it is picked as many times as possible. Therefore o\in a sampling where each member of the population is chosen more than once is called sampling with replacement, while a situation where a member of the population or a sampling unit can not be chosen more than once is called sampling without replacement.

Sampling with replacement is associated with an infinite population because a sample of any size can be drown. Theoretically such populations is considered an infinite population because, a member of the population the chance of been selected so many times. For instance, in a but containing thirty black balls and 25 red balls, a black ball can be chosen more than sixty time when replace while a red ball can be picked more than fifty times when replaced, this makes it an infinite population. But if the ball can only be drawn thirty time which makes it a finite population.

Example:

Given a population consisting the following: 2,3.6.8.11

a). Draw a sample of size two with replacement
b). Determine the mean of the population
c). Determine the standard deviation of the population
d). Determine the mean of the sampling distribution of the mean.
e). Determine the standard deviation of the sampling distribution of the mean.

(A) 5(5) = 25 possible sample of size two with replacement can be drown as follows:

(2,2)	(2,3)	(2,6)	(2,8)	(2,11)
(3,2)	(3,3)	(3,6)	(3,8)	(3,11)
(6,2)	(6,3)	(6,6)	(6,8)	(6,11)
(8,2)	(8,3)	(8,6)	(8,8)	(8,11)
(11,2)	(11,3)	(11,6)	(11,8)	(11,11)

(B) Population = 2,3,6,8,11.
Population mean = $\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.0$

(C) Population standard deviation variance α^2

$$\frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-2)^2}{5}$$

$$= \frac{16 + 9 + 0 + 4 + 25}{5} = 10.8$$

Therefore standard deviation

$$\alpha = \sqrt{\alpha 2} - \sqrt{10.8} = 3.29$$

(D The mean of the sampling distribution
From (a) above the corresponding sample means are

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

$$\mu_x = \frac{\Sigma\Sigma \overline{X}}{N} = \frac{\text{sum of all sample mean above}}{\text{Total number of possible sample}}$$

This proves the fact that ($\mu_x = \mu$) the mean of the sampling distribution of the mean is equal to the population mean

(A) $\alpha^2_x = \frac{\Sigma(X-\overline{X})2}{N} = \frac{135}{25} = 5.40$

This demonstrate the fact that for finite population involving sampling with replacement (or infinite population. i.e $\alpha^2_x = \alpha^2/n$ since variance above divided by 2 10.8/2 is also = 5.40.

Example 2: using the population above, (2,3,6,8,11.)

(a) Draw a sample without replacement

(b) Determine the mean of the sampling distribution of the mean

(c) Determine the variance of the sampling distribution of the mean.

5x2 possible sample can be drown without replacement

(a) (2,3), (2,8), (2,11), (3,6), (3,8), (3,11), (6,8), (6,11), (8,11).

Note that the solution of (2,3) is the same as (3,2) etc.

(b) The corresponding sample means are
2.5, 4.0, 5.0, 6,5, 4.5, 5.5, 7.0, 7.0, 8.5, 9.5

And the mean of sampling distribution of mean is given by.

$$\mu x = \frac{\Sigma \overline{X}}{N} = \frac{2.5+4.0+5.0+6.5+4.5+5.5+7.0+7.0+8.5+9.5}{10}$$

$$= 6.0$$

(c) Standard deviation α_x
Variance

$$\alpha^2_x = \Sigma (X-\overline{X})2 = \frac{(2.5-6.0)2+(4.0-6.0)2+\ldots+(9.0-6.0)}{10} = 4.05$$

This illustrate the fact that $\alpha^2_x = \frac{\alpha^2}{n}\frac{(N-n)}{N-1}$ since $10.8\ (5-2) = 4.05$

Sampling Distribution

sampling distribution can be refers to the computed values of th various statistical measures such as the means, variances and standar deviation together with their relative frequencies determined from th particular sample. In this case we can have the sampling distribution o mean, sampling distribution of variance as well as the samplin distribution of the standard deviation. Therefore each item in th sampling distribution is a particular statistic of a sample. The samplin distribution tends to be closer to the normal distribution when th sample size become larger.

It should be noted that the mean of the sampling distribution is the sam as the mean of the population. Therefore the mean of the samplin distribution can be taken as the mean of the population.

Sampling distribution of mean (normal population)

(1) When the population is infinite, the sampling distribution ar given thus:

Mean = μ = same as the population

Variance = $\frac{\alpha^2}{n}$ where n = sample size

standard deviation = $\frac{\alpha}{\sqrt{n}}$ = standard error of the mean

The standard error represents the true deviation of the various sample means from the actual population mean. The smaller the standard error the closer the sample means from the population mean.

The population variable is ...

$\mu \sim n(\mu, \alpha^2)$ hence, $\bar{X} \sim n(\mu, \alpha^2/n)$.

therefore the standard normal statistic for $\bar{X}$ is given by

$$Z_x = \frac{\bar{X} - \mu}{\alpha/\sqrt{n}}$$

(2) When the population is finite then the variance of the sampling distribution is given by variance = $\frac{\alpha^2}{n}$

however, if the sample size is large i.e of size N>30 then the variance is given by

$$\text{Variance} = \frac{\alpha^2}{n} \cdot \frac{N-n}{N-1}$$

Where (N-n) (N-1) = finite population multiplier which differentiates the variance of sampling distribution under finite population with that of infinite population $\bar{X}$ under finite population is enperted to follow normal distribution with its means = μ (same with distribution under infinite population).

and

$$\text{variance} = \frac{\alpha^2}{n} \cdot \frac{N-n}{N-1}$$

re, the standard normal statistic for $\bar{X}$ under finite tion is given by

$$Z_x = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

(3) When the population variance is not known.

If the variance of the population is not known then, the sampling distribution can be defined using the student t distribution. In this case the sample variance (S^2) can be used as an estimate of the population variance.

Therefore the standard normal statistic is given by

$$Z_x = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = t_x = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad \text{with (n-1) degree of freedom}$$

sampling distribution of variance

there exist two sampling distribution of variance

1. The chi-square (x^2) distribution.
2. the F distribution

(a) The chi-square (X^2) distribution

In this type of sampling distribution of variance, the variance of the distribution (S^2) is given by

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

Where x = observation

$\bar{X}$ = sample mean

n = sample size

degree of freedom = n-1, hence,

$$x2 = \frac{\Sigma (x-x)^2}{\sigma^2} = \frac{(n-1)S2}{\sigma^2} \quad \text{with (n-1) degree of freedom}$$

(b) The F – distribution

This distribution is the ratio of two cli-squares distribution variable. This means that it involves two independent samples each from an independent but normal population. hence if two independent sample of size are n_1 and n_2 were drawn with variances S^2_1 and S^2_2. if the populations from which the samples were drawn follow no variance distribution with variances σ^2_1 and σ_2, the F distribution is defined as.

$$F = \frac{(n_1-1)\, S^2_1 / \sigma^2_1}{n_1 - 1}$$

$$\frac{(n_2 - 1)\, S^2_1 / \sigma^2_2}{n_2 - 1}$$

with (n_1-1) and (n_2-2) degree of freedom

sampling distribution of proportion

given sample of n successive trail, let

p = probability of success

x = number of success

then x follows binomial distribution with mean

$\mu = np$ and variance

$\sigma^2 np(1-p) = npq$

where $q = (1-p)$

if the sample size is large, then, the binomial distribution can be approximated to normal distribution

$x\sqrt{N}(\mu, \sigma^2)$ hence the propotion of success p is given by

$p = x/n$

as the sample size become larger than proportion p also approaches normal distribution with mean.

mean, $P = E(x/n) = 1/n\ E(x) = 1/n\ np = p$

and its variance

variance, $P = var(x/n) = 1/n^2\ var(x)$

$= 1/n^2 np(1-P) = \frac{P(1-p)}{n}$

$= P\ v\ n\ [P, \frac{p(1-p)}{n}]$

CHAPTER TEN

ESTIMATION

10.1 Estimation is a systematic way of approximation of a particular (variable) population parameter through its sample values. It is associated with inferential statistics such that a sample value of a given statistic can be used to approximate the population value in such a manner that inferences can be made (about the future of such population) on the bases of the sample values of the population parameter

10.2 Estimator

An estimator is the representative of the population such that. it approximates the population parameter. It is a sample statistic which can be used to approximate or have the characteristics of the population.

A good example of an estimator is the sample mean, mode, and median. However, the sample mean usually provides the best estimate of the population. Therefore an estimate is a specific value or range of values that have the characteristics of the population and that can be used to approximate the population.

The sample mean is a better estimator of the population because of the following reasons:

1) The distribution of sample mean x tend to be more consistent (with less variance) than the distribution of the statistic.
2) The distribution of the sample mean x do not in any way underestimate or over estimate the population mean μ and hence it is an unbiased estimator of the population mean μ. That it tend to target the population mean μ.

10.3 Destrable properties of a good estimator

1) **Unbiaseness**
2) **Least variance**
3) **Efficiency**
4) **Sufficiency**

a) **Unbias Estimator**

An unbias estimator is the estimate that tends towards the true population parameter as the sample size gets larger. This means that it is an estimator that approximates the true parameter such that its expectation is equal to the true parameter i.e. $E(x) = \mu$.

b) **Minimum Variance**

An estimator is best estimator when it has the smallest variance as compared with any other estimate of the true parameter. Therefore, it is not enough for an estimator to be unbias to the true parameter; it must have minimum variance.

c) **Efficient Estimator**

An estimator is efficient when it is unbiased and has minimum variance. Therefore, it must possess both properties together for it to be efficient such that it will become the best (minimum variance) estimator within the class of all unbiased estimators.

d) **Sufficient estimator**

A sufficient estimator is the estimator that utilizes all the information a sample contains about the true population parameter such that it uses all the observations of the sample compared with any other estimatior.

10.4 Point Estimate

A point estimate is a single value used to approximate a population parameter. That is to say that it is an estimate of a population parameter that is given by a single number of a population mean μ. It is the best point estimate of the population mean μ. The sample mean is the best point estimate of the population mean.

10.5 Interval estimate

An interval estimate is an estimate which lies between two numbers. This means that the estimate of a population parameters, is given by two number within which the parameter may be considered or found.

Confident Interval

The sample mean x is known to be the best estimator of the true population mean μ especially when small sample is involved. Hence,

scholars developed another estimator that may ensure confidence in making inferences about the population. The estimate is called the confidence interval estimate which consists of range or an interval or values instead of just a value.

Therefore, a confidence interval or confidence estimate is a range of interval of values that is likely to contain the true value of the population parameter. It is associated with the degree of confidence which is a measure of how certain we are that the interval contains the population parameter.

The degree of confidence

The degree of confidence is the probability $1 - \alpha$ (expressed as the equivalent percentage value) that the confidence interval contains true value of the population parameter. It is also called the confidence level or confidence coefficient denoted by alpha (α). It can be located at the extreme side of a normal curve.

The degree of confidence, level of significance or confidence level is therefore the degree to which a researcher is confident or the degree to which he agrees to make an error.

The common choice for the degree of confidence are 90% with $\alpha = 0.10$, 95% with $\alpha = 0.05$ and 99% with $\alpha = 0.01$. the choice of a 95% is most common because it provides a good balance between precision, as reflected in the width of the confidence interval and reliability as expressed by the degree of confidence.

10.6 Critical Value

The notation $Z_{\alpha/2}$ is the positive Z value at the vertical boundary separating an area. The area of $-Z_{\alpha/2}$ is at the vertical boundary for the area of $\alpha/2$ in the left tail.

A critical value is the number in the borderline separating sample statistic that is likely to occur. The number $Z_{\alpha/2}$ is a critical value that is a Z score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution. Therefore, there is an area of $1-\alpha$ between the vertical border lines at $-Z_{\alpha/2}$ and $Z_{\alpha/2}$

The common degree of confidence and their critical values

Degree of confidence		Critical value Z
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.574

10.7 Margin of Error (E)

The critical value as defined above is closely related to the interval estimate. Given a sample, we can obtain the sample mean x which is typically different from the population mean μ. Therefore, the difference between the sample mean x and the population mean μ is seen as an error. Using the sample standard error and notation ($Z_{\alpha/2}$) of the critical value we can define the margin of error.

When a sample data used to estimate a population mean μ, the margin of error denoted by E, is the maximum likely difference between the observed sample mean x and the true value of the population mean μ. The margin of error E is also called the maximum error of the estimate and can be found by multiplying the critical value and the standard deviation of the sample means.

$$E = Z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

The error margin above indicates that there is a probability of 1 – α that sample mean will be in error by not more than E and there is a probability of α that a sample mean will be in error by more than E.

Calculating E when δ is unknown

If n > 30, we can replace population standard deviation δ in E by the sample standard deviation S. But if n ≤ 30, the population must have a normal distribution. We may use the formula for E above.

Confidence interval for population mean when n > 30

$X - E < \mu < x + E$ or $\mu = x \pm E$

Where $E = Z_{\alpha/2} \cdot \sigma/\sqrt{n}$

The value x – E and x + E are confident limits.

Procedure for constructing a confident interval for the population mean μ when n > 30

1) Find the critical value $Z_{\alpha/2}$ that corresponds to the desired degree of confidence 9 for example, if the degree of confidence is 95%, the critical value is $Z_{\alpha/2} = 1.96$).
2) Evaluate the margin of error $E = Z_{\alpha/2}.\sigma/n$." If the population standard deviation θ is unknown, use the value of the sample standard deviation S provided that N > 30.
3) Using the value of margin of error and the value of the sample mean x, find the value of x – E and x + E (i.e. confidence limits). Then substitute those values in the general format of the confidence interval/

   $X - E < \mu < X + E$

   Or

   $\mu = x \pm E$
4) Round the resulting value round off for confidence interval for (n > 30) large samples with replacement or infinite population the confidence is given by

   $X \pm Z_{\alpha/2} \cdot \sigma/\sqrt{n} \cdot \sqrt{\frac{N-n}{N-1}}$

   Since the population standard deviation is usually not known, hence the sample standard deviation δ.

   Confidence interval for (n < 30) small sample (for population mean) is given by

   $X \pm t_{\alpha/2} \cdot \delta/n$

   T distribution is used for small sample

   Confidence interval for proportions n ≥ 30

   The confidence limit is given by

   $P \pm Z_{\alpha/2}\sqrt{pq/n} = P \pm Z_{\alpha/2}\sqrt{P(1-P)/n}$

   For sampling with replacement or infinite population the confidence limit is given by

   $P \pm Z_{\alpha/2}\sqrt{pq/n} \sqrt{\frac{N-n}{N-1}}$

CHAPTER ELEVEN

11.1 LOGIC OF HYPOTHESIS TESTING

A hypothesis is an assumption about the frequency distribution of a random variable. It is a tentative statement that can be subjected to test. Hypothesis testing is concerned with developing rules or procedures for deciding whether to reject or accept the hypothesis. It is a way of determining whether a given observation or finding is compatible with some stated assumptions. That is, whether the finding is sufficiently close to the hypothesized value so that we do not reject the stated hypothesis (preposition).

If on a prior basis, we believe that the time value of β is unity and the sample estimated value is 0.498, then is this value consistent with the stated hypothesis?

A research hypothesis is a statement of assumption that is made in order to test the logical and empirical reality of study. Hypothesis is very important such that, it should be specific to the research at hand. Hypothesis can be classified into the Null hypothesis and the Alternative hypothesis

The null hypothesis is set in such a way that available evidence will nullify it and hence, it is set against the direction of the research or observation such that it can be nullified easily. Example, H_0: $\beta = 0$. The alternative hypothesis which is rival hypothesis is set in line with the theoretical postulation.

As earlier stated, the Null hypothesis (usually denoted by Ho) is set in such a way that small evidence can nullify it. For instance, if we are comparing the performance of 100 level and 200 level students in terms of their attitude towards learning, and we state that their attitudes are the same then, we have set a null hypothesis, since small evidence in terms of different can nullify this assumption.

The null hypothesis is tested against (a rival) an alternative which is also called maintained hypothesis denoted by H_1 An alternative (assumption) hypothesis is set according to the observation

or the theoretical foundation of the work at hand, for instance theory specified that quantity demand is a negative function of price hence, the higher the price the lower the quantity demanded. While a null hypothesis will state that price does not have any impact on quantity demand. However, an alternative hypothesis will state that price has a negative impact on quantity demand.

The role of hypothesis is to guide the researcher by delimiting the area of search and keeping him on the right track as well as shapens his thinking so as to focus his attention on the important facets of the problem at hand. Hypothesis is also important because it indicate the type of data required, the approach to the study, as well as the analytical techniques to be used.

11.2 Steps involved in hypothesis testing

There are about (6) six steps involved in testing of hypothesis.

1. **Formullation of hypothesis**
2. Choose the level of significance.
3. Choose the location of the critical and or rejection region.
4. Choose the appropriate test statistic.
5. Compute the sample value of the choosen test statistic.
6. Compare the actual sample value with the critical or table value of the statistic.

Step 1 **Formulation of hypothesis.**

(a) H_0: Null Hypothesis is set against a rival hypothesis in such a way that available evidence will nullify it.

(a) Hi: Altanative or rival hypothesis is set again Ho as a rival hypothesis in order to direct the researcher.

E.g. *Ho:μ=30 and*

Ho: $\mu = 30$ - then, it is a two tail test

The two tail is divied into two

If *$\mu < 30$* - *One tail test to the left*

$\mu > 30$ - *One tail test to the right*

A hypothesis is an assumption which we make about a population parametre.

In the test of hypothesis, we begin by making some assumptions about the true value of the population or estimate to decide whether or not the assumption is acceptable.

The hypothesis which would react to test is called the null hypothesis, it is set in such a way that available evidence will nullify it.

It is called a null hypothesis because, it implies that there are no difference between the true popparamere and the (sample information or estimate) hypothesized value.

The alternative hypothesis represent a rival assumption about the population parametre. In otherwords, it is a counter proposition to the null hypothesis.

The Null hypothesis is denoted by Ho while alternative hypothesis is denoted by Hi.

The alternative hypothesis implies that the population parametre is not equall to the specified (sample) value.

The alternative hypothesis is important in defining the location of the rejection or the critical region of the test.

For the purpose of conducting a test of hypothesis about a population parametre, the whole sets of the values of the population is devided into two regions.

-The acceptance region and the rejection or critical region.

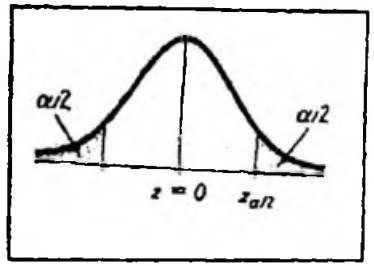
The acceptance region includes the values of (x) the population which have

high probabilities of being observed while the reject or critical region include the value of (x) which are not likely to occur.

These highly unlikely values of x are those values whose total probability is less than 5% or less than 1% i.e those that define an area of the probability distribution equall to 0.05 and 0.01.

These probabilities are known as the level of significance of the test.

Step 2. Choose level of significance of the test



In making a decision, one can never be 100% sure that he/she will make(or has made) the right decission, in this respect therefore, the researcher may choose the level of error at which he will be comfortable. This acceptable limit at which the researcher is willing to make mistake is called level of significance. In making any decision one is likely to commit the following types of errors.

1. Type one error occurs when one reject the null hypothesis when it is actually true.
2. Type two error occurs when one accept null hypothesis when it is actually not true.

Fact / *Decission*	*Ho. is true*	*Ho is False*
Accept Ho, Ho is true	*Correct decission ($1 - \alpha$)*	*Type II (β)*
Reject Ho Ho is False	*Type I Error a*	*Correct decision ($1 - \beta$)*

The probability that these errors will occur is given by

P(Type I error) = P(reject Ho/Ho is true) = α

and

P(Type II error)= P(accept Ho/Ho is False) = β

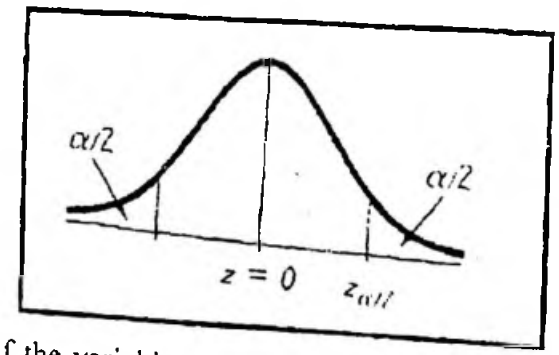
Of course, the researcher would like α and β to be as small as possible, hence, α and β are called significance level i.e. The likely hood of making any of the errors.

It is obvious that we want these probabilities of being wrong to be small (ie to be minimum).

That is why we usually assign low values to the probabilities (ie probabilities of been wrong) and we call it level of significance of the test.

Choosing a certain level of significance involves our specifying the probability of encountering a type one error and in most cases we specify a level of significance of 5% and in rear cases 1%.

*Step 3.*Choosing the location of the critical or rejection region.



The value of the variable x which has a very low probability of been observed i.e. a total proberbility of 5% or 1%.In other words, the critical region includes the area which correspond to the level of significance of the test.The critical region may be choosen either at right tail or left tail of the distribution.

Where we have the critical region entirely on the right tail or

left tail we are said to conduct a one tail test where the critical divides on the two tails or ends of the distribution, then we have a two tail test.

The decision of whether to choose a one tail or two tail critical region depend on the form in which the altanative hypothesis is expressed. That is to say that the tail in which the critical or rejection region lie is determined by the theoretical foundation of the research in question.

$Hi: \mu$? 30	*located in the two tails*
$Hi: \mu > 30$	*right tail*
$Hi: \mu < 30$	*left tail*

Step 4. Choosing the apropriate test statistic

The most commonly used statistic is the t, z and f statistic (ie statistic which are used in testing hypothesis).

The choice among the test statistic depends on the type of test you want to cary out, the size of the sample and the information which we have about the population variane. For instance, if the sample size $n > 30$, then it is appropriate to use the *z statistic,* but if the sample size $n < 30$, it is appropriate to use the *student t test statistic.*

Step 5 Compute the sample value of the choosen test statistic

Using the sample information, we compute the sample value or the observed value of the choosen statistic which is denoted with asteric (z* t* f*) to distinguish it from the critical value otherwise known as theoritical value or tabular value.

Step 6 .Compare the computed sample value with the critical or table value of the statistics.

The choice among the test statistic depends on the type of test you want to cary out, the size of the sample and the information which we have about the population variane. For instance, if the sample size $n > 30$, then it is appropriate to use the *z statistic*, but if the sample size $n < 30$, it is appropriate to use the *student t test statistic*.

Step 5 Compute the sample value of the choosen test statistic

Using the sample information, we compute the sample value or the observed value of the choosen statistic which is denoted with asteric (z* t* f*) to distinguish it from the critical value otherwise known as theoritical value or tabular value.

Step 6 Compare the computed sample value with the critical or table value of the statistics.

Example 11.1

The average weight of horses was found to be 100kg from previous records. Their standred deviation was 20kg. A new diet or horse-feed was inovated and was given to a large number of horses. A random sample of 35 horses was selected and the average weight is found to be 110kg. Using a 5% level of significance, can you say that there is a significant gain in the weight of the horses?

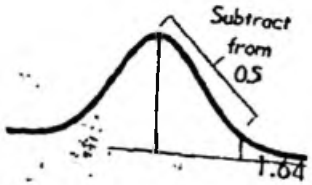
Step 1

$Ho : \mu = 100$

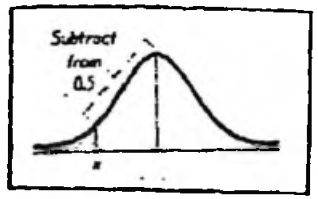
$Hi : \mu > 100$

Step 2 *Level of significance is 5%*

Step 3 - $Z_{0.05} = 1.640$



Step 3



Step 4 t table

Step 5 $t^* = \dfrac{x - \mu}{\frac{s}{\sqrt{n}}} = \dfrac{42 - 50}{\frac{11.9}{\sqrt{12}}}$

$t^* = \dfrac{-8}{\frac{11.9}{3.46}} = \dfrac{-2.33}{3.44} = -2.72$

Step 6 *Discission*

Since -2.72 > - 2.33, we accept H_o and conclude

A coruse in statisstics is thought to 2 seperate groups of students by two different method. In the first group 12 students were thought by the conventional classroom method. The second group consisting of 10 is thought by a new method.

At the end of the semester, the same examination was given to each group. The 12 students in the first group made an average grade of 85 with standard deviation of 4 while the 2nd of 10 students made an average grade of 81 with standard deviation of 5.

Test the hypothesis that the two methods are equal using 5% degree level of significance. Assuming that the population is normally distributed.

Step 1 *Ho : $\mu = \mu 2$ or $\mu i - \mu 2 = 0$*

Step 1 $Ho: \mu = \mu 2$ or $\mu i - \mu 2 = 0$

$Hi: \mu \neq \mu 2$ or $\mu 1 - \mu 2 \neq 0$

Step 2 level of using $\alpha = 0.05\%$

Step 3 two tail test

Step 4 $t_\alpha = t_{0.025}$ with degree of freedom

Where $df\ v =$

$$\frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2 - 1}} \qquad \frac{\left[\frac{16}{12} + \frac{25}{10}\right]^2}{\frac{\left[\frac{16}{12}\right]^2}{12-1} + \frac{\left[\frac{25}{10}\right]^2}{10-1}}$$

$$= \frac{\left[\frac{(1.33 + 2.2)}{1.78 + 2.25}\right]^2}{\frac{11}{9}} \qquad = \frac{(3.83)^2}{0.16 + 0.69}$$

$$= \frac{14.67}{0.85} \qquad = 17.26$$

$t^* < -t_{0.025}\ (v = 17)$ or $t^* > +t_{0.025}\ (v = 17)$

$t^* < -2.110$ and $t^* > +2.110$

Step 5 *Computation*

$$t^* = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1^2)}{n_1} + \frac{(s_2^2)}{n_2}}}$$

$$= \frac{85 - 81 - \mu 1 - \mu 2}{\sqrt{\frac{(16)}{(12)} + \frac{(25)}{(10)}}}$$

Since according to our hypothesis $\mu_1 - \mu_2 = 0$

$$= \frac{85 - 81 - 0}{\sqrt{\frac{16}{12} + \frac{25}{10}}}$$

$$\frac{4}{\sqrt{1.33 + 2.5}} = \frac{4}{3.83} = \frac{4}{1.96} = 2.04$$

Step 6 Accept H_o and conclude that the two method of teaching are basically the same.

Example 5

Using eg 4, test the hypothesis that the 2 methods of learning are equally at 01/0 than at 10% level of significance. Assuming that the pop is aproximately normally distributed with equall variances.

$$Z = \frac{s = \mu s}{\sigma s}$$

For Means *Here $s = x$, the sample mean*

$\mu s = \mu x = \mu$ - *Pop mean and*

$\sigma s = \sigma x = \sigma/\sqrt{N}$

Where σ = pop st div a N = sample

Hence, $Z = \frac{x - \mu}{\sigma/\sqrt{N}}$

11.3 Some aspects of Hypothesis Testing

The mean tests

When the population variance is known and the population is infinite,

a. One tailed test of a simple mean

$H_0: \mu \geq 0$

$Hi: \mu > 0$

The following formula can be used

1. $P = \mu + Z\alpha \frac{\sigma}{\sqrt{n}}$

2. $Z\alpha \frac{\sigma}{\sqrt{n}}$

b. *Two tailed*

$Ho: \mu = 0$

$Hi: \mu \geq 0$

The following formulas can be used

1. $P = \mu \pm Z\alpha \frac{\sigma}{\sqrt{n}}$

2. $Z\alpha \frac{X - \mu}{\theta/\sqrt{n}}$

When the population variance is known and the population is finite

a. One tailed test of single mean

$Ho: \mu = 0$

$Hi: \mu = 0$

$$Z_a = \frac{X - \mu}{\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$$

b. Two tailed of a single mean

$Ho: \mu = 0$

$Hi: \mu = 0$

$$Z_\alpha = \frac{X - \mu}{\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$$

When the population variance is unknown and the sample size is large

a. *One tailed of a single mean*

$Ho: \mu = 0$

$Hi: \mu = 0$

For infinite population

$$Z\alpha \frac{X - \mu}{5/\sqrt{n}}$$

For a finite population $\sigma^2 x = \frac{\sigma^2 N - n}{N - n} = \frac{5^2 N - n}{N - n}$

Therefore

$$Z_a = \frac{X - \mu}{\sigma\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$$

Two tailed test for a single mean

$Ho: \mu = 0$

$Hi: \mu = 0$

For infinite population

$$Z\alpha = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{X - \mu}{5/\sqrt{n}}$$

For finite population

$$Z\alpha = \frac{X-\mu}{\sigma\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$$

$$= \frac{X-\mu}{s\sqrt{n}\sqrt{\frac{N-n}{N-1}}}$$

When the population variance is unknown and the sample size is small

$t = \frac{X-\mu}{s/\sqrt{n}}$ *with (n - 1) degree of freedom*

The same for two failed except for the hypothesis setting

$H_o: \mu = 0$

$H_1: \mu = 0$

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

Hypothesis testing of difference between means

When the population variances are known

a. *One tailed test*

Ho: $\mu_1 - \mu_2 = 0$

Hi: $\mu_1 - \mu_2 > 0$

$$Z\alpha_1 - X_2 = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sigma_{x2} - X_2}$$

where $\sigma_{x1} - X_2 = \frac{\sigma^2_{x1}}{n_1} + \frac{\sigma^2}{n_2}$

b. *Two tailed test*

Ho: $\mu_1 - \mu_2 = 0$

Hi: $\mu_1 - \mu_2 > 0$

The Z statistics is the same as the one failed test

When the population variance is unknown and the sample sizes are large

Ho: $\mu_1 - \mu_2 = 0$

Hi: $\mu_1 - \mu_2 > 0$

$$Z\alpha_1 - X_2 = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sigma_{x2} - X_2}$$

Where $\sigma_{x1} - X_2 = \frac{S^2_1}{n} + \frac{S^2_2}{n}$

Since the population variance is unknown

Two taile

Ho: $\mu_1 - \mu_2 = 0$

Hi: $\mu_1 - \mu_2 > 0$

$$Z\alpha_1 - X_2 = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\theta_{x2} - X_2}$$

Where $\sigma_{x1} - X_2 = \frac{S^2_1}{n} + \frac{S^2_2}{n}$

Since the sample sizes are large

When the population variance is unknown and the sample size is small

a. *One tailed test*

$Ho: \mu_1 - \mu_2 = 0$

$Hi: \mu_1 - \mu_2 > 0$

Since the sample sizes are small, the pooled variance become (i.e $S^2 = \sigma^2$)

$$S^2_p = \frac{(n_1 - 1) S^2_1 + (n_2 - 1) S^2_2}{n_1 + n_2 - 2}$$

And the standard deviation becomes

$\sigma_{x1} - X_2 = Sp \sqrt{1/n_1 + 1/n_2}$

$$Therefore\ t = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sigma_{x1} - X_2}$$

$$= \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{sp\sqrt{(^1/n_1) + (^1/n_2)}}$$

Two tailed test

$Ho: \mu_1 - \mu_2 = 0$

$Hi: \mu_1 - \mu_2 = 0$

The t statistic is the same as the one tailed test

Test of hypothesis about proportion

a. One tailed test

$Ho: p = 0$ or $Ho: p = 0$

$Hi: p = 0$ $Hi: p = 0$

$$Zp = \frac{p - p}{\sigma p} \quad or \quad Zp = \frac{\beta - \beta}{\sigma \beta}$$

$Where\ \sigma p = \sqrt{\rho(1 - \rho)}$

Two tailed

$Ho: \rho = 0$

$Hi: \rho = 0$

The Z statistics and the θ are the same one failed test

Test of hypothesis

c. One tailed

$Ho: \rho_1 - \rho_2 = 0$

$Hi: \rho_1 - \rho_2 \neq 0$

$$Z\rho_1 - \rho_2 = \frac{(\rho_1 - \rho_2) - (\rho_1 - \rho_2)}{\sigma\rho_1 - \rho_2}$$

$$where\ \theta\rho_1 - \rho_2 = \rho_1 \frac{(1 - p_1) + p_2 (1 - p_2)}{n_1\ n_2}$$

d. *Two tailed test*

$Ho: p - p = 0$

$Hi: p - p \neq 0$

CHAPTER TWELVE

12.1 ANALYSIS OF VARIANCE (ANOVA)

The basic aim of analysis of variance is to split total variation of a variable (around its mean) into components which may be attributed to specific (additive) causes.

It is used in its statistical form to answer the question.

1. *does a significant difference exist any where among three or more sample?*
2. *did all samples come from the same population*
3. *did at least one sample come from different population?*

Anova is sometimes used along side regression analysis in order to:

1. *Test the ovral significant of regression.*
2. *To obtain the test of improvement in the fit by introducing additional explanatory variable in the function.*
3. *To test equality of coeficient obtained from different samples.*

It is significantly the same with regression analysis since both of them are aimed at spliting total variation into additive components.

Total variation = Explain variation + unexplained

$TSS = ESS + RSS$

However, the main difference between the two methods is that, apart from the above, regression provides numerical value for the influence of various explanatory factors.

12.2 ANOVA Test of difference between means.

$TSS = ESS + RSS$

To analyse these variation we first of all look at the causes and sorces of variation which is rooted in the mean.

Therefore we look at within group means and between group mean.

To do this we have to formulate our hypothesis

$H_o : \mu_1 = \mu_2 = \mu_3$ or $H_o : F = 0$

$H_i : \mu_1, \mu_2, \mu_3$, *are not equal* or $H_i : F \neq 0$

If the mean are same, i.e. if Ho is true, the entire population is considered normally distrributed with mean $\mu(=\mu_1, \mu_2, \mu_3)$ and standard deviation Ó i.e

$y \sim N(\mu, Ó)$,

Therefore, we estimate the common mean

$$\mu = \frac{\Sigma Y}{N} = \frac{\Sigma\Sigma_{ji} Y_{ji}}{N} = y$$

Between group $$\sigma^2 = \frac{\Sigma nj\,(Y-Y)^2}{K-1} = MSSB$$

Within group $$\sigma^2 = \frac{\Sigma\Sigma\,(Y_{ji}-Y_j)^2}{N-K} = MSSW$$

Hence $\Sigma_{nj}\,(Y-Y)^2 + \Sigma\Sigma(Y_{ji} - Y_{ji})2 = TSS = \Sigma\Sigma\,(Y_{ji} - Y)$

$SSB + SSW = TSS$

(Between variation) + (within variation) = Total variation

$$F^* = \frac{\Sigma nj(Yji - Y)^2/K-1}{} = \frac{MSSB}{}$$

$\Sigma\Sigma(Y_{ji} - Y_{ji})^2 MSSW$

Where nj = size of j sample

$N = \Sigma nj$ = *size of pooled (enlarge) sample*

K = *no of samples*

F* = Estimate of variance from between the mean variation

Estimate of variance from within the sample variance

ANOVA TABLE

(1)	Sum of square (2)	Degree of freedom (3)	Mean square (4) = (2): (3)	F (5)
Between the means	$\Sigma uj\ (Y_{ji}-Y)^2$	$V = K-1$	$\Sigma nj\ (Y-Y)^2$	$F = \Sigma nj\ (Y-Y)^2/K-1$
Within the samples	$\Sigma\Sigma(Y_{ji} - Y_j)^2$	$V=N-K$	$\Sigma\Sigma(Y_{ji}-Y)^2$	$\Sigma\Sigma\ (Y_{ji}-Y_j)/N-K$
Total variation	$\Sigma\Sigma\ (Yji - Y)^2$	$V= N-1$		

Example 12.1

Suppose we want to find out whether or not there is a significant difference in performance in statistic among the five departments that made up the faculty of social science, suppose also you conducted a statistics test in all the debts and obtained the follwing results.

Soc.	Psy	Geo	Eco	Pol
8	7	5	7	3
3	9	4	6	3
5	5	4	8	6
9	4	6	9	3
9	6	2	5	5
8	6	9	7	4
$\Sigma x_1 = 42$	$\Sigma x_2 = 36$	$\Sigma x_3 = 30$	$\Sigma x_4 = 42$	$\Sigma x_5 = 24$
Σx_1^2 324	$\Sigma x_2^2 = 234$	$\Sigma x_3^2 = 178$	$\Sigma x_4^2 = 304$	$\Sigma x_5^2 = 104$
$x_1 = 7$	$x_2 = 6$	$x_3 = 5$	$x_4 = 7$	$x_5 = 4$

$H_o : F = 0$ = *The five departments are not likely to be different in terms of students statistics scores.*

$H_o : F \neq 0$ = *The five departments are likely to be different in terms of student statistics scores.*

ANOVA

SSW = $\Sigma\Sigma(x-x)^2$

(x_1-x_1)	$(x_1-x_1)^2$	(x_2-x_2)	$(x_2-x_2)^2$
8-7=1	1	7-6 =1	1
3-7= -4	16	9-6 = 3	9
5-7= -2	4	4-6 =-2	4
5-7 = 2	4	4-6 =-2	4
9-7= 2	4	6-6 = 0	0
8-9 =1	1	6-6 = 0	0
Total	30		18

(x_3-x_3)	$(x_3-x_3)^2$	(x_4-x_4)	$(x_4-x_4)^2$	(x_5-x_5)	$(x_5-x_5)^2$
5-5 = 0	0	7-7 = 0	0	3-4 = -1	1
4-5 =-1	1	6-7 = 1	1	3-4 = -1	1
4-5 =-1	1	8-7 = 1	1	6-4 = 2	4
6-5 = 1	1	9-7 = 2	4	3-4 = -1	1
2-5 =-3	9	5-7 = -2	4.	5-4 = 1	1
9-5 = 4	16	7-7 = 0	0	4-4 = 0	0
Total	28		10		08

SSw = 30 + 18 + 29 + 10 + 08 = 94

MSSw = $\frac{\Sigma\Sigma(x-x)}{N-k} = \frac{94}{}$

SSB = $\Sigma n(x-x)^2$

$6(7-5.8)^2 = 6(1.2)^2 = 6(1.44) = 8.64$

$6(6-5.8)^2 = 6(0.2)^2 = 6(0.04) = 0.24$

$6(5-5.8)^2 = 6(-0.8)^2 = 6(0.64) = 3.84$

$6(7-5.8)^2 = 6(1-2)^2 = 6(1.44) = 8.64$

$6(4-5.8)^2 = 6(-1.8)^2 = 6(3.24) = 19.44$

40.8

MSSB = $\frac{SSB}{k-1} = \frac{\Sigma n(x-x_T)^2}{k-1}$

Therefore, $F^* = \frac{MSS\ B}{MSS\ W} = \frac{40.8}{94} = 2.71$

Total variable $\Sigma(x-xT) = \Sigma ni(x-xT) + \Sigma\Sigma(x-x)2 = 134.80$

Short method ***but not necessarily easy mehod.***

It start by calculating the following:

$I\ \frac{(\Sigma\Sigma x)^2}{\Sigma N} = \frac{(40+36+30+42+24)^2}{30} = \frac{(174)^2}{30} = \frac{30276}{30} = 10092$

$\Sigma\Sigma x$ = *the* Σ *of all sample accross all groups*

ΣN = *sum of sample sizes*

$II.\ \Sigma\Sigma x^2 = (\Sigma x^2{}_1 + \Sigma x^2{}_2 + \ldots\ldots + \Sigma x^2{}_s) = (324+234+\ldots\ldots+104) = 1144$

$\Sigma\Sigma x^2$ = *the sum of sum ofx squares*

III. $\frac{\Sigma(\Sigma x)^2}{n} = ((\Sigma x)^2 + (\Sigma x2)^2 + ... + (\Sigma x)^2) = \frac{6300}{6} = 1050$

$$SSB = \frac{\Sigma(\Sigma x)^2}{N} - \frac{(\Sigma\Sigma x)^2}{\Sigma n}$$

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

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

Having obtained the values of *SSB, SSW, and SST*, we procceed to set up the *ANOVA table*

12.3 ANOVA Summary Table

Sources of variation	*Df*	*ss*	*ms*	*F**
Between groups	*4*	*40.8*	*10.20*	*2.71*
Within groups	*25*	*94.00*	*3.76*	
Total	*29*	*134.80*		

NB: *df between = k-1 where k = no of groups under comparism*
Df Within = N-k where K = no of groups, N = no of cases across all samples (i.e..6x5 = 30).

Since, we have now obtained our *F* to be 2.71*, we then measure *F* against* F_T

In the table, the difference between are horizontal, while degree of freedom (df) and the sums of squares (ss) can be located on the vertical axis.

Df between = 4, the df within = 25, P = 0.05 $F_T = 2.76$. Hence we accept Ho since $F^* < F_T$ and conclude that there is not diff among the five departments in terms of the students' statistics scores. Therefore, the observed difference between groups is attributed to chance and or sample error.

CHAPTER THIRTEEN

13.1 CURVE FITTING

In most cases we wish to express the relationship that exists between two or more variables through curves and mathematical equation that may best approximate these expressions to a set of data. Therefore, curve fitting is all about finding equation that approximates the curve which fits a given set of data.

It is a way of finding not only the equation but obtaining the values of both the known and the unknown variables within the equation and putting them together. Curve fitting is important because it gives a clear picture of the relationship between variable and enhances accuracy in inferences.

13.2 Steps involved in Curve Fitting

1) Collect data indicating the corresponding values of variables under study. Example x = price, y = quantity. Therefore, the sample of n quantities and their prices will be revealed
   $y_1, y_2, y_3, \ldots\ldots y_n$ quantity and $x_1, x_2, x_3, \ldots\ldots x_n$ price.
2) Plot the point of intersection between variables starting from the first quantity and price on a positive curve or a rectangular coordinate system. Therefore we may have a scattered diagram.
3) Use the scattered diagram to visualize a kind of a curve approximating the data. This kind of curve is called an approximating curve.

The approximating curve may appear to approximate the data so well, in which case, we can say that linear relationship exists between the two variables. It may appear not to represent the data well, in which case there is a non-linear relationship between the variables.

If the relationship is linear, the equation will be a straight line

$Y = a + bx$

However if the relationship is non-linear, then the equation may be parabola or quadratic

$Y = a + bx + cx^2$

The main purpose of curve fitting is to estimate the dependent variable from the independent variable. However, two method of fitting a curve to a set of points or data exist.

Least square method and the non-least square method. However, our main concern here is the least square method because of its flexibility and simple for statistical as well as mathematical application.

Recall that in a scattered diagram, one can visualize a relationship between the variables under study. That is, the scattered points may tend towards a particular direction; however, a closer look may reveal more than one line.

In this case, therefore, the least square method of fitting a curve to a set of point states that one should choose a curve whose sums of square deviation is minimum and to do that the least square method outlined some steps.

Steps in least square method of fitting a curve

1) Find the error
2) Sum the error
3) Square the sums of the error
4) Minimize the error to obtain the normal equations
5) Use normal equation to fit the equation.
6) Estimate the parameter in the equation
7) Fit the equation with the values of the variables.

The above steps can be handled thus;

Given the model $Y = \beta o + \beta_1 X_1 + U$

(a) Find the error; if $Y = \beta o + \beta_1 X_1 + U$, *then the error*

$U = Y - \beta o - \beta_1 X_1$

(b) Sum the error, that is $\Sigma U = \Sigma (Y - \beta o - \beta_1 X_1)$

(c) Square the sum of the error, that is $\Sigma U^2 = \Sigma (Y - \beta o - \beta_1 X_1)^2$

(d) Then minimize the error, by taking partial derivative in respect of βo *and* β_1.

These estimate or coefficient can be obtained by minimizing the sum of squared residuals. i.e $?e_i^2 = \Sigma(Y_i - Y_i)^2 = \Sigma(Y - \beta_0 - \beta_1 X_1)^2$

We obtain by

(1) Finding the error, (in our model)

$$U_1 = Y - \hat{\beta}_0 - \hat{\beta}_1 X_1$$

(2) Squaring the error

$$U_i^2 = (Y - \beta_0 - \beta_1 X_1)^2$$

(3) Summing the squared error

$$\Sigma U_i^2 = \Sigma (Y \quad \beta_0 - \beta_1 X_1)^2$$

To minimize we take partial derivation in respect of β_0, and β_1 respectively so as to get our normal equation and subsequently, the coefficient β_0, and β_1 i.e. take partial derivation and equate it to zero.

$$\frac{d\Sigma Ui^2}{d\beta_0} = (-1)(2)\,\Sigma(Y - \beta_0 - \beta_1 X_1) = 0$$

$$= -2\,\Sigma\,(Y - \beta_0 - \beta_1 X_1) = 0$$

Divide through by -2

$$= \frac{-2\,\Sigma\,(Y \quad \Sigma - \Sigma X_1)}{-2} = \frac{0}{-2}$$

$$= \Sigma(Y \quad \Sigma_o - \Sigma Y_1) = 0$$

Where $\frac{0}{-2} = 0$

Open the bracket

Open bracket

$$= \Sigma Y - \Sigma\beta_0 - \Sigma\beta_1 X_1 = 0$$

$$= \Sigma Y - N\beta_0 - \beta_1 \Sigma X_1 = 0 \quad \text{.....(1) Normal equation}$$

Where $\Sigma\,\beta_0 = \beta_0 + \beta_0 + \beta_0 + \ldots\ldots = N\,\beta_0$ *i.e. Number of* β_0

To find our β_0, we multiply through by N and make β_0 the subject of the formula.

$$\frac{\Sigma Y}{N} = \frac{N\beta_0}{N} + \beta_1 \frac{\Sigma X_1}{N}$$

$$\overline{Y} = \beta_0 + \overline{\beta_1 X_1}$$

Therefore

$$\beta_0 = Y - \beta_1 X_1$$

To get the 2nd normal equation we take partial derivation in respect of the β_1

$$\frac{d\Sigma Ui^2}{d\beta_1} = (-X_1)\,(2)\,\Sigma(Y - \beta_0 - \beta_1 X_1) = 0$$

$$= (-2X)\,\Sigma\,(Y - \beta_0 - \beta_1 X_1) = 0$$

To eliminate -2 we divide through by -2

$$\frac{(-2X)\Sigma(Y - \beta_0 - \beta_1 X_1)}{-2} = \frac{0}{-2}$$

$$X_1\Sigma\,(Y - \beta_0 - \beta_1 X_1) = 0$$

$$\Sigma X_1 Y - \beta_0 \Sigma X_1 - \beta_1 \Sigma\, X_1^2 = 0$$

$$\Sigma X_1 Y - \beta_0 \Sigma X_1 - \beta_1 \Sigma\, X_1^2 \text{....(2) normal equation}$$

$$\beta_0 = \Sigma Y - N\beta_0 + \beta_1 \Sigma X_1 \quad \text{...(1)}$$

$$\beta_1 = \Sigma X_1 Y = \beta_0 \Sigma X_1 + \beta_1 \Sigma X_1^2 \quad \text{...(2)}$$

These equations are called normal equations and are used to obtain the coefficient β_0, *and* β_1 in other words they are used to fit the regression line. To do that we use crammer rule matrix as follows

Given the above *OLS* estimators we need to also obtain the standard error of these estimates which are important to establish confidence interval and to test the statistical hypothesis. The relevant formulas are as follows

$$\begin{bmatrix} A \end{bmatrix} = \begin{bmatrix} \Sigma Y \\ \Sigma X_1 Y \end{bmatrix} = \begin{pmatrix} N & \Sigma X_1 \\ \Sigma X_1 & \Sigma X_1^2 \end{pmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\begin{bmatrix} A \end{bmatrix} = N\Sigma X_2^2 - (\Sigma X_1)^2$$

To determine β_1 we can replace the middle column with the sums on the right hand of the equations and divide by the determinants

$$\beta_1 = \frac{\begin{pmatrix} N & \Sigma Y \\ \Sigma X_1 & \Sigma X_1 Y \end{pmatrix}}{N\Sigma X_1^2 - (\Sigma X_1)^2}$$

$$= \frac{N\Sigma X_1 Y - \Sigma X_1 \Sigma Y}{N\Sigma X_1^2 - (\Sigma X_1)^2}$$

Therefore, given the model $Y = \beta_0 + \beta_1 X_1 + U$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1$$

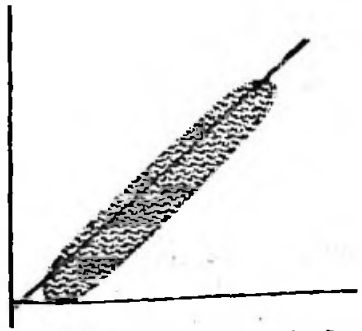
$$\beta 1 = \frac{N\Sigma X_1 Y - \Sigma X_1 \Sigma Y}{N\Sigma X_1^2 - (\Sigma X_1)^2}$$

CHAPTER FOURTEEN

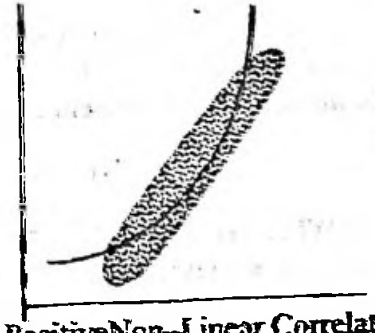
14.1CORRELATION THEORY

Correlation is the technique used to measure the relationship between economic variables. It measures the strength of a linear relationship between economic variables. It is defined as the degree of relationship existing between two or more variables. A correlation between two variables is called simple correlation, while correlation between three or more variables is called multiple correlations. Correlation can be positive or negative.

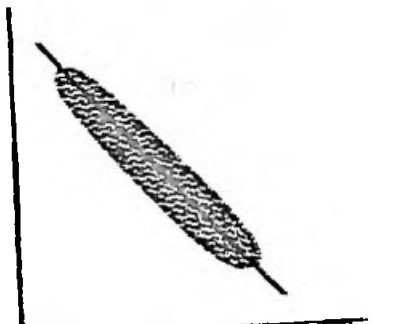
A positive correlation occurs when the variables tend to change together in the same direction. That is when they tend to increase or decrease together. Example the higher the price, the higher the quality demanded and the lower the price the lower quantity demand.



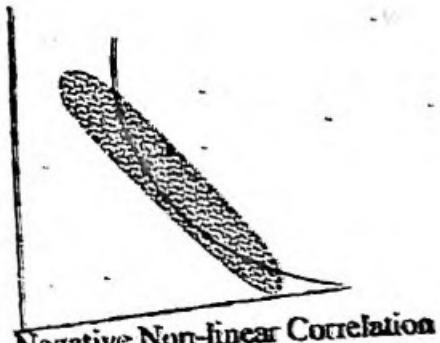
Positive Linear Correlation



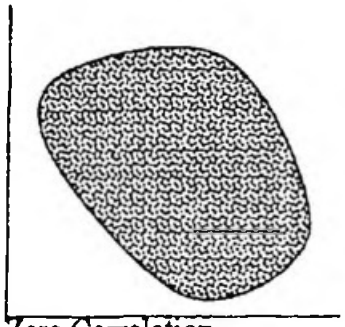
PositiveNon–Linear Correlation



Negative Linear Correlation



Negative Non-linear Correlation



Zero Correlation

Correlation can be linear, non linear or there is no relationship at all. Correlation has serious limitation, because no matter the value of the correlation coefficient, it does not establish causation. However, it is relevant in establishing linear and non linear relationship. Correlation measures only data gotten from a normal distribution. That is in correlation is only useful if data are normally distributed.

For correlation between X and Y the population correlation coefficient is represented by P_{xy} and its sample is estimated by r_{xy} hence the sample coefficient is given by

$$Rxy = \frac{\Sigma xiyi}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}}$$

Where $xi = Xi - Xi$, and $y = Yi - Yi$ these can be refered to as the variation of x and y. The sum of product of x and y is positive for instance, if $(Xi - Xi)(Yi - Yi) = \Sigma xiyi$ the correlation between *x and y* can be said to be positive. If the sum of the product of deviation between x and y is negative that *is* $(xi - xi)(yi - yi) = \Sigma xiyi < 0$ The correlation between *x* and *y* can be said to be negative.

However, this measure of correlation has two defects.

(1) It is affected by the number of observation. The greater the number of observation, the greater the number of product will be, and therefore it affect the sum of *x and y* that is Σxy will be different. The higher the number of observation will make the correlation appear stronger even if it were not supposed to be strong.

The sum of *x and y* is affected by the unit of measurement. Supposing x represent height and y represent weight, because of the unit of measurement, the correlation may be higher when it not true or lower when is not true since height is measured in meter and weight in kilometers.

To correct the first defect, the sum of product of the deviation is divided by the number of observation i.e. $\frac{\Sigma xiyi}{n} = S_{xy}$

Where S_{xy} = covariance of x and y and it is a better measures of correlation because it will not change or vary directly with the number of observations.

The covariance can still be affected by unit of measurement of x and y as stated in the second defect. This can be correlated by dividing the covariance by the standard deviations of the variables. That is

$$\frac{\Sigma xy}{nS_xS_y} = \frac{Sxy}{S_xS_y}$$

Where S_{xy} = covariance of X and Y = $\frac{\Sigma x_iy_i}{}$

S_x = Standard deviation of X_i = $\sqrt{\Sigma(x_i - x)^2/n}$ = $\sqrt{\Sigma x_i^2/n}$

S_y = Standard deviation of Y_i = $\sqrt{\Sigma(y_i - y)^2/n}$ = $\sqrt{\Sigma y_i^2/n}$

Therefore R_{xy} $\frac{\Sigma x_iy_i}{\sqrt{\Sigma x_i^2/n}\,\Sigma y_i^2/n}$ = $\frac{\Sigma x_iy_i}{\sqrt{(\Sigma x_i^2)\ (\Sigma y_i^2)}}$

For actual values of observations

$$R_{xy} = \frac{n\Sigma(X_iY_i) - (\Sigma x_i)(\Sigma y_i)}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

For deviation

$$R_{xy} = \frac{\Sigma x_iy_i}{\sqrt{\Sigma x_i^2}\ \sqrt{\Sigma y_i^2}}$$

Worked Example

Given the following data

$X = 1, 2, 3, 4, 5, 7.$

Y = 8, 10, 12, 11, 12

Where

X = price of a commodity

Y = quantity of commodity

Determine the correlation coefficient and test its statistical significance at 5% level of significance

X	Y	X^2	Y^2	XY
1	8	1	64	8
2	10	4	100	20
4	12	16	144	48
5	11	25	121	55
7	12	49	144	84
19	53	95	573	215

$$Rxy = \frac{N\Sigma(XY) - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$Rxy = \frac{5(215) - (19)(53)}{\sqrt{5(95) - (19)^2}\sqrt{5(573) - (53)^2}}$$

$$= 68/79.89 = 0.85$$

The student t test statistics is used to test the significance or otherwise of sample correlation if the sample size is less than thirty (30)

$$t^* = \frac{r}{\sigma_r} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{r\sqrt{n-2}}{1-r^2}$$

$$r \simeq N(0, \delta) = \sqrt{(1-r^2)/(n-2)}$$

H_0: $t^* < t_T$ or $r = 0$: r is not significant

H_1: $t^* > t_T$ or $r \neq 0$: r is significant

Level of significant – two tailed test)

Critical region: $t^* \leq -t_{0.025}$ or $t^* \geq +t_{0.025}$

Degree of freedom (V) = n – 2 = 5 – 2 = 3

Test statistic = $t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.85\sqrt{5-2}}{\sqrt{1-(0.85)^2}}$

$$= \frac{0.85\sqrt{3}}{\sqrt{1-0.7225}}$$

$$t^* = \frac{0.85(1.75)}{\sqrt{0.2775}} = \frac{1.47}{0.53} = 2.7$$

Decision Rule

If $t^* > t_T$, reject H_0

If $t^* < t_T$, accept H_0

$t_T = 3.182$

since $t^* 2.77 < t_T = 3.182$, we reject H_0 and conclude that P is significant.

X	Y	x $(X - \bar{X})$	x^2 $(X - \bar{X})^2$	y $(Y - \bar{Y})$	y^2 $(Y - \bar{Y})^2$	xy
1	8	-2.8	7.84	-2.6	6.76	7.28
2	10	-1.8	3.24	-0.6	0.36	1.08
4	12	0.2	0.04	1.4	1.96	0.28
5	11	1.2	1.44	0.4	0.16	0.48
7	12	3.2	10.24	1.4	1.96	0.48
19	53		22.8		11.2	13.6

$\bar{X} = \frac{19}{5} = 3.8$

$\bar{Y} = \frac{53}{5} = 10.6$

$\Sigma x_i^2 = 22.8$

$\Sigma y^2 = 11.2$

$\Sigma xy = 13.6$

$r = \frac{\Sigma xy}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y^2}} = \frac{13.6}{\sqrt{22.8}\sqrt{11.2}} = \frac{13.6}{4.78(3.3..)}$

$= \frac{13.6}{15.996...} = 0.85$

14.2 *The Rank Correlation Coefficient*

The correlation coefficient discussed was based on the assumption that the variables are quantitative and in most case they have accurate data for their measurement and such data is normally distributed. In a situation where the variables are qualitative such that it cannot be measured numerically such as religion political affiliation preference for a commodity profession etc. these variable are categorical variable that may not have a precise values, hence can only be measured using another statistic. The rank correlation coefficient (developed by Spearman) is use by ranking the variables in a specific sequence. That is either in order of importance or size using number 1,2,3,.. etc and measure the relationship between the ranks instead of their numerical values.

In ranking the variables, it does not matter if they are ranked in ascending or descending order so long as they apply the same on both sides. If two variables have the same value we assign to their mean rank. This is a non-parametric measure of correlation or association between two economic variables; it is also used when the data are feared not to be normally distributed. That is to say that this measure does not depend on any kind of distribution. The spearman rank correlation coefficient is given by

$$r = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

Where;

D = difference between ranks of corresponding paves of X and Y

n = number of observation

Given the following preference on the commodity X and commodity Y

Time Periods days	Qty Y_i	Price X_i	Rank of Y_i	Rank of X_i	D	D^2
1	10	2	10	10	0	0
2	20	4	9	9	0	0
3	50	6	6.5	8	-1.5	2.25
4	40	8	8	7	0.5	0.25
5	50	10	6.5	6	0	0
6	6	12	5	5	0	0
7	80	14	4	4	0	0
8	90	16	2.5	2	-0.5	0.25
9	90	18	2.5	2	0.5	0.25
10	120	20	1	1	0	0
						4

Therefore, $r = 1 - \frac{6(4)}{10(10^2-1)} = 1 - \frac{24}{990} = 1 - 0.00242$

$= 0.98$

14.3 *Limitations of Analysis of Correlation*

1. Correlation is only a measure of linear relationship. That is, it can only be applied when the relationship is linear. Two variables can be strongly related but may be non-linear.

2. Although, it is a measure of association, it does not imply a functional relationship between the variables. That is, correlation does not imply causal relationship. It does not suggest whether one variable cause the other to change.

CHAPTER FIFTEEN

15.1 SIMPLE REGRESSION MODEL

This is the simplest possible regression analysis because it involves only two variables. It is a situation where the dependent variable is related to a single independent variable. Other names for the dependent and independent variables are

- *Dependent → Independent*
- *Regressand → Regressor*
- *Explained → Explanatory*

For instance, a change in the price of a commodity (independent variable) can explain fluctuations in the quantity demanded for a particular commodity (dependent variable) at a particular time

PURPOSE: The purposes of regression analysis are

(a) To split total variation of a variable (around its mean) into components, attributed to specific (additive) causes (explained and unexplained variation).

(b) To estimate the influence of independent variable on the dependent variable.

(c) To predict the population (mean value) of the dependent variable given some values of the explanatory variables.

15.2 Conditional Model

Suppose we have 50 families who are divided into 10 income groups, say *N100, N200, N300, N400, N500, N600, N700, N800, N900 and N1,000 per week.*

Suppose also that under each income group, we have five different consumption expenditures that depend on the respective incomes.

- *Here, income is the independent variable → denoted as X*

- *Consumption expenditure is the dependent variable → denoted as Y*

Clearly $Y = f(X)$

When we get the mean consumption expenditure corresponding to the weekly level of income of *N100*, we call this value the conditional expected value, the same apply for income level of *N200, N300, N400, etc.*

These mean values are called conditional expected values because they depend on the given values of the (conditional) variables X, i.e. *N100, N200, N300, etc.*

They are denoted as $E(Y/X_i)$ i.e. the expected value of (consumption) Y given the value of income level X.

However, when we get the mean consumption expenditure for all the 50 families, it is unconditional mean, since no income levels: *N100, N200, etc* are dependent variables

It is denoted as $E(Y)$ that is, the expected value of Y. Thus, when we say expected value of Y, we refer to the "Mean value of Y.

When we deal with conditional expected values, $E(Y/X)$ we are dealing with conditional model.

Therefore, it is clear that

$$E(Y/X_i) = f(X_i)$$

This equation is called conditional model, conditional expectation function (CEF) or population regression function (PRF).

The general form of CEF is

$$E(Y/X_i) = \alpha + \beta X_i$$

Which is a linear function of X, where α and β are fixed parameters.

Where α = *intercept* while

β = *slope of the linear CEF*

The regression model can be written as

$$Y_i = E(Y/X_i) + U_i$$

Substituting for $E(Y/X_i)$ we have:

$Y_i = \alpha + \beta X_i + U_i$ this is the actual or true relationship

Where

Y_i = *dependent variable at period i*

X = *independent variable at period i*

U_i = *the deviation or the unobservable random variable (positive or negative) at period i*

It is technically called the stochastic disturbance or error term.

We can see that

$$Y_i = \underbrace{(\alpha + \beta X_i)}_{\text{Systematic or Deterministic component}} + \underbrace{(U_i)}_{\text{Non-systematic or Stochastic component}}$$

It then follows that

$U_i = Y_i - E(Y/X_i)$

The estimated relationship is given as

$$Y_i = \hat{\alpha} + \beta X_i + \hat{e}$$

Where

$\hat{\alpha}$ = *estimate of the true intercept* α

β = *estimate of the true parameter* β

e = *estimate of the true value of* U_i

The estimated regression line is given as

$$Y_i = \hat{\alpha} + \hat{\beta} X_i$$

Since e stand for stochastic error term, it is the difference between the actual value of Y_i and its estimated value, $\hat{Y}_i$, then:

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

We shall use this later to determine the *OLS* equations

3.3 *The underlying assumptions regarding the error term and other assumptions of the ordinary least square (OLS) method regarding the CEF.*

It is clear that the linear classical estimation function CEF is divided into two components, the explained and the unexplained. Therefore, it is important we look at the underlying assumptions in establishing the CEF. We have two forms of assumption, the stochastic assumptions (features of the random variable, u) and the other assumptions

Stochastic Assumptions

1. U_i is a random variable, that is, for each value X_i it can assume any value, positive, negative or zero.
2. Zero mean value of U_i: $E(U_i) = 0$ technically, the conditional mean value of $U_i = 0$ which means U_i does not systematically affect the mean value of Y, this is because the positive values of U_i offset the negative ones so that the average or mean of U_i is zero.
3. U_i is homoscedastic i.e. the variance of U_i is constant. It is means that for all values of X_i, the U_i will show the same dispersion around their mean.

Symbolically, $Var(U_i/X_i) = E[U_i - E(U_i/X_i)]^2$

$= E(U_i^2/X_i)$

$= \sigma^2$

Or $\delta^2 = k$

4. U_i is normally distributed. we can symbolically summarize the above four assumptions as

$U_i \sim N(0, \sigma^2)$

- U_i is a normally distributed random variable with *mean* = zero and constant variance σ^2.

5. There is zero covariance between U_i and X_i of different period.

Symbolically:

$$Cov(U_i X_i) = E[(X - E(X))(U_i - E(U_i))]$$

$$= E[X - E(X)]E(U_i)$$

Since $E(U_i) = 0$

$$= E[X - E(X)](U_i)$$

$$= E[XU_i - E(X)U_i]$$

$$= E(XU_i) - E(X)E(U_i)$$

$$= E(XU_i) - E(X).E(U_i)$$

$$= XE(U_i) = 0$$

$$= X(0) = 0$$

$$= 0, \text{ since } X \text{ is fixed.}$$

There is also zero Covariance between U_i's of different period.

That is they do not affect together, they are independent of one another.

Formally, is the same way

$Cov(U_i U_j) = E[(U_i - E(U_i))(U_j - E(U_j))] = 0$

All errors are only accumulated by U_i, meaning that the explanatory variables are measured without errors.

Non – Stochastic Assumptions

- The regression model is linear in the parameter

  E.g: $Y_i = \alpha_i + \beta X_i + U_i$

- X_i is assumed to be non-stochastic that is, they are fixed in repeated sampling.

- The number of observations n must be greater than the number explanatory variables or parameters.+

- X_i is variable in a given

- sample.

- The regression model or relationship is correctly specified – no specification bias or error in the model

- No perfect multicollinearity among the explanatory variables – no perfect linear correlation.

- The macro-variables are correctly aggregated – individual units are correctly summed up.

Having reviewed the assumptions, we consider the normal equation of OLS.

15.4 The normal OLS equations

First: let us establish the expected value of the dependent variable Y_i and its variance.

Recall: $Y_i = \alpha_i + \beta x_i + U_i$

Then: $E(Y_i) = E(\alpha_i + \beta x_i + U_i)$

$$= E(\alpha_i + \beta X_i) + E(U_i)$$

But by our second stochastic assumption above:

$$E(U_i) = 0$$

Therefore, $E(Y_i) = \alpha_i + \beta X_i + 0$

Hence, $E(Y_i) = \alpha_i + \beta X_i$

This is the mean value of Y

For the variance

$$E(Y_i) = \alpha_i + \beta X_i + U_i$$

$$Var(Y_i) = E[Y_i - E(Y_i)]^2$$

$$But, E(Y_i) = E(\alpha_i + \beta X_i + U_i - \alpha_i + \beta X_i)^2 = E(U_i)^2$$

In our third stochastic assumption assumption $E(U_i)^2 = \delta^2$

Hence, $Var(Y_i) = \delta^2$

Now let's find the *OLS* normal equations:

Recall:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

And $e = Y_i - \hat{Y}_i$

Substituting

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

$$= Y_i - \hat{\alpha} - \hat{\beta} X_i$$

$$e = Y_i - \hat{\alpha} - \hat{\beta} X_i$$

Square both sides

$$e^2 = (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

ake their sums

$$\Sigma e^2 = \Sigma (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

according to one of our stochastic assumption

$$\Sigma e^2 = 0$$

r rather, Σe^2 is the deviation associated with the influence of U_i, the larger s value (variance), the less reliable the value of the parameters or efficients. So, we have to minimize Σe^2.

o do so, we differentiate it with respect to α and β.

$$\Sigma e^2 = \Sigma (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

let $U =$ $(Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$

Since it is a composite function, *(note* $U \neq$ *our* U_i*)*, then

$$\Sigma e^2 = \Sigma (U)^2$$

More precisely: $U = Y_i - \hat{\alpha} - \hat{\beta} X_i$ ...(i)

$$\Sigma e^2 = \Sigma (U)^2 \quad \text{...(ii)}$$

From (i)

$$\delta U / \delta Q = -1$$

From (ii)

$$\delta \Sigma e^2 / \delta U = 2\Sigma U$$

Then,

$$\delta \Sigma e^2 / \delta \alpha = \delta \Sigma e^2 / \delta u \times \delta U / \delta Q$$

$$= -1 \times 2\Sigma U$$

Therefore,

$\delta\Sigma e^2/\delta\alpha = \delta\Sigma e^2/\delta\alpha \times \delta U/\delta Q$

$= -2\Sigma (Y_t - \hat{\alpha} - \hat{\beta}X_t)$

But according to first-order condition

$\delta\Sigma e^2/\delta\alpha = 0$

$$-2\Sigma(Y_t - \hat{\alpha} - \hat{\beta}X_t) = 0$$

Dividing by -2

$$\Sigma(Y_t - \hat{\alpha} - \hat{\beta}X_t) = 0$$

Opening bracket,

$$\Sigma Y_t - \Sigma\hat{\alpha} - \Sigma\hat{\beta}X_t = 0 \qquad \text{... Normal Equation (1)}$$

Applying the usual summation rule Σ = n to α

$$\Sigma Y_t - n\alpha - \Sigma\beta X_t = 0$$

Re-arranging

$$\Sigma Y_t - \Sigma\hat{\beta}X_t = n\hat{\alpha}$$

$$\Sigma Y_t - \hat{\beta}\Sigma X_t = n\hat{\alpha}$$

Dividing through by n

$$\Sigma Y_t/n - \hat{\beta}\Sigma X_t/n = n/n\hat{\alpha}$$

Therefore

$$\hat{\alpha} = \Sigma Y_t/n - \hat{\beta}\Sigma X_t/n$$

But, $\Sigma Y_t/n = \bar{Y}_t$ *and* $\Sigma X_t/n = \bar{X}_t$.

Then $\alpha = Y_t - \beta\Sigma x_t$

In general

$$\hat{\alpha} = \bar{Y} - \beta\bar{X}$$

We also differentiate Σe^2 with respect to $\hat{\beta}$

Given $U = Yt - \hat{\alpha} - \beta X_t$

$\delta U/\delta\beta = -X_t$

And $\Sigma e^2 = \Sigma U^2$

$\delta\Sigma e^2/\delta U = 2\Sigma U$

Similarly, $\delta\Sigma e^2/\delta\beta = \delta U^2/\delta\beta \times \delta\Sigma e^2/\delta u$

$= -X_t$

$\delta\Sigma e^2/\delta\beta = -2X_t \Sigma(Y_t - \hat{\alpha} - \beta X_t)$

and

$\delta\Sigma e^2/\delta\alpha = 0$

$-2X_t \Sigma(Y_t - \alpha - \beta X_t) = 0$

Dividing by -2

$X_t\Sigma (Y_t - \hat{\alpha} - \beta X_t) = 0$

Opening the bracket,

$\Sigma X_t Y_t - X_t\Sigma\hat{\alpha} - \Sigma\hat{\beta}X_t^2 = 0$ *....mormal equation (2)*

Substituting$\alpha = (\bar{Y}_t - \beta\bar{X}_t)$ *to determine the value* β

Then $\Sigma X_i Y_i - \Sigma X_i(\bar{Y}_i - \hat{\beta}\bar{X}_i) - \hat{\beta}\Sigma X_i^2 = 0$

$\Sigma X_i Y_i - \Sigma X_i \bar{Y}_i - \hat{\beta}\Sigma X_i \bar{X}_i - \hat{\beta}\Sigma X_i^2 = 0$

$\Sigma X_i Y_i - \bar{Y}_i \Sigma X_i - \bar{X}_i \hat{\beta}\Sigma X_i - \hat{\beta}\Sigma X_i^2 = 0$

Recall that $\Sigma \bar{Y}_i = n\bar{Y}_i$ and $\Sigma X_i = n\bar{X}_i$ then

Then $\Sigma X_i Y_i - n\bar{Y}_i\bar{X}_i + \hat{\beta}n\bar{X}_i^2 - \hat{\beta}\Sigma X_i^2 = 0$

Collecting like terms in $\hat{\beta}$.

$$\Sigma X_i Y_i - n\bar{Y}_i\bar{X}_i = \hat{\beta}\bar{X}_i^2 - \hat{\beta}\Sigma n\bar{X}_i^2$$

$$\Sigma X_i Y_i - n\bar{Y}_i\bar{X}_i = \hat{\beta}(\bar{X}_i^2 - \Sigma n\bar{X}_i^2)$$

Therefore,

$$\hat{\beta} = \frac{\Sigma X_i Y_i - n\bar{Y}_i\bar{X}_i}{X_i^2 - \Sigma \bar{X}_i^2}$$

Now substituting $\bar{Y}_i = \Sigma Y_i/n$ and $\bar{X}_i = \Sigma X_i/n$.

$$\hat{\beta} = \frac{N\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{N\Sigma X_i^2 - (\Sigma X_i)^2}$$

Therefore

$$\hat{\beta} = \frac{N\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{N\Sigma X_i^2 - (\Sigma X_i)^2}$$

In deviations

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2}$$

And

$$\hat{\alpha} = \bar{Y}_i - \hat{\beta}\bar{X}_i$$

Note that

$\Sigma x^2 = \Sigma X^2 - n\bar{X}^2$ and $x = X - \bar{X}$

$$\Sigma y^2 = \Sigma Y^2 - nY^2 \qquad y = Y - Y$$
$$\Sigma xy^2 = \Sigma XY - nXY^2$$

It should also be noted that in demand function, the elasticity can be derived from the estimated regression line.

Given

$\hat{Y}_i = \hat{\alpha} - \hat{\beta}\hat{X}_i$

$\delta\hat{Y}/\delta X_i = \hat{\beta}$

By definition, price elasticity is given as

$S/p = \frac{\delta Y}{\delta X} \cdot X/Y$

Where X = price, Y = quantity and $\delta Y/\delta X = \hat{\beta}$

Then $s/p = \hat{\beta}.X_i/\bar{Y}_i$

$\bar{\hat{Y}}_i$ = *mean value of the estimated quantity*

$\bar{X}_i$ = *the average price in the sample.*

But $\bar{\hat{Y}}_i = \bar{Y}_i$

Therefore

$$S/p = \hat{\beta}.\bar{X}_i/\bar{Y}_i$$

More precisely, according to the OLS method, the best way to fit a curve in a set of points is to find the line whose sum of square deviation is minimum. The method outline the following steps

(a) *Find the error; if* $Y = \beta o + \beta_1 X_1 + U$, *then the error* $U = Y - \beta o - \beta_1 X_1$

(b) *Sum the error, that is* $\sum U = \sum(Y - \beta o - \beta_1 X_1)$

(c) *Square the sum of the error, that is* $\sum U^2 = \sum(Y - \beta o - \beta_1 X_1)^2$

(d) Then minimize the error, by taking partial derivative in respect of βo and β_1.

These estimate or coefficient can be obtained by minimizing the sum of squared residuals. i.e $\sum e_i^2 = \Sigma(Y_i - Y_i)^2 = \Sigma(Y - \beta_0 - \beta_1 X_1)^2$

We obtain by

(4) Finding the error, (in our model)

$U_1 = Y - \hat{\beta}_0 - \hat{\beta}_1 X_1$

(5) Squaring the error

$$U_i^2 = (Y - \beta_0 - \beta_1 X_1)^2$$

(6) Summing the squared error

$$\Sigma U_i^2 = \Sigma (Y - \beta_0 - \beta_1 X_1)^2$$

To minimize we take partial derivation in respect of β_0, and β_1 respectively so as to get our normal equation and subsequently, the coefficient β_0, and β_1 i.e. take partial derivation and equate it to zero.

$$\frac{d\Sigma Ui^2}{d\beta_0} = (-1)(2)\,\Sigma(Y - \beta_0 - \beta_1 X_1) = 0$$

$$= -2\,\Sigma (Y - \beta_0 - \beta_1 X_1) = 0$$

Divide through by -2

$$= \frac{-2\,\Sigma (Y - \beta_0 - \beta_1 X_1)}{-2} = \frac{0}{-2}$$

$$= \Sigma (Y - \beta_0 - \beta_1 X_1) = 0$$

Where $\frac{0}{-2} = 0$

Open the bracket

$$= \Sigma Y - \Sigma\beta_0 - \Sigma\beta_1 X_1 = 0$$

$$= \Sigma Y - N\beta_0 - \beta_1 \Sigma X_1 = 0 \quad \text{.....(1) Normal equation}$$

Where $\Sigma \beta_0 = \beta_0 + \beta_0 + \beta_0 + \text{.....} = N\beta_0$ *i.e. Number of* β_0

To find our β_0, we multiply through by N and make β_0 the subject of the formula.

$$\frac{\Sigma Y}{N} = \frac{N\beta_0}{N} + \frac{\beta_1 \Sigma X_1}{N}$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}_1$$

Therefore

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1$$

To get the 2nd normal equation we take partial derivation in respect of the β_1

$$d\Sigma Ui^2 = (-X_1)\,(2)\,\Sigma(Y - \beta_0 - \beta_1 X_1) = 0$$
$$d\beta_1$$

$$= (-2X)\,\Sigma\,(Y - \beta_0 - \beta_1 X_1) = 0$$

To eliminate -2 we divide through by -2

$$\frac{(-2X)\Sigma(Y - \beta_0 - \beta_1 X_1)}{-2} = \frac{0}{-2}$$

$$X_1\Sigma\,(Y - \beta_0 - \beta_1 X_1) = 0$$

$$\Sigma X_1 Y - \beta_0 \Sigma X_1 - \beta_1 \Sigma X_1^2 = 0$$

$$\Sigma X_1 Y - \beta_0 \Sigma X_1 - \beta_1 \Sigma X_1^2 \quad \text{....(2) normal equation}$$

$$\beta_0 = \Sigma Y = N\beta_0 + \beta_1 \Sigma X_1 \quad \text{...(1)}$$

$$\beta_1 = \Sigma X_1 Y = \beta_0 \Sigma X_1 + \beta_1 \Sigma X_1^2 \quad \text{...(2)}$$

These equations are called normal equations and are used to obtain the coefficient β_0 and β_1 in other words they are used to fit the regression line. To do that we use crammer rule matrix as follows

Given the above *OLS* estimators we need to also obtain the standard error of these estimates which are important to establish confidence interval and to test the statistical hypothesis. The relevant formulas are as follows

$$[A] = \begin{bmatrix} \Sigma Y \\ \Sigma X_i Y \end{bmatrix} \quad \begin{bmatrix} N\Sigma X_i & \beta_0 \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix} = \begin{bmatrix} \beta_1 \end{bmatrix}$$

$$[\,] = N\Sigma X_2^2 - (\Sigma X_i)^2$$

To determine β_1 we can replace the middle column with the sums on the right hand of the equations and divide by the determinants

$$\hat{\beta}_1 = \frac{\begin{vmatrix} N & \Sigma Y \\ \Sigma X_i & \Sigma X_i Y \end{vmatrix}}{N\Sigma X_i^2 - (\Sigma X_i)^2}$$

$$= \frac{N\Sigma X_i Y - \Sigma X_i \Sigma Y}{N\Sigma X_i^2 - (\Sigma X_i)^2}$$

Given the above OLS estimates, we need to also obtain the standard error of these estimates which are important to establish confidence interval and to test the statistical hypothesis. The relevant formulas are as follows

β_1 is the linear estimator or weithed average of Yi with ki serving as its weith. The properties of ki are: 1. the ki are stochastic, like x, 2. $\Sigma ki = 0$, 3. $\Sigma ki^2 = 1/\Sigma xi^2$,

4. $\Sigma k_i x_i = \Sigma k_i X_i = 0$

15.5 *Total, explained and residual sums of sqares*

Recall:

$Y_i = \alpha + \beta X_i + U_i$ → *Actual relationship*

$Y_i = \hat{\alpha} + \hat{\beta} X_i + e$ → *estimated relationship*

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ → *estimated regression line*

- Total derivation in Y_i from its mean is given as $\bar{Y}_i$

  $y = Y_i - \bar{Y}_i$

  This is broken into

- Explained Derivation/variation, which is the deviation of the regressed value from the meun.

  It is given by

  $\hat{y} = \hat{Y}_i - \bar{Y}_i$

- Residual/unexplained variation which is the deviation of the observation Y_i from the mean.

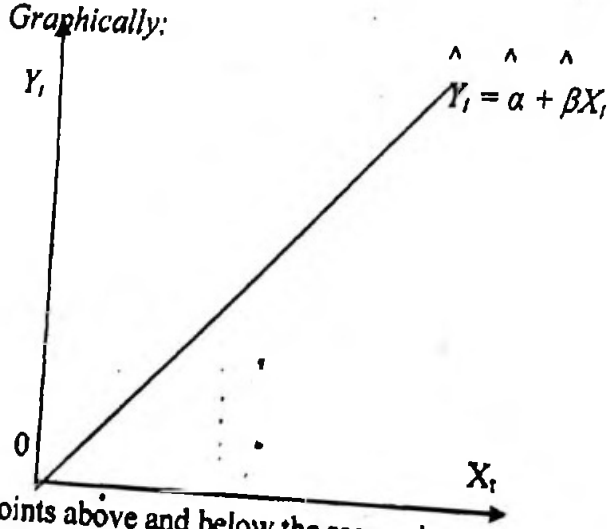
  It is given by

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

  $\hat{y}$ is the part of total variation which is the derivation of the observation Y_i which is explained by the regression line.

  e is that part of total variation of Y_i which is attributed to the random variable U_i.

Graphically:



Points above and below the regression line are expected by e.
It is clear that:
$y_i = \hat{Y}_i - \hat{e}$

We now find the value of $\hat{y}_i$
Recall that:
$\hat{y} = \hat{Y}_i - \overline{Y}_i$

but $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$

and $\overline{Y}_i = \hat{\alpha} + \hat{\beta}\overline{X}_i$

(From the first normal OLS equation)

Substituting:

$\hat{y}_i = \hat{\alpha} + \hat{\beta}X_i - (\hat{\alpha} + \hat{\beta}\overline{X}_i)$

$\hat{y}_i = \hat{\alpha} + \hat{\beta}X_i - \hat{\alpha} - \hat{\beta}\overline{X}_i$

$\hat{y}_i = \hat{\beta}X_i - \hat{\beta}\overline{X}_i$

Factorizing

$\hat{y}_i = \hat{\beta}(X_i - \overline{X}_i)$

But, $X_i - \overline{X}_i = x$

Therefore

$\hat{y}_i = \hat{\beta}X_i$

Now, the total deviation is

$y = \hat{y}_i + e$

Taking the sum of their squares

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e \qquad \text{... (1)}$$

$\sum y_i^2$ = *TSS*; $\sum \hat{y}_i^2$ = *ESS*; $\sum e$ = *RSS*

This can also be written as

$$\sum (Y_i - \overline{Y}_i)^2 = \sum (\hat{Y}_i - \overline{Y}_i)^2 + \sum (Y_i - \hat{Y}_i)$$

Where; *TSS = Total Sum of Square*

ESS = Explained Sum of Square

RSS = Residual Sum of Square

Now substitute the value $\hat{y}_i = \hat{\beta}x_i$ in equation (1)

$$\sum Y_i^2 = \sum (\hat{\beta}X_i)^2 + \sum e \quad \text{... (2)}$$

We want to know the *ESS* and *RSS* as a percentage of *TSS*

ESS/TSS = Explained Variation/Total Variation

$$= \frac{\sum (\hat{\beta}X_i)^2}{\sum Y_i^2}$$

$$= \frac{\sum(\beta)^2 (X_i)^2}{\sum Y_i^2}$$

But recall that $\beta = \frac{\sum xy}{\sum x^2}$

Therefore $\frac{ESS}{TSS} = \frac{(\beta)^2 \sum (X_i)^2}{\sum x^2}$

$$= \frac{[\sum xy]^2}{\sum x^2} \cdot \frac{\sum (X_i)^2}{\sum Y_i^2}$$

$$\frac{ESS}{TSS} = \frac{\sum (\beta x)^2}{\sum y^2}$$

Therefore $\frac{\sum y^2}{\sum y^2} = \frac{ESS}{TSS} = \frac{\sum (xy)^2}{\sum x^2 \cdot \sum y^2} = \frac{\sum (\beta x_i)^2}{\sum y^2}$

Recall that,

The $r^2 = \frac{\sum (xy)^2}{\sum x^2 \cdot \sum y^2} = \frac{\sum y^2}{\sum y^2}$

Therefore r^2 = is the explained variation expressed as a percentage of total variation

r^2 = co-efficient of determination, which is the test of the goodness of fit of the regression line. Hence the better the goodness of fit the closer the observation to the regression line,

r^2 shows the percentage change in Y_i that is brought about by the explanatory variable X_i. The closes it is to 1, the better the goodness of fit and vice versa.

Symbolically $0 \leq r^2 \leq 1$

From (2)

$$\sum Y_i^2 = \sum (\beta X_i)^2 + \sum e$$

To find the remaining/residual/unexplained variation in terms of r, divide through by $\sum y_i$

$$\frac{\sum y_i^2}{\sum y_i^2} = \frac{\sum (\beta x_i)^2}{\sum y_i^2} + \frac{\sum e^2}{\sum y_i^2}$$

$$1 = r^2 + \sum e^2 / \sum y_i^2$$

Re arranging

$$\sum e^2 / \sum y_i^2 = 1 - r^2$$

$$\sum e^2 = \sum y_i^2 (1 - r^2)$$

Hence, $\sum e^2 = \sum y_i^2 (1 - r^2)$

Conversely, if we know $\sum e^2 / \sum y_i^2$, we can find r^2 as follows

$$r^2 = 1 - \sum e^2 / \sum y_i^2$$

Based on this, we deduce that:

- *If* $\sum e^2 / \sum y_i^2 = 0$, $r^2 = 1$; *this implies that all the observations lie on the regression line and there will be no scatter of points. This means that* Y_i *is perfectly explained by* X_i
- *If* $r^2 = 0$, $\sum e^2 / \sum Y_i^2 = 1$; *this means that variations in* Y_i *are not at all explained by* X_i
- *If* $r^2 < 1$, $\sum e^2 / \sum y_i^2 > 0$ *but* < 1, *this means that some parts of the variations is explained by the regression line and some part is unexplained.*

It should be noted that r^2 is very much related to the slope of the regression line, β

Recall that: $r^2 = \frac{(\sum xy)^2}{\sum x^2 \cdot \sum y^2}$

$$= \frac{(\sum xy)(\sum xy)}{}$$

$$= \frac{\sum x^2 \sum y^2}{\sum xy \cdot \sum xy} = \frac{\sum xy}{\sum x} \quad \frac{\sum xy}{\sum y}$$

But, $\beta = \frac{\sum xy}{\sum x^2}$

Therefore, $r^2 = \beta_1 \frac{\sum xy}{\sum y^2}$

The r^2 is the coefficient of determination of a simple regression Given the above OLS estimates, we need to also obtain the mean, variances, and standard error of these estimates which are important to establish confidence interval and to test the statistical hypothesis. The relevant formulas are as follows

15.6 Mean and variance of OLS estimates

Mean of $\hat{\beta}$

If all the possible sample of size n are taken from the population of X_i and Y_i, then the mean value of β will be its expected value: Mean $\beta = E(\beta)$

Recall: $\beta = \frac{\sum xy}{\sum x^2}$

But, $y = Y - \bar{Y}$

Therefore, $\hat{\beta} = \frac{\sum xY - \bar{Y}}{\sum x^2}$

$$\hat{\beta} = \frac{\sum x(Y - \bar{Y})}{\sum x^2}$$

$$\hat{\beta} = \frac{\sum xY - \bar{Y}\sum x}{\sum x^2}$$

$$= \frac{Y\sum x}{\sum x^2} - \frac{\bar{Y}\sum x}{\sum x^2}$$

Taking the second part $\frac{\bar{Y}\sum x}{\sum x^2}$ *and by definition* $\sum x = 0$

Therefore, $\frac{\bar{Y}\sum x}{\sum x^2} = 0$

Or rather $\frac{\bar{Y}\sum x}{\sum x^2} = \frac{Y\sum(X - \bar{X})}{\sum X^2} = \frac{\sum XY - \sum X\bar{Y}}{\sum X^2}$

But since $n = \sum$ *and* $\sum X = nX$,

$$\frac{Y\sum x}{\sum x^2} = \frac{n\bar{X}\bar{Y} - n\bar{X}\bar{Y}}{\sum X^2} = \frac{0}{\sum X^2}$$

Therefore, $\frac{\bar{Y}\sum x}{\sum x^2} = 0$

We are left with

$\frac{\sum xY}{\sum x^2} = \sum k_i Y$ *since* $k = \frac{x_i}{\sum x^2}$

We now take the expected value of $\hat{\beta}$

$E(\hat{\beta}) = E(\sum k_i Y_i)$

But $Y_i = \alpha + \beta X_i + U_i$

$E(\hat{\beta}) = E[(\sum k_i)(\sum Y_i)]$

$= E[(\sum k_i)(\alpha + \beta X_i + U_i)]$

$= E[\alpha \sum k_i + \sum k_i \beta X_i + \sum k_i U_i]$

$= E[\alpha \sum k_i + \beta \sum k_i X_i + \sum k_i U_i]$

but originally, $\sum k = 0$

Since, $\frac{\sum x_i}{\sum x^2} = 0$

i.e. $\frac{\sum x}{\sum x^2} = \frac{\sum(\bar{X} - X)}{\sum x} = \frac{(\sum \bar{X} - \sum X)}{\sum x^2}$

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

But $\sum k_i x_i = 1$, since $\sum k_i X_i = \frac{\sum x}{\sum x^2} X$

Because $\frac{\sum(X - \bar{X})}{\sum x^2} = \frac{\sum \bar{X} - \sum X\bar{X}}{\sum X}$

$$= \sum \bar{X}^2 - \frac{n\bar{X}\bar{X}}{\sum X^2}$$

$$= \frac{\sum X^2 - n\bar{X}^2}{\sum X^2 - n\bar{X}^2} = 1$$

Therefore,

$$E(\beta) = E[\beta(0) + \beta(1) + \sum k_i U]$$

$$= E[\beta + \sum k_i E(U)]$$

But by one of our assumption $E(U) = 0$

Therefore

$E(\hat{\beta}) = E\beta + \sum k_i(0)$

$E(\hat{\beta}) = E\beta$

Hence

$E(\hat{\beta}) = \hat{\beta}$

This means that the *OLS* estimate $\hat{\beta}$ is equal to true value of the population parameter β. Hence *OLS* is unbiased.

The Variance of $\hat{\beta}$

By definition $Var(\hat{\beta}) = E[\hat{\beta} - E(\hat{\beta})]^2$

But since OLS is unbiased, we use.

$Var(\hat{\beta}) = Var(\sum k_i Y)^2$

$= \sum k_i^2 Var(Y)$

And $Var(Y) = E[Y - E(Y)]^2$

And also $Y = \alpha + \beta X - U$, *while*

$E(Y) = \alpha + \beta X$

Therefore, $Var(Y) = E[\alpha + \beta X + U - (\alpha + \beta X)]^2$

$= E[\alpha + \beta X + U - \alpha - \alpha X)]^2$

$= E(U)^2$

Therefore, $Var(\beta) = \sum k_i^2 Var(Y)$

$= \sum k_i^2 E(U)^2$

But $E(U)^2$ *is given as* $\delta^2 = \sum e^2/n-K$

And $\frac{\sum k_i^2}{\sum x^2} = \frac{(\sum x)^2}{\sum x^2 \sum x} = \frac{(\sum x)^2}{(\sum x)^2}$

$= 1/\sum x^2$

Therefore, .

$Var(\hat{\beta}) = \sum k_i^2 \delta^2 E(U)^2$

$= 1/\sum x^2 . \sum e^2/n-K$

Hence,

$Var(\hat{\beta}) = 1/\sum x^2 . \sum e^2/n-K$

The Mean of α

Recall $\hat{\alpha} = \overline{Y} - \beta \overline{X}$

Putting $\beta = \sum k_i Y$

$$\hat{\alpha} = \overline{Y} - \sum k_i \overline{Y} \overline{X}$$

We have $\hat{\alpha} = \overline{Y} - \overline{X} \sum k_i Y$

$$\hat{\alpha} = \sum (1/n - \overline{X} k_i) Y \qquad \text{since } Y = \sum Y/n$$

By factorizing $\sum$ *and* Y.

Taking the expected value

$E(\hat{\alpha}) = E(Y) \sum (1/n - \overline{X} k_i)$

$= \sum (1/n - \overline{X} k_i) (\alpha + \beta x) \qquad \text{since } E(Y) = \alpha + \beta x$

Opening the bracket

$E(\hat{\alpha}) = \sum (\alpha/n - \alpha \overline{X} \sum k_i + \beta x/n - \beta \overline{X} \sum k_i)$

$E(\hat{\alpha}) = \sum \alpha/n - \alpha \sum \overline{X} \sum k_i + \sum \beta x - \beta \sum \overline{X} \sum k_i X$

$E(\hat{\alpha}) = \sum \alpha/n - \alpha \sum \overline{X} \sum k_i + \beta \sum x - \beta \sum \overline{X} \sum k_i X$

Taking $\sum = n$, $\sum x = n\overline{X}$ and knowing that $\sum k_i = 0$ *and* $\sum k_i X = 1$

$E(\hat{\alpha}) = \frac{n\alpha}{n} - \alpha \sum \overline{X}(0) + \beta n \overline{X} - \beta n \overline{X}(1)$

$= \alpha - 0 + \beta n \overline{X} - \beta n \overline{X}$

$= \alpha - 0 + 0$

Hence.

$E(\hat{\alpha}) = \alpha$

This means that the *OLS* estimate α is exactly equal to the true value of the population parameter α

The Variance of α

$Var(\hat{\alpha}) = E.[\hat{\alpha} - E(\hat{\alpha})]^2$

But $E(\alpha) = \alpha$

Therefore

$Var(\hat{\alpha}) = E[\hat{\alpha} - \alpha]^2$

Recall that:

$\hat{\alpha} = \overline{Y} - \beta \overline{X}$. *and* $\beta = \sum k_i Y$

Therefore.

$\hat{\alpha} = \overline{Y} - \sum k_i Y \overline{X}$

$= \frac{\sum Y}{n} - \frac{\sum k_i Y \overline{X}}{1}$

Factorizing Y *and* $\sum$

$= \sum [1/n - \sum X/Y]$

Taking the variance

$Var(\hat{\alpha}) = Var[\sum (1/n - \overline{X}) Y]$

$Var(\hat{\alpha}) = [\sum (1/n - k_i \overline{X})]^2 Var(Y)$

$= \sum (1/n - k_i \overline{X}) (1/n - k_i \overline{X}) . Var(Y)$

$= \sum[1/n^2 - k_i\overline{X}) - k_i\overline{X} + k_i^2\overline{X^2}].Var(Y)$

$= \sum[1/n^2 - 2\overline{X}\sum k_i + \sum k_i^2\overline{X^2}].Var(Y)$

$= 1/n^2 - 2\overline{X}(0) + (\sum x/\sum x^2)\overline{X^2}].Var(Y)$

$= 1/n - (1/\sum x^2)\overline{X^2}\ .\ Var(Y)$

$= 1/n - \overline{(X^2}/\sum x^2)\ .\ Var(Y)$

$= \frac{\sum X^2 + n\overline{X^2}}{n\sum X^2}.\ Var(Y)$

But, recall also that, $\sum x^2 = \sum X - n\overline{X^2}$

$= \frac{\sum\overline{X^2} - n\overline{X^2} + n\overline{X^2}}{n\sum X^2}.\ Var(Y)$

$= \frac{\sum X}{n\sum X^2}\ .\ Var(Y)$

But $Var(Y) = \delta^2 = \frac{\sum e^2}{n-K}$

Therefore,

$Var(\hat{\alpha}) = \frac{\sum X}{n\sum X^2}\ .\ \frac{\sum e^2}{n-K}$

Hence,

$Var(\hat{\alpha}) = \frac{\sum X^2}{n\sum x^2}\ .\ \frac{\sum e^2}{n-K}$

These mean values and variances lead us to another aspect of econometric

15.7 Test of significance of the estimates

We can test hypothesis by

Simple Test: the rule of thumb or the standard error test.

We compare $s(\hat{\alpha})$ *and* $s(\hat{\beta})$ *with* $\hat{\alpha}$ *and* $\hat{\beta}$. *In terms of* $s(\hat{\alpha}) \rightarrow \hat{\alpha}/2$ *and* $s(\hat{\beta}) \rightarrow \hat{\beta}/2$.

Our hypothesis

H_0: $\hat{\alpha}\ \hat{\beta} = 0$

H_1: $\hat{\alpha}\ \hat{\beta} \neq 0$

We accept H_0 *if:*

$\hat{\alpha}/2 < s(\hat{\alpha})$ *or* $s(\hat{\alpha}) > \hat{\alpha}/2$

and

$\hat{\beta}/2 < s(\hat{\beta})$ *or* $s(\hat{\beta}) > \hat{\beta}/2$

And conclude that and α and $\hat{\beta}$ are statistically insignificant in explaining the relationship between *X and Y*.

We reject H_0 if:

$\hat{\alpha}/2 > s(\hat{\alpha})$ *or* $s(\hat{\alpha}) < \hat{\alpha}/2$

And if,

$\hat{\beta}/2 > s(\hat{\beta})$ *or* $s(\hat{\beta}) < \hat{\beta}/2$

We conclude that $\hat{\alpha}$ *and* $\hat{\beta}$ are significant.

Here the t-statistic (calculated) is reduced to

$t_{cal}(\hat{\alpha}) = \hat{\alpha}/s(\hat{\alpha})$ note that $s(\hat{\alpha}) = \sqrt{var(\hat{\alpha})}$

And

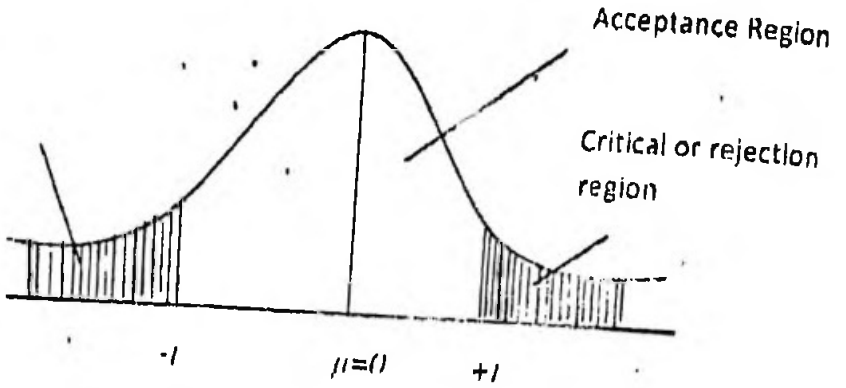
$t_{cal}(\hat{\beta}) = \hat{\beta}/s(\hat{\beta})$ And the null hypothesis

H_0: $\hat{\alpha}\ \hat{\beta} = 0$

Is tested against the alternative hypothesis

$H_1: \hat{\alpha}\ \hat{\beta} \neq 0$

Standard error test is a two-tailed test graphically shown as



The calculated t is compared with the theoretical t which is found from the t-table at *n-k* degree of freedom

Where *n = size of the sample*

K = number of parameters

The confidence interval is usually *99%* or *95%* which is *1% or 5%* significance level.

Since it is two-tailed test 1% and 5% have to be divided by 2.

i.e. 1% = 1%/2 = 0.005

or 5% = 5%/2 = 0.025

i.e. lower and lower respectively.

- **The Decision Rule**

According to rule of thumb, we

Reject H_0: $\hat{\alpha}\ \hat{\beta} = 0$, if $t_{cal} > t$

- *Accept H_1: $\hat{\alpha}\ \hat{\beta} = 0$, if $t_{cal} < t$*

15.8 The Gauss Markov Theorem

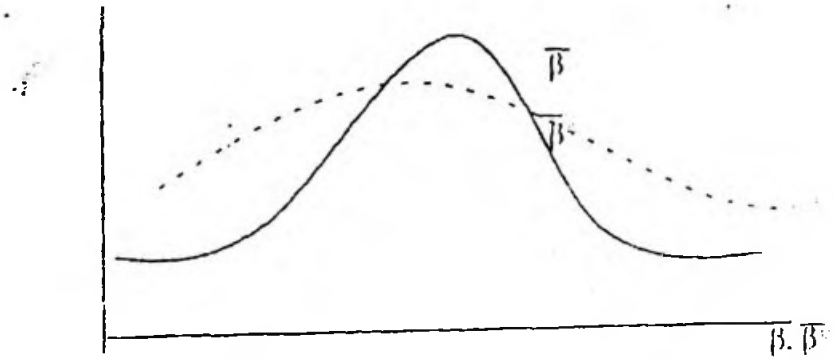
The Gauss Markov Theorem captures the basic properties of OLS estimates. The theorem states that given the assumption of the classical linear regression model, the least –square estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

This means that estimators $\hat{\alpha}$, $\hat{\beta}$, etc are said to be BLUE of $\hat{\alpha}$, $\hat{\beta}$, etc. if the following hold

- *It is linear → linear function is a random variable, such as the dependent variable Y_i in the regression model.*
- *It is unbiased, that is, its average or expected value $E(\beta)$, is equal to the BLUE value of β*
- *It has minimum variance in the class of such linear unbiased estimators.*
- *unbiased estimators with the least variance is known as an efficient estimator*

Note that

Minimum variance of β means that β is more likely to be closer to β it is depicted below.



It shows that the distribution of β^* is more widespread around the mean (longer variance) than that of β.

An estimator is sufficient if it utilizes all the information a sample contains about the true parameters. An *OLS* estimator is also sufficient.

Alternatively, the assumptions or properties of the OLS can be stated as follows:

1. It is a conditional model such that $E(Y/X) = \alpha + \beta T + U$

   This means that the expected value of Y is conditioning upon the value of X

2. *U's are iind* ~ $(0,\delta^2)$

   Where I = identical

   I = independent

   Nd = normally distributed

   Implying

   $E(U_i) = 0$
   $E(U_i)^2 = \delta^2$
   $E(U_i\ U_j) = 0$
   } "White Noise assumptions"

   All of which have already been stated under the stochastic assumptions.

3. $E(U_i\ U_j) = 0$ *This is also called the fixed regressor assumption, which means that* U_i *and* X_i *are correlated.*

When these assumptions hold also, $\hat{\alpha}$ *and* $\hat{\beta}$ *are said to be best, linear unbiased estimators (BLUE) as envisaged by Gauss Markov Theorem.*

15.9 Summary of OLS Simple regression model

Let us summarize the procedures of OLS estimation

Obtain α *and* $\hat{\beta}$ *using:*

$$\hat{\alpha} = \bar{Y}_i - \hat{\beta}\bar{X}_i$$
$$\hat{\beta} = \Sigma x_i y / \Sigma x_i^2$$

- *Obtain* r^2 *as*

$$r^2 = \hat{\beta}\Sigma xy/\Sigma y^2 = (\Sigma xy)^2/\Sigma x^2.\Sigma y^2$$

- *Obtain* Σe^2 *as*

$$\Sigma e^2 = \Sigma y^2 (1-r^2)$$

- *Obtain* δ^2 *as*

$$\delta^2 = \Sigma e^2 / n-K \text{ where } K = 2$$

- Obtain Standard errors of $\hat{\alpha}$ and $\hat{\beta}$ as

$$s(\hat{\alpha}) = \sqrt{\Sigma X^2/n\Sigma x^2}\ .\ \delta^2$$

$$s(\hat{\beta}) = \sqrt{1/\Sigma x^2}\ .\ \delta^2$$

- Obtain t calculated as

$$t_{cal(\alpha)} = \alpha/s(\alpha)$$
$$t_{cal(\beta)} = \beta/s(\beta)$$

Which are compared with t theoretical at a given level of significance with

H_0: $\hat{\alpha}$ $\hat{\beta} = 0$

H_1: α $\beta \neq 0$

H_0 *is accepted when* $t_{cal}(\alpha\ \beta) < t^*$

H_0 *is rejected when* $t_{cal}(\hat{\alpha}\ \hat{\beta}) < t^*$

The following computational expressions are also useful

$$\Sigma x^2 = \Sigma X^2 - n\bar{X}^2$$

$\sum x^2 = \sum Y^2 - n\bar{Y}^2$

$\sum xy = \sum XY - n\bar{X}\bar{Y}$

Worked Example:

Ex(1): Given Y = quantity supply and X = price

$Y = 3,5,3,4,6,8,8,6,9,10,11$

$X = 2,4,3,5,5,7,6,4,7,9,8$

Estimate the linear regression model of the above information and conduct the r^2 and standard error test

Solution

X	Y	X^2	Y^2	XY
2	3	4	9	6
4	5	16	25	20
3	3	9	9	9
5	4	25	16	20
5	6	25	36	30
7	8	49	64	56
6	8	36	64	48
4	6	16	36	24
7	9	49	81	63
9	10	81	100	90
8	11	64	121	88
60	73	374	561	454
$\sum X$	$\sum Y$	$\sum X^2$	$\sum Y^2$	$\sum XY$

$\bar{X} = \sum X/n = 60/11 = 5.43$

$\bar{Y} = \sum Y/n = 73/11 = 6.64$

Recall that

$Y = \alpha + \beta X + U$

$\hat{Y} = \hat{\alpha} + \hat{\beta} X$

$\hat{\beta} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - nX^2} =$

$= \frac{459 - 11(5.45)(6.64)}{374 - 11(5.45)^2}$

$= 1.183$

$\hat{\beta} = 1.183$

$\hat{\alpha} = \bar{Y} + \hat{\beta}\bar{X}$

$= 6.64 - (1.183)(5.45)$

Therefore,

$\hat{\alpha} = 0.1913$

The regression line is, therefore

$\hat{Y} = 0.1913 + 1.18X$

Interpretation: On a priori bases $\hat{\alpha}$ and β have conform to a priori expectation this is because both of them are positively signed.

The value of $\alpha = 0.1913$, indicates that in the absence of the independent variable, quantity supply remain at 0.1913.

However, the value of $\beta = 1.183$ indicates that, if price change by one unit quantity supply tend to increase by 1.183

To find r^2

$r^2 = (\sum xy)^2 / \sum x^2 . \sum y^2)$

$= \hat{\beta} . \sum xy^2 / \sum y^2$

But $\Sigma xy = \Sigma XY - n\bar{X}\bar{Y} = 484 - 11(5.45)(6.69) = 55.93$

$\Sigma y^2 = \Sigma Y^2 - \Sigma \bar{Y}^2$

$= 561 - 11(6.64)^2 = 76.01$

Therefore,

$r^2 = \hat{\beta} \cdot \Sigma xy^2 / \Sigma y^2$

$$1.183 \times \frac{55.93}{76.01}$$

$= 0.8706$

Therefore $r^2 = 0.87$

This implies that the equation has a strong power of explaining Y, *since 87% variation in Y is brought about by change in X.*

To obtain Σe^2

$\Sigma e^2 = \Sigma y^2 (1 - r^2)$

$= 76.01 (1 - 0.87)$

Therefore $\Sigma e^2 = 9.88$

To find δ^2

$$\delta^2 = \frac{\Sigma e^2}{n-K} = \frac{9.88}{11-2} = 1.098$$

Therefore, $\delta^2 = 1.1$

The standard error

$$s(\hat{\alpha}) = \sqrt{\Sigma X^2 / n\Sigma x^2 \cdot \delta^2} = \sqrt{\frac{374}{(11 \times 47.27)} \times 1.098} = 0.89$$

$$s(\beta) \sqrt{1/\Sigma x^2 \cdot \delta^2} = \sqrt{1/47.27 \times 1.098} = 0.15$$

We can summarize the results as below

$\hat{Y} = 0.1913 + 1.18X$
$(0.89) \quad (0.15)$
$n = 11, \quad k = 2$

Testing at 99% confidence interval:

$t_{cal(\hat{\alpha})} = \hat{\alpha}/s(\hat{\alpha}) = 0.1913/0.89 = 0.215$

$t_{cal(\hat{\beta})} = \hat{\beta}/s(\hat{\beta}) = 1.18/0.15 = 7.87$

The degree of freedom is $n-k = 11-2 = 9$

From the t-table, $t_{0.005} = 3.25$

Now:

Reject H_0: $\hat{\alpha}$, $\hat{\beta} = 0$

Accept H_0: $\hat{\alpha}$, $\hat{\beta} \neq 0$

We can see *that* $t_{cal(\hat{\alpha})} < t_{0.005}$

We therefore accept H_0 and conclude that $\hat{\alpha}$ is statistically insignificant in explaining Y and

$t_{cal(\hat{\beta})} > t_{0.005}$

We therefore reject H_0 and conclude that $\hat{\beta}$ is significant in explaining Y

CHAPTER SIXTEEN

SOME ASPECTS OF SIMPLE REGRESSION MODEL

16.1 Time as a Regressor; the Linear Trend Model

Most economic variables exhibit trend, that is, they are affected by the passage of time, or they change with time. In this situation, time is an important variable – regressor – that determine the expected value of the dependent variable.

When Y is regressed on time, T, the model is called a *Linear Trend Model*, and the time variable T is known as the *Trend Variable*. The linear trend model can be formulated as

$$Y_t = \alpha + \beta T - U_t$$

As in the CEF $\hat{\beta}$ stands for the slope which determines the direction of the trend.

- *If $\hat{\beta}$ is positive, there is an upward trend in Y → Y increases as time goes on*
- *If $\hat{\beta}$ is negative, there is a downward trend in Y → Y decreases as time goes on.*

The time variable (t) can be used as describe time period → *1, 2, 3, 4, 5,.....n*. It can also be used as years eg: *1990, 1991, 1992..... 2010*

What is important is that, in linear trend model, time is considered as the regressor – the main determinant or cause of change in the regress and Y.

Linear trend model is also like *CEF*, so we can write it as

$$E(Y/T) = \hat{\alpha} + \hat{\beta}T - U_t$$

Since time is a regressor.

The estimated linear trend model is

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta}T + U_t$$

Where:

$$\hat{\alpha} = \bar{Y}_t - \hat{\beta}\bar{T}$$

$$\hat{\beta} = \sum ty / \sum t^2$$

The linear Trend Model is very useful in determining or finding the role of growth of certain economic variables such as

- Population
- GNP/GNP per capita
- Employment
- Productivity
- Money Supply
- Trade Surplus/Deficit
- Government expenditure, etc

16.2 Constant growth models/ the log-lin model

Non-Constant Growth Rate:

The linear Trend Model involves non-constant growth rate

For example, at period 1, *Y will be* Y_1. To find change in Y, we:

$\Delta Y = Y_t - Y_{t-1}$ as it is in difference equation.

But the $E(Y/T) = \alpha + \beta T$ - this represents the slope of the trend line, which shows the change in *Y as T* changes.

Then $\Delta Y = \beta = Y_t - Y_{t-1}$ which is the growth (positive or negative) in Y. The rate of growth is given as:

$$\frac{Y_t - Y_{t-1}}{Y_t} = \frac{\Delta Y_t}{Y_t} = \frac{\Delta Y}{Y} = \frac{\beta}{Y}$$

Hence the rate of growth is β → *which implies that the growth in Y cannot be constant so:*

(a) When $\beta > 0$ i.e. positive → Y will be an increasing series, which will increases at a decreasing rate since the rate of growth will be declining as implied by the fraction β/Y.

(b) When $\beta < 0$ i.e. it is negative → Y will be a declining series and the rate of growth will be increasing very fast. Y will be declining at an increasing rate.

The linear trend model does not therefore exhibit constant growth rate.

- **Constant Growth Rate:**

Suppose we want to find out the growth rate of productivity in firm A for the period *1990 – 2010.*

Y_t → *the productivity in ('000 N)*

t → *the time from 1990 – 2010*

We can use the compound interest formula to determine this

Recall:

$$A = P(1+r)^n$$

Where A = TV (trend value) = Amount

n = years = t

r = rate of interest

P = initial deposit/principal

We can substitute these variables with our concerned variables

$A = Y_t$ = *productivity*

$P = Y_o$ = *Initial productivity*

$n = T$ = *the time*

$r = g$ = *the compound (overtime) rate of growth of Y*

Our model becomes: $Y_t = Y_o(1+g)^t$

We can see that the model is non-linear, as it is raised to the power of T.

To transform it to linear, we take the natural logarithm of both sides

$$lnY_t = lnY_o + ln(1+g)^t$$

Recall that in log, if log_eA^b can be written as:

$$log_eA^b = b.log_eA$$

Therefore, $lnY_t = lnY_o + T.ln(1+g)$

This is a linear model, which can also be reduced to

$$Y_t = \alpha + \beta T + U_t$$

By assuming or letting

$$lnY_o = \alpha$$

$$ln(1+g) = \beta$$

Therefore

$lnY_t = \alpha + \beta T$

Since the compound interest formula will not hold exactly, we add the error term.

Therefore,

$lnY_t = \alpha + \beta T + U_t$

Which is like any other linear regression model such that the parameters α and β are linear.

The difference between this model and other linear model is that the regressant is the logarithm of Y and the regressor is time which takes values, 1,2, 3, ... etc

This model is called semi-log model or log-lin model. The model gives constant proportional or relative change in Y for a given absolute change in the regresor T, which is measured by the slope β.

Therefore, β = relative change in regressand/absolute change in regressor

$$\beta = \frac{Y_t - Y_{t-1}/Y_{t-1}}{T_t - T_{t-1}}$$

$$\beta = \frac{Y_t - Y_{t-1}/\Delta Y_{t-1}}{T_t - T_{t-1}}$$

This is the relative (growth rate) change in Y for an absolute change in T.

Note that, β x 100 = growth rate in Y = Semi-elasticity of Y with respect to T.

- **Types of Growth Rate**
- *Instantaneous rate of growth → the rate of growth in Y at a period in time. It is given by the slope, β*
- *Compound rate of growth → the rate of growth of Y over a period of time.*

It is given: $e^{\beta} - 1$

i.e. antilog β-1 x 100

This is found from our compound rate of growth, g, in our original -representation.

Recall that, $\beta = \ln(1+g)$

So, $1+g = -\beta.1/\ln$

$$g = \frac{\beta.1}{\ln 1}$$

Therefore, $g = e^{\beta} - 1$, which is the compound rate of growth

Computational Procedure

Given the model

$$\ln Y_t = \alpha + \beta T$$

$$\hat{\alpha} = \ln\overline{Y_t} - \hat{\beta}\overline{T}$$

$$\ln Y_t = \sum \ln Y_t / n$$

$$\hat{\beta} = \frac{\sum y_t}{\sum t}$$

Example 16.1

From the quarterly expenditure on services of a particular country for 5 years, the following growth model is estimated

$$\hat{\ln Y_t} = 7.7890 + 0.0073t$$

$$(0.0023)\ (0.00017) \qquad r^2 = 0.99$$

Note that:

Y = expenditure on service

The period is between 1993:1 1.

1998:3

I.e. From the first quarter of 1993, to the third quarter of 1998

We can give the following interpretations

β = 0074.3

The instantaneous rate of growth is *(0.00743) x100 = 0.743*. This means that between the periods *1993:1 to 1993:3*, there has been a quarterly increase in expenditure on service by *0.743 percent*.

Yearly, there is roughly a *(0.743 x 4) = 2.97* percent increase in expenditure on services. This value is elastically significant.

The intercept *7.7840* gives the beginning value of service expenditure (at the end of the fourth quarter of 1992) when we take its antilog, since

Ln of Y = 9.9890 at the period.

Therefore, antilog of *9.9890 = 2413.90* at the beginning of *1993 is 2413.90* million or billion – depending on the unit.

The compound rate of growth is

$$e^{.00743} - 1$$

$$= (Antilog\ 0.00743 - 1) \times 100$$

$$= 0.00746 \times 100$$

$$= 0.746$$

Hence, the quarterly compound rate of growth in expenditure on services is *0.746* percent. This value is slightly greater than the instantaneous rate of growth because of the compounding effect.

To differentiate the log-lin model from the linear trend model, we gave the estimated linear trend model of the same data (without using log) as follows:

$$Y_t = 2405.848 + 19.6920t$$

$$r^2 = 0.9843$$

This is a linear model:

The interpretation is as follows:

the slope 19.6920 implies that over the period 1993:1 to 1993:3 on average, there is an absolute increase in the expenditure on service (absolute rate of growth) of 19.69 (approx. 20) million or billion per quarter. There was an upward trend, thus, in the expenditure on services.

Hence the main difference between the log-lin model and linear model is that: log-lin gives the relative rate of growth.

Example 16.2

The per capita incomes of Nigeria between 1996 to 2005 (hypothetical) are given below.

Yr	[illegible]	[illegible]	[illegible]	'99	'00	'01	'02	'03	'04	'05
PCI	[illegible]	[illegible]	[illegible]	560	591	632	685	750	794	866

Estimate the log-lin and linear trend models and interpret them.

Solution

	Y	Y^2	T	YT	lnY	lnY	lnYT
1	484	234356	1	484	6.182	38.218	6.182
2	504	254016	4	1008	6.223	38.720	12.445
3	520	270400	9	1560	6.254	39.100	18.761
4	560	313600	16	2240	6.328	40.043	25.312
5	591	349281	25	2955	6.382	40.728	31.909
6	632	399424	36	3792	6.449	41.588	38.693
7	685	469225	49	4795	6.529	42.633	45.706
8	750	562500	64	6000	6.620	43.825	52.961
[illegible]	[illegible]	630436	81	7146	6.6[illegible]	44.583	60.639
[illegible]	[illegible]	[illegible]	[illegible]	8664	6.764	45.750	67.639
[illegible]	6386	4233094	385	38640	64.408	415.198	359.702

$\sum T = 55$ $\quad \sum YT = 38640$ $\quad \sum lnYT = 359.702$

$\sum Y = 6386$ $\quad \sum lnY = 64.408$ $\quad \sum Y^2 = 4,233,094$

$\sum lnY^2 = 415.198$

Linear Trend Model

$\hat{Y} = \hat{\alpha} + \hat{\beta}T$

$\hat{\beta} = \frac{\sum YT - n\bar{Y}\bar{T}}{\sum T^2 - n\bar{T}^2} = \frac{\sum yt}{\sum t^2}$

$= \frac{38640 - (10)(6386/10)(55/10)}{385 - 10(55/10)^2}$

$\hat{\beta} = \frac{38640 - (10)(638.6)(5.5)}{385 - 10(5.5)^2}$

$= \frac{38640 - 33123}{385 - 302.5} = \frac{3517}{82.5}$

$\hat{\beta} = 42.63$

$\hat{\alpha} = \hat{Y} + \hat{\beta}\bar{T}$

$\hat{\alpha} = 638.6 - (42.63)(5.5)$

$= 638.6 - 234.467$

$= 404.133$

≈ 404

Therefore, the estimated regression line is

$\hat{Y} = 404.133 + 42.63t$

To determine the r^2

$r^2 = \frac{\hat{\beta}.\sum xt}{\sum y^2}$

$\sum xt = 3517$ *above*

$\sum y^2 = \sum y^2 - n\bar{Y}^2$

$= 4.233094 - (10)(638.6)^2$

$= 154994.4\ (1-0.967)$

$= 5063.62$

$\delta^2 = \sum e^2/n-K = 5063.62/10-2 = 5063.62/8$

$s(\hat{\beta}) = \sqrt{\sum T^2/\sum t^2}\ .\ \delta^2$

$s(\alpha) = \sqrt{\frac{385}{10(82.5)}} \times 632.9525$

$\delta^2 = \sum e^2/n-K = 5063.62/10-2 = 5063.62/8$

$s(\hat{\alpha}) = \sqrt{295.3778} = 17.187$

$s(\hat{\beta}) = \sqrt{1/\sum t^2}\ .\ \delta$

$= \sqrt{1/82.5 \times 632.9525} = \sqrt{7.672}$

$s(\hat{\beta}) = 2.77$

$t_{cal}\ (\hat{\beta})/s(\hat{\beta}) = 15.37$

$t_{cal}\ (\hat{\alpha})/s(\hat{\alpha}) = 404.133/17.187 = 23.514$

We can summarize our results as follows:

$\hat{Y}_t = 404.133 + 42.63T$

s.e. = (17.187) (2.77)

t = (23.514) (15.39) $r^2 = 0.969$

At 95% our theoretical = 2.306 (8d.f)

Hence, both $\hat{\alpha}$ and $\hat{\beta}$ are statistically significant.

- *The intercept 404.133 implies that the per capita income in 1995 was 409.133*

- *The slope, $\beta = 42.63$ implies that on average there is a yearly absolute growth in per capita income of 42.63*
- *$r^2 = 0.967$ implies that time explains about 97% variation in per capita income.*

Hence, there was an upward trend in the per capita income.

TheLog-lin model

$\ln\hat{Y}_t = \alpha + \beta T$

$$\hat{\beta} = \frac{\sum \ln YT - n\overline{T}\overline{\ln Y}}{\sum T^2 - n\overline{T}^2}$$

$$= \frac{359.702 - (10)(5.5)(64.408/10)}{385 - 10(5.5)}$$

$$= \frac{359.702 - 354.244}{82.5}$$

$$= \frac{5.458}{82.5}$$

$$= 0.0662$$

$\alpha = \ln\hat{Y}_t + \hat{\beta}T$

$= 6.4408 - (0.0662)(5.5)$

$= 6.4408 - 0.364$

$\alpha = 6.077$

$\sum \ln Y^2 = \ln Y_1^2 - n\overline{\ln Y}^2$

$= 415.198 - (10)(6.4408)^2$

$= 418.198 - 414.839$

$= 0.389$

Therefore $r^2 = \hat{\beta}\frac{\sum \ln Y_1}{\sum \ln Y^2}$

$$= 0.0662 \times \frac{5.458}{0.369}$$

$$= \frac{0.3613}{0.369}$$

$= 0.199$

The model is:

$\ln Y_t = 6.077 + 0.66T$

- *The antilog of the intercept gives the per capita income in 1993 i.e.*

  Antilog 6.077 = 435.12

- The slope B gives the instance rate of growth

  i.e. there is annual rate of growth of

  0.0662 x 100 = 6.62%.

  The compound rate of growth is

  Antilog 0.0662 – 1

  = 1.0684 – 1

  = 0.0684 x 100

  = 6.84% per annum

16.3 The lin-log model

In the log-lin model, we find the percentage growth in Y for an absolute change in X. In the lin-log model, we find the absolute change in Y for a

percentage change in X. in this case. all regressand Y is not in log. but our regresor X is in Log.

$$Y_i = \alpha + \beta lnX + U$$

The slope β gives the absolute change in Y for a percentage change in X. Therefore. β can be written

$$\beta = \frac{Change\ in\ Y}{Relative\ change\ in\ X}$$

$$\beta = \frac{\Delta Y}{\Delta X/X} = \frac{Y_2 - Y_{2-1}}{X_2 - X_1 / X_{2-1}}$$

In making interpretation β is divided by 100% just as it is multiplied by it in log-lin model.

Again. the absolute change in Y (ΔY) is $\beta(\Delta X/X)$.

Note that our regresor X is no longer time. The lin-log model is commonly used in estimating the relationship between a particular type of expenditure and the total (whole) expenditure. For example. it is used in the Engel expenditure model to estimate or show that the total expenditure that is devoted to food tends to increase in AP as total expenditure increases in G P.

16.4 *The log linear model: The log- log model*

A model in which both the regressor and regressand are in log is called double-log. log – log. ln-linear or constant elasticity model.

The original model is the exponential regression model of the form

$$Y_i = \alpha X_i^{\beta} e^{U_i}$$

This can be written in log-linear form by taking both sides with ln.

$$lnY_i = ln\alpha X_i^{\beta} e^{U_i}$$

$$lnY_i = ln\alpha + \beta lnX_i^{\beta} + e^{U_i}$$

$$lnY_i = ln\alpha + \beta lnX_i + e^{U_i}$$

$$lnY_i = ln\alpha + \beta lnX_i + U_i(1)$$

Therefore.

$$lnY_i = ln\alpha + \beta lnX_i + U$$

Taking lnα as α^* we have

$$lnY_i = \alpha^* + \beta lnX_i + U$$

To estimate the regression line. we can set

$$\alpha = \alpha$$

$$lnY_i = Y$$

$$lnX_i = X^*$$

So that we have $Y_i = \alpha + \beta X + U$

This can be estimated using OLS method. since the model is linear in the parameters.

α and β are unbiased.

The slope β. measures the elasticity of Y with respect to X. accordingly. the elasticity coefficient is constant. i.e. change in ln Y per unit. Change in lnX remains the same no matter at which ln X we measure it.

Note that the simplest way to determine whether bivariate model fits log-log model is to plot the scatter diagram of ln Y against ln X and see if the graph is approximately a straight line.

Example16.3 Given the following data. estimate the demand function

$$Y = \alpha X^{\beta} e^{U}$$

Where Y = *demand for rice*

X = *price of rice in arbitrary units*

Y	X	lnY	lnX
543	61	6.2971	4.1109
580	54	6.3631	3.0890
618	50	6.7265	3.9120
695	43	6..5939	3.7612
724	38	6.5848	3.6376
812	36	6.6995	3.5835
887	28	6.7879	3.3322
991	23	6.8987	3.1353
1186	19	7.0685	2.9445
1940	1-	7.5785	2.3026

Using OLS estimation, we have estimated the model as follows:

$$Y = \alpha X^{\beta} e^{u}$$

$$Log_e Y = Log_e \alpha + Log_e \beta + U$$

Or

$$lnY = ln\alpha + \beta lnX$$

Setting $ln\alpha = 0$

$$lnY = 0 + \beta lnX$$

$$lnY = \hat{0} + \hat{\beta} lnX$$

From our estimation we have got

$$Log_e Y = 9.121 - 0.69 logX \text{ or}$$

$$lnY = 9.121 - 0.69 lnx$$

$$(0.07)\ (0.02)\ r = 0.992$$

Writing the model in exponential regression form we have:

$$Y = \alpha X^{\beta} \text{ or}$$

$$Q = \alpha P$$

$$Q = \alpha P^{-0.69}$$

Where $\beta = -0.69$ which is the elasticity of demand. The demand for rice is inelastic.

The intercept $\alpha = Antilog 0$

$= Antilog 9.121$

$= 9145.34$

$= 9145$

Meaning that at zero prices the consumers are willing to demand for *9145* units of rice.

Note: the simple interpretation (non technical) is that 1 units change (increase) in price of rice will lead to *0.69* decreases in the quantity of rice demanded

CHAPTER SEVENTEEN

17.1MULTIPLE REGRESSIONS MODEL.

In the real world, the dependent variable can be determined by two or more variables for instance consumption is determined not only by income but by other variables such as taste, weather and so on therefore to determine the influence of these variables on consumption; a researcher may have to estimate a partial regression coefficient. Economic theory states that consumption is a function of income, taste, etc.

$Y = f(X_1, X_2, \ldots X_n)$

Where

Y = *consumption*

X_1 = *income*

X_2 = *Taste*

Mathematical economist will say that

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

This relationship means that, the *variations in consumption Y are fully explained by changes in income* X_1 *and taste* X_2.

An econometrician will go further than the relationship because there are other variables that may influence consumption that are not specified in this equation. To take care of such variables, therefore, we introduce the random error term.

$$Y = (\beta_0 + \beta_1 X_1 + \beta_2 X_2) + U$$

Systematic Component — *random components*

Where β_1 β_2 are partial regression coefficients while the intercept β_0 gives the mean or average effect on Y of the entire variable excluded from the

model. Although its mechanically interpreted as the average value of Y, *when* X_1 *and* X_2 *are set equal to zero.*

To make this model more meaningful it is important to operate within the frame work of the assumptions underlying the estimation the parameters involve. Most especially

- *The randomness of U*
- *The zero means value of U i.e. E (Ui/X1i X2i) = 0 for each I*

  No serial correlation i.e. cov (Ui, Uj) = 0 i ≠ j

  Homoscedasticity i.e. var (Ui) = 0

  Normality of U i.e. Ui VN (0, 0)
- *Zero covariance between Ui and each Xi variable. i.e cov (Ui X1i) = cov (Ui, X2i) =0*
- *No specification bias*
- *No multicollinearity*
- *Linearity in parameter assumption*

Having specify the model, the next thing is to use sample observation of Y, X_1 *and* X_2 *to estimate the true parameter* β_0, β_1, β_2

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

Where $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ are estimates of the true parameters β_0, β_1, β_2 of the consumption relationship.

These estimate or coefficient can be obtained by minimizing the sum of squared residuals, *i.e.* $\sum \hat{e}^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2$

We obtain by

(4) Finding the error, (in our model)

$U_i = Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2$

(5) *Squaring the error*

$$U_i^2 = (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2$$

(6) *Summing the squared error*

$$\Sigma U_i^2 = \Sigma (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2$$

To minimize we take partial derivation in respect of β_0, β_1 and β_2 respectively so as to get our normal equation and subsequently, the coefficient β_0, β_1 and β_2 i.e. take partial derivation and equate zero.

$$d\Sigma U_i^2 / d\beta_0 = (-1)(2)\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)(1) = 0$$

$$= -2\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2) = 0$$

Divide through by -2

$$= \frac{-2\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)}{-2} = \frac{0}{-2}$$

$$= \Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2) = 0$$

Where $\frac{0}{-2} = 0$

Open the bracket

$$= \Sigma Y - \Sigma\beta_0 - \Sigma\beta_1 X_1 - \Sigma\beta_2 X_2 = 0$$

$$= \Sigma Y - N\beta_0 - \beta_1 \Sigma X_1 - \beta_2 \Sigma X_2 = 0 \quad \ldots (1)$$

Where $\Sigma\beta_0 = \beta_0 + \beta_0 + \beta_0 + \ldots = N\beta_0$ i.e. Number of β_0

To find our β_0 we multiply through by N and make β_0 the subject of the formula.

$$\frac{\Sigma Y}{N} = \frac{N\beta_0}{N} + \frac{\beta_1 \Sigma X_1}{N} + \frac{\beta_2 \Sigma X_2}{N}$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2$$

Therefore

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

To get the 2nd normal equation we take partial derivation in respect of the β

$$d\Sigma U_i^2 / d\beta_1 = (-X_1)(2)\{\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)\} = 0$$

$$= (-2)\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2) = 0$$

To eliminate -2 we divide through by -2

$$\frac{(-2X_1)\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)}{-2} = \frac{0}{-2}$$

$$X_1\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2) = 0$$

$$\Sigma X_1 Y - \beta_0 \Sigma X_1 - \beta_1 \Sigma X_1^2 - \beta_2 \Sigma X_2 X_1 = 0$$

$$\Sigma X_1 Y - \beta_0 \Sigma X_1 - \beta_1 \Sigma X_1^2 - \beta_2 \Sigma X_2 X_1 \quad \ldots (2)$$

To determine the third normal equation differentiate ΣU_i^2 in respect of

$$d\Sigma U_i^2 / d\beta_2 = (-X_2)(2)\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 \Sigma X_2) = 0$$

$$= (-2X_2)\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 \Sigma X_2) = 0$$

Divide through by -2

$$= \frac{(-2X_2)\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 \Sigma X_2)}{-2} = \frac{0}{-2}$$

$$X_2\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 \Sigma X_2) = 0$$

$$\Sigma X_2 Y - \beta_0 \Sigma X_2 - \beta_1 \Sigma X_1 X_2 - \beta_2 \Sigma X_2^2 = 0$$

$$\Sigma X_2 Y = \beta_0 \Sigma X_2 + \beta_1 \Sigma X_1 X_2 + \beta_2 \Sigma X_2^2$$

$$\beta_0 = \Sigma Y = N\beta_0 + \beta_1\Sigma X_1 + \beta_2\Sigma X_2 \qquad ...(1)$$

$$\beta_1 = \Sigma X_1 Y = \beta_0\Sigma X_1 + \beta_1\Sigma X_1^2 + \beta_2\Sigma X_1X_2 \qquad ...(2)$$

$$\beta_2 = \Sigma X_2 Y = \beta_1\Sigma X_2 + \beta_1\Sigma X_1X_2 + \Sigma X_2^2 \qquad ...(3)$$

These equations are called normal equations and are used to obtain the coefficient β_0, β_1, and β_2 in other words they are used to fit the regression line. To do that we use crammer rule matrix as follows

Given the above OLS-estimators we need to also obtain the standard error of these estimates which are important to establish confidence interval and to test the statistical hypothesis. The relevant formulas are as follows

$$\begin{bmatrix} A \end{bmatrix} = \begin{pmatrix} \Sigma Y \\ \Sigma X_1 Y \\ \Sigma X_2 Y \end{pmatrix} = \begin{pmatrix} N & \Sigma X_1 & \Sigma X_1 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\begin{bmatrix} A \end{bmatrix} = \Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2$$

To determine β_1 we can replace the middle column with the sums on the right hand of the equations and divide by the determinants

$$\hat{\beta}_1 = \frac{\begin{pmatrix} N & \Sigma Y & \Sigma X_1 \\ \Sigma X_1 & \Sigma X_1 Y & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_2 Y & \Sigma X_2^2 \end{pmatrix}}{\Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2}$$

$$= \frac{\Sigma X_1^2 Y (\Sigma X_2^2) - \Sigma X_1^2 Y (\Sigma X_1 X_2)}{\Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2}$$

$$\hat{\beta}_1 = \frac{\begin{pmatrix} N & \Sigma X_1 & \Sigma Y \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 Y \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2 Y \end{pmatrix}}{\Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2}$$

$$= \frac{\Sigma X_2^2 Y (\Sigma X_1^2) - \Sigma X_2^2 Y (\Sigma X_1 X_2)}{\Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2}$$

Given the above OLS estimates, we need to also obtain the standard error of these estimates which are important to establish confidence interval and to test the statistical hypothesis. The relevant formulas are as follows

$$Var(\beta_0) = [1/n + X_1^2 \Sigma X_2^2 + X_2^2 \Sigma X_1^2 - 2\bar{X}_1\bar{X}_2 \Sigma X_1 X_2] \delta U^2$$

$$Se(\beta_0) = + \sqrt{var\beta_0}$$

$$Var(\hat{\beta}_1) = \frac{\Sigma x_2^2}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x1\, x2)} \delta U_i^2$$

Or equivalently

$$(\beta_1) = \frac{\delta U i^2}{\Sigma x_1^2 (1-r^2)}$$

$$Se(\hat{\beta}_1) = + \sqrt{var(\beta_1)}$$

$$Var(\hat{\beta}_1) = \frac{\Sigma X_1^2}{\Sigma X_1^2 \Sigma X_2^2 - (\Sigma X_1 X_2)^2} \delta U i^2$$

Or equivalently

$$Var(\hat{\beta}_2) = \frac{\delta U i^2}{\Sigma x 2^2 (1-r^2)}$$

$$Se(\beta_2) = + \sqrt{var(\beta_1)}$$

and it is given by

$$\sigma u^2 = \frac{\Sigma Ui^2}{n-k} = \frac{\Sigma Ui^2}{n-k}$$ *while the covariance* (β_1, β_2)

$$= \frac{-r\,23\delta^2}{\sqrt{(1-r^2)\Sigma X_1^2 X_2^2}}$$

Where, r23 = correlation between X_1 *and* X_2.

17.2 The multiple coefficient of multiple determination

As in the two variables case, the R^2 shows the percentage of total variable of Y that is explained by the regression plane. The notation r^2 is extended to the two variable models while in three variable model we would like to know the proportion of the variation in Y that is explained by variable X_1 and X_2 jointly and therefore we may extend the capital letter R^2 with sub script. of the variables whose relationship is being studied. In the three or more variable model, the square multiple correlation coefficient (which is another name for R^2) is denoted by $R^2_{Y\,X_1X_2}$. It measures the goodness of fit of the regression equation. To derive $R^2_{Y\,X_1X_2}$, we establish in the previous chapter that *TSS = ESS + RSS*

i.e. $$\frac{\Sigma\hat{Y}^2}{\Sigma y^2} = \frac{\Sigma y^2}{\Sigma y^2} + \frac{\Sigma ei^2}{\Sigma y^2}$$

$$1 = R^2 + \frac{\Sigma e_i^2}{\Sigma y^2}$$

Therefore $R^2 yx_1x_2 = \frac{\Sigma\hat{y}^2}{\Sigma y^2} = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{\Sigma ei^2}{\Sigma y^2}$

$$= \frac{\Sigma y^2 - \Sigma e_i^2}{\Sigma y^2}$$

We have also established that

$e_i = y - \hat{y}$ *and*

$$y = \beta_1 X_i + \beta_2 X_2$$

The square residuals are

$\Sigma e_i^2 = \Sigma e_i (y - \hat{y})$ *Since* $\Sigma e_i^2 = \Sigma e_i y$

$= \Sigma e_i (yi - \beta_1 X_1 + \beta_2 X_2)$

$= \Sigma e_i yi - \beta_1 \Sigma e_i x_1 + \beta_2 \Sigma e_i x_2$

$\Sigma e_i^2 = \Sigma e_i yi$

Since, $\Sigma eiyi = 0$, $\Sigma eix_2 = 0$ *and* $\Sigma ei = yi - \hat{y}i$

Therefore. $\Sigma ei^2 = \Sigma(y - \hat{y})y$

$= \Sigma y (y - \beta_1 X_1 + \beta_2 X_2)$

Where $\hat{y} = \beta_1 X_1 + \beta_2 X_2$

When we open the bracket, we have

$\Sigma ei^2 = \Sigma yi2 - \beta_1 \Sigma yx_1 + \beta_2 \Sigma yx_2$

Therefore. $Ryx_1x_2 = \frac{\Sigma yi^2 - (\Sigma yi^2 - \beta_1 \Sigma yx_1 + \beta_2 \Sigma yx_2)}{\Sigma y2}$

$= \frac{\beta_1 \Sigma yx_1 + \beta_2 \Sigma yx_2}{\Sigma y2}$

Note that R^2 like the r^2 lies between zero and 1. That is $0 < R^2 < 1$. If it is =1, the fitted regression line explain 100 percent of the variability in Y. if it is zero, the model did not explain any of the variation. However, the fit of the model is better when R^2 is closer to 1. .

If $R^2 = 0.9$ for instance, it is closer to 1, but this value must be tested to establish its statistical significance. For example we first of all set our own hypothesis

H_o: $R^2 = 0$

H_1: $R^2 \neq 0$

At $\alpha = \alpha_o$

The statistic for testing R^2 is the F statistics given by

$$F^* = \frac{R^2/k-1}{(1-R^2)/n-k}$$

This is another way of testing the overall significance of the model.
Where R^2 = the variability in Y that is explained by the model.

$1 - R^2$ = *the proportion of that fraction of the variation that is not explained by model.*

That is R^2 = *explained proportion of*

$1 - R^2$ = *unexplained proportion of*

Not also that

$R^2 \sim X^2$ *with* $K - 1$ *degree of freedom*

$1 - R^2 \sim X^2$ *with* $n - k$ *degree of freedom*

This means that R^2 tends towards or is equivalent to chi-square with $k - 1$ *degree of freedom*

Decision rule

If $F_c > F_t$ *reject* H_0

If $F_c < F_t$ *accept* H_0

Note that, if the purpose of estimating your model is for policy analysis or decision making, the parameters becomes the focus of the model. If the parameters are significant, no matter the value of R^2, the relevance of the model is justified.

However, if the purpose of the model is to forecast or prediction, what is important is the R^2, not the parameters. Therefore, the value of your estimated model depends on the problem it intend to solve.

Worked example 17.2

Given that model $Y = B_0 + B_1X_1 + B_2X_2 + U$, supposing $R^2 = 0.95$, $n = 10$ $\alpha = 0.05$, test the significance of R^2.

Solution

$H_0 : R^2 = 0$

$H_1 : R^2 = 0$

$$Fc = \frac{R^2/(K-1)}{(1-12^2)/n-k} \qquad \text{where } v_1 = K - 1 = 3 - 1 = 2, \quad v_2 = n - k = 10 - 3 = 7$$

$$= \frac{0.95/3-1}{(1-0.95)/10-3}$$

$$= \frac{0.95/2}{0.05/7}$$

$$= \frac{0.48}{0.007}$$

$$= 68.57$$

Decision rule

If $Fc > F_t$ *rejects* H_0 *otherwise accept* H_1

since Fc *38.57* $> F_t$ *3.46 we reject* H_0 *and conclude that* R^2 *is significantly different from zero*

The adjusted $\overline{R}^2$

R^2 vary directly with increase in the number of explanatory variables. That is to say that R^2 increase the number of explanatory variable increases. R^2 statistic never decreases if a researcher increases the quantity of his model by simply adding additional regression. The value of the numerator of the R^2 statistics is increased by that amount while the denominator remains the same.

R^2 can be adjusted by the value of degree of freedom to connect the likely defect. The adjusted coefficient of determination is denoted by R^2 with a bar, i.e

$$\overline{R^2} = 1 - \frac{\Sigma ei/n-k}{\Sigma x^2/(n-1)}$$

$$= 1 - (1 - R^2)\frac{n-1}{n-k}$$

From the above it is apparent that $\overline{R^2} = R^2$ if $k = 1$ which implies that, the X variables increase, the *adjusted* R^2 increase less than R^2

Therefore, if $K = 1$, then $(\overline{R^2} = R^2)$ the coefficient of determination will be equal to the adjusted coefficient of determination.

For example since

$$\overline{R^2} = 1 - (1 - R^2)\frac{(n-k)}{(n-1)}$$

if $k = 1$ then

$$\overline{R^2} = 1 - (1 - R^2)$$

$$= 1 - (1 - R^2) \text{ by opening the brackets}$$

Therefore, $\overline{R^2} = R^2$

Comparing two R^2

For the researcher to compare two models on the basis of their coefficient of determination, the sample size and the dependent variable must be the same, no matter the form of the explanatory variable and whether it is adjusted or not.

For example, if your have estimated two models one about consumption and the other about production as follows

$Yi = 0.25 + 0.67X_1 + 0.26X_2$ with $R^2 = 0.86$

$Y2 = 0.12 + 0.57X_1 + 0.34X_2$ with $R^2 = 0.95$

As long as the dependent are not the same as above, you cannot compare R^2 but, when you have a dependent variable that are the same, then R^2 can be compared. That is comparism becomes relevant when the dependent variables are the same.

e.g. $Y_1 = 0.25 + 0.67X_1 + 0.26X_2$ $R^2 = 0.86$

$Y_2 = 0.12 + 0.57X_1 + 0.34X_2$ $R^2 = 0.95$

Multiple regression (worked example)

17.3Hypothesis testing in multiple regression

To adequately analyze a multiple regression estimates, a researcher is expected to test the following

(1) Test hypothesis about the individual regression estimates

(2) Test the overall significance of the estimated regression model, ie find out if all the slopes are simultaneously equal to or greater than, or at least one of them is less than or equal to zero.

(3) Test that two or more coefficients are equal to one another

(4) Test that the partial regression coefficients satisfy certain restrictions

(5) Test for the stability of the regression model overtime or in different cross sectional unit

(6) Test the functional form of regression the model

7.4 Hypothesis Testing About individual Regression Coefficients

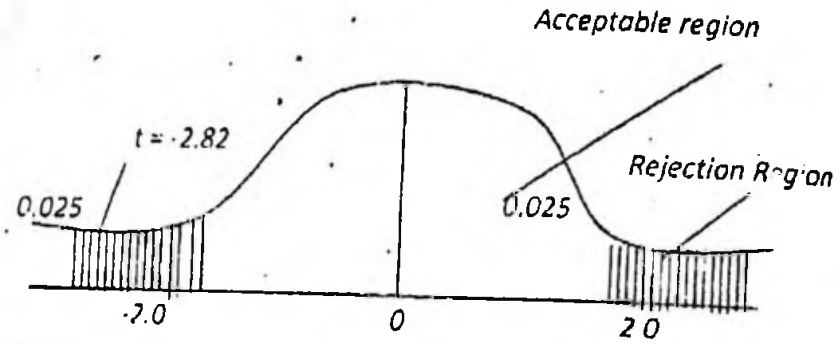
Given the normality assumption, we can use the t – statistic to test a hypothesis about any individual partial regression coefficient. To do so, a researcher must set the hypothesis

$H_0: \beta_i = 0$

$H_1: \beta_i = 0$

Choose the level of significance. That is either 0.05 or 0.025

Choose critical region depending on whether the test is a two tailed or one tailed test. For instance if our H_1 says $\beta_1 \neq 0$, then it is a two tailed if it is either $>$ or $<$, it is a one tailed test.



After that the researcher determine the value of the t- statistic which is given by

$$t^* = \frac{\beta_1 - \beta_1}{\delta\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\delta}\beta_1}$$

The researcher will then compare the value of t^* and the table value of t, if $t^* > tT$, we reject H_0 otherwise we accept H_0.

17.5 *Test for Parameter Restriction*

Another type of test is the test for parameter restriction which is different from the test of significance of the partial regression coefficients. In this type of test the hypothesis are set as follow

$H_0: \beta_1 + \beta_2 = 0$

$H_1: \beta_1 + \beta_2 = 1$

There are occasions where economic theory may suggest that the coefficients in a regression model satisfy some linear equality restriction. For instance, consider the following equation

$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} e^{U}$

Where Y = *output*, X_1 = *labour inputs*, X_2 = *capital inputs*, if the above equation is transform using log, we have

$$\ln Y = \alpha_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + e^{U}$$

Where $\alpha = \ln \beta_0$

Hence if there are constant return to scale (equiproportional change in output for in equiproportionate change in the inputs) economic theory of hypothesis would suggest that

$$\beta_1 + \beta_2 = 1$$

This $\beta_1 + \beta_2$ is an example of a linear equality restriction.

This hypothesis can be tested by t statistic

$$tc = \frac{(\hat{\beta}_1 + \hat{\beta}_2) - (\beta_1 + \beta_2)}{Se(\beta_1 + \beta_2)}$$

$$Then\ t = \frac{(\hat{\beta}_1 + \hat{\beta}_2) - 1}{\sqrt{Var(\hat{\beta}_1) + var(\hat{\beta}_2) + 2\,cov(\hat{\beta}_1 + \hat{\beta}_2)}}$$

Since $(\beta_1 + \beta_2) = 1$ *according to* H_0

It can also be tested using the F statistics

$$Where\ F = \frac{R^2_{UR} - R^2_R / M}{1 - R^2_{UR} / (n - k)}$$

where $R^2_{UR} = R^2$ *unrestricted*

R^2 *restricted*

M = *number of linear restriction*

K = *number of parameters in the unrestricted regression*

n = *number of observation*

17.6 Test of overall significance of the regression

This is the test of joint significance of B_1 and B_2 and the hypothesis is set as follows

$$H_0 : \beta_1 = \beta_2 = 0 + \beta_2$$

$$H_1 : \beta_1 \neq \beta_2 \neq 0$$

$$F = \frac{(\beta_1 \Sigma xy + \beta_2 \Sigma y_1 / K - 1}{\Sigma e_i^2 / n - k} = \frac{ESS/df}{RSS/df}$$

That is the analysis of variance approach or

$$F = \frac{R^2 / k - 1}{(1 - R^2)/(n - k)}$$

If $F > F_t$ reject H_0 otherwise, accept H_0.

6.7 Test of equality of two regression coefficients.

In this type of test the researcher will want to test the following hypothesis

H_0 $\beta_2 = \beta_3$ or $(\beta_2 - \beta_3) = 0$

H_1 $\beta_2 = \beta_3$ or $(\beta_2 - \beta_3) \neq 0$

Given that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$

The above hypothesis, state that, the two slope coefficient B_2 and B_3 are equal. Such hypothesis is important because it expresses

(1) The similarity of the two slopes

(2) It also express how closely related the two slopes are. e.g X_1 – income of consumer and X_2 – wealth of the consumer. The hypothesis try to shows that the income and wealth coefficients are the same. If the Y and the Xs are expressed in logarithm, the hypotheses mean that the elasticity of wealth and income are the same.

$$t = \frac{(\beta_2 - \beta_3) - (\beta_2 - \beta_3)}{Se\,(\beta_2 - \beta_3)}$$

Where $se(B_2 - B_3) = \sqrt{var(\beta_2) + var(\beta_3) + 2\,cov(\beta_2 + \beta_3)}$

If H_o is true

$$t = \frac{\beta_2 - \beta_3}{\sqrt{var(\beta_2) \quad a\beta_3) + 2\,cov(\beta_2 + \beta_3)}}$$

The test procedure involves the following

(1) Estimate β_2 and β_3

(2) Obtain the variance, covariance and the standard error of the estimates.

(3) Obtain the t ratio, Note that null hypothesis is $(\beta_2 - \beta_3) = 0$

(4) If the t variable computed is greater than the critical t value, at a given level of significant, then reject H_0

17.7 Chow test

This is a test for structural or parameter stability of regression models, since the value of parameters of a model do not remain the same through the entire time period, when one is dealing with time series data. This may be due to structural changes emanating from external forces. Or due to policy changes thus *if* $Y = a + B_1X_1 + B_2X_2 + B_3X_3 + U$ and the researcher collect data from say *1970 – 1990*. The explicit assumption is that $B_1\,B_2\,B_3$ has remained all the same over the years. If they change then there has been a structural break.

Now how does the researcher find out that a structural change has occurred? The researcher can find out by obtaining the *OLS* regression of Y and the Xs and save the residuals. By so doing he is maintaining or assuming that, the relationship by Y and the Xs has not change much over the span of (21) twenty one years that is from *1970 to 1990*.

After that the researcher can divide the data into two, may be from *1970 to 1979* that is (10) ten years and from *1980 to 1990* on another side. Then run an OLS regression of the two data separately and save their residuals.

Time period 1970 – 1979: $Y = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \alpha_3X_3 + U : n = 10$ *(1)*

Time period 1980 1990: $Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + U : n = 11$ (2)

Time period 1970 – 1990: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + U: n = n_1 + n_2 = 21$ (3)

Regression equation 3 assumes that there is no difference between the two time period and therefore estimates the relationship between Y and the Xs for the entire time period consisting of 21 observations. In other words this regression assumes that the intercepts as well as the slope coefficients remain the same over the period That is there is no structural change. If that happened then $a_1 = n_1$ $a_2 = X_2$ *and* $a_3 = X_3$ *(i.e. restriction).*

However, equation (1) and (2) assumes otherwise that the intercept and the slope coefficient are not the same, over that period under study.

CHAPTER EIGHTEEN

18.1 STATISTICAL BIAS

Statistical bias is the difference between the expected value of the estimate and the true parameter. When a multiple regression model is estimated. it is assumed that all relevant variables have been included. Thus the question: what happens when a relevant variable is omitted. This will amount to misspecification of the true model. and therefore. there will be bias. i.e. one of the parameters in the model may not live up to its expected value.

18.2 Omission of relevant variable or model misspecification

Omission of important variable in a model and the misspecification of a model are the sources of statistical bias. If a researcher omit an important variable in a model. a vacuum is created as a result of which the parameters may misbehave by coming up with an additional value which is unlike the parameter. I.e. not expected of the parameter. and therefore. it will give a misleading result. For instance.

Consider $Y = \alpha + \beta_1X_1 + \beta_2X_2 + u$ *- (1) and we estimated*

$$Y = \alpha + \beta_1X_1 \quad \text{- (2)}$$

What is the implication. looking at the difference between the correct model (1) and the estimated model (2).

The question to ask is will the $E(\beta) = \beta$ or will β remain an unbiased estimate? The problem is alternatively called model miss-specification.

To check this. we will consider the equations without the intercept

i.e. $Y = \beta_1X_1 + \beta_2X_2 + u$ - (1)

$Y = \beta_1X_1$ - (2)

Unbiased as in the $E(\beta_1) = \beta_1$

From the estimated equation (2) $\beta_1 = \frac{\Sigma x_1Y_1}{\Sigma x_1^2}$

We need to find $E(\beta)$, *if* $(E\beta) = \beta$ *the relationship holds.*

$E\beta_1 = \beta$ = *unbiased*

Steps to Establish the Above

$\beta_1 = \frac{\Sigma x_1 Y}{\Sigma x_1^2}$ = *Co-efficient of* x_1 *in the estimated equation*

But recall that the true value of $Y_i = \beta_1 X_1 + \beta_2 X_2 + U$, thus substituting we have

$$\beta_1 = \frac{\Sigma x_1 (\beta X_1 + \beta_1 x_2 + U)}{\Sigma x_1^2}$$

Taking the expected value thus, we arrive at

$$E\beta_1 = \frac{E(\Sigma x_1 (\beta_1 x_1 + \beta_2 x_2) + E(u_i)}{\Sigma x_1^2}$$

NB. Taking the (E) of the formula applies only to $v = 0$

Solve the bracket

$$= \frac{\Sigma x_1^2 \beta_1 + \Sigma \beta_2 x_1 x_2}{\Sigma x_1^2} \quad or \quad = \frac{\beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2}{\Sigma x_1^2}$$

$$= \frac{\beta_1 \Sigma x_1^2}{\Sigma x_1^2} + \frac{\beta^2 \Sigma x_1 x_2}{\Sigma x_1^2}$$

$$= \frac{\beta 1 \Sigma x_1^2}{\Sigma x_1^2} + \frac{\beta_2 \Sigma x_1 x_2}{\Sigma x_1^2}$$

$E\beta_1 = \beta_1 + \beta_2 \frac{(\Sigma x_1 x_2)}{\Sigma x_1^2}$ = *The scope of coefficient of the regression of* x_2 *on* x_1

Recalling that estimated equation $Y = \beta x_1 + u$

Clearly the steps are thus:

Given the model $Y = \beta_1 x_1 + \beta_2 x_2 + u$

And the miss-specified model as $Y = \beta_1 x_1 + u$

And $\beta_1 = \frac{\Sigma x_1 Y}{\Sigma x_1^2}$

Thus the question is β (as estimated unbiased? to answerthe question we substitute for the true value of Y and take the expected value. Simplified and taken the second part we arrive at

$$E(\beta_1) = \beta_1 + \beta_2 \frac{(\Sigma x_1 x_2)}{\Sigma x_1^2}$$

Note that. Unbiased means $E(\beta) = \beta$. hence, this shows clearly that β_1 is biased as the 2nd part is added and it is not equal to zero, because x_2 is an important variable. Thus if x_1 and x_2 are close substitutes and we know now that it is biased thus we may want to find out if it was over/under evaluated or biased.

Note that, β_2 is the co-efficient of the omitted variable and what is in the bracket is the slope of the co-efficient in the regression of x_2 and x_1. Thus our bias is the co-efficient of the omitted variables and the (explanation of bracket item).

The extent of biasness depends on $\beta_2 \frac{(\Sigma x_1 x_2)}{\Sigma x_1^2}$. *Hence, if* β_2 *is positive and* $(\Sigma x_1 x_2)$ *is also positive, then* β_1 *is over estimated and therefore, positively bias*

Consider an earnings function

$W_i = \alpha_0 + \alpha_1 S_i + \alpha_2 A_i + e$. Correct equation

W_i = *wage,* S_i = *years of schooling,* A_i = *ability (and our estimated equation is*

$w_i = \alpha_1 + \alpha_2 S_i + e$

Therefore, α_2 will be biased thus

$$E(\alpha_2) = \alpha_2 + \alpha_3 \frac{\Sigma S_i A_i}{\Sigma S_i^2}$$

To sign the bias (+) (+)

Thus we will end up over estimating (+) α_2. Thus the omission of relevant variable has implications.

18.3 Consequence of omission of important variable (under fitting or miss-specification)

(1) If the omitted variable (x_2 for instance) is related to the included variable x_1, the correlation coefficient between the two variables x_1 and x_2 will be non-zero. Hence, β_1 and β_2 will be bias and inconsistent. And the bias does not disappear as the sample size gets larger.

(2) If x_1 and x_2 are uncorrelated, only $\beta_1$will remain bias, while β_2 is unbiased.

(3) The variance of the disturbance term will be (too large) incorrectly estimated.

(4) The conventional measure of variance of β_1 estimate will be a bias estimator of the variance of the true β_1

(5) Consequently, the usual confident interval and the test of hypothesis procedure are likely to give misleading conclusions about the statistical significance of the estimated parameter.

(6) The forecast and its confident interval will be unreliable.

CHAPTER NINETEEN

19.1 THE X^2 CHI-SQUARE

The chi-square X^2 is a probability distribution that tries to answer the question "How well does a model depict reality" In other words it is a test of goodness of fit of a model to a particular problem or situation. It is also used to test the independence or otherwise of an event.

It is a technique for comparing these theoretical or expected frequencies with actual or observed frequencies.

Properties of Chi- square x^2

1. *X^2 is a random variable that cannot assume negative value.*

   Example we could let $u = x^2$ and call it u distribution. However, squaring 'chi" remained as that x^2 can only be zero or greater.

2. *The distribution of chi- square (x^2) depends only on the degree of freedom v.*

3. *X^2 distribution is always skewed to the right. The larger the Y the less the skeweness.*

4. *The mean of x^2 is given by the degree of freedom $E(x^2) = v$.*

5. *The variance is twice the degree of freedom: var (x^2) = 2 v.*

   The chi-square statistic is given by

$$X^2 = \frac{\Sigma(f_o - f_t)^2}{f_t} \quad \text{or} \quad X^2 = \frac{\Sigma(f_{oij} - f_{tij})^2}{f_{tij}}$$

19.2 X^2 Test of goodness of fit

Steps

1. *State the null and alternative hypothesis, e.g.*

   H_o : the sample size have not in any way described the populations

   H_i : the sample size have described the populations

2. *Determine the size of random sample.*

3. *Specify the number of frequency distribution*

4. *Calculate the theoretical frequencies.*

5. *Calculate the degree of freedom.*

6. *Choose the level of significance (i.e. the probability of type I error)*

7. *State the decision rule*

8. *Calculate the X*

9. *Compare the calculated value with the table value*

Worked example 19.2

When a coin activated dispenser releases a volume of liquid which is called a fill. Due to particular constructions of the dispenser, we know that the fill is normally distributed with standard deviation of 0.10 ounce regardless of the fill setting or average volume of liquid dispensed per fill. A random sample of 49 fill produces the following data:

7.85	7.98	8.06	8.11
7.86	8.99	8.06	8.11
7.87	8.99	8.07	8.12
7.87	8.01	8.07	8.16
7.88	8.03	8.07	8.16
7.89	8.03	8.08	8.17
7.92	8.04	8.09	8.19
7.94	8.05	8.09	8.21
7.95	8.05	8.09	8.21
7.96	8.05	8.10	8.22
7.97	8.05	8.10	8.24
7.97	8.06	8.10	8.26

Testing ..... that the population is normally distributed with standard deviation of 0.10.

Step I

Ho : Population is not normally distributed

Hi : Thepopulation is normal distributed.

Step II $n = 49$

Step III

Observation X	*Frequency*
7.80 to less than 7.90 ounce	*6*
7.90 - 8.00	*9*
8.00 - 8.10	*19*
8.10 - 8.20	*10*
8.20 - 8.30	*5*
Observed frequency	*49*

Step IV

Less than 7.90 ounce

7.9 To less than 800

8.00 To less than 8.10

8.10 To less than 8.20

8.20 ounce or more

Lower class limit XI	*upper class limit* XI	*standard normal deviations* $=Z_1 = \frac{X_1 - 8.05}{0.10}$	$=\frac{X_1 - 8.0}{0.10}$
-	7.90	-	1.50 =
7.90	- 8.00	-1.50	-0.50 =
8.00	- 8.10	-0.50	+0.50 =
8.10	- 8.20	+0.50	+1.50 =
8.20	- x	+1.50	+ x =

Area of class	theoretical frequency (fi)
$P(Z_1 < Z < Z_0)$	$49\ p(Z_1 < Z < Z_0)$
0.6669	3.27312
0.2417	11.8433
0.38.30	18.7670
0.2417	11.8433
0.668	3.2732
10000	49.0000

Step 5

Calculate df v.

$v = k - l - m$

K = no. of class interval

M = no of parf estimate

Therefore. $5 - 1 - 1 = 3$

Step 6 $\alpha = 0.05$

Since at $v = 3$ and $\alpha = 0.05$ the critical value = 7.81

Step 7 Decision rule

If $X^2 > 7.81$ reject Ho

$X^2 < 7.81$ accept Ho

Step 8 Calculate $X^2 = \frac{\Sigma(fo - fi)^2}{Fi}$

Event	Observed frequency fo	theoretical frequency fi	(fo - fi)	(fo-fi)²	(fo-fi)²/fi
1	6	3.2732	2.7268	7.4354	2.27
2	9	11.8433	-2.8433	8.0844	0.68
3	19	18.7670	0.2330	0.0543	0.00
4	10	11.8433	-1.8433	3.3978	0.29
5	5	3.2732	1.7268	2.9818	0.91
		49.0000	0.0000		4.15

Since $4.15 < 7.81$ we accept null hypothesis and conclude that the sample data supports the assumption that the population is normally distributed.

If $X^2 = 0$, the observed and the theoretical frequency agree exactly

If $X^2 > 0$ they do not agree exactly

The larger the value of X^2, the greater is the discrepancy between the observed frequency and the theoretical frequency.

19.3 X^2 Test of independence of events

"Association" and "Independence" are two important concepts in Economics. If two events are independent, then the occurrence of one event does not affect the probability of the other event occurring.

X^2 test is also a technique that is used to test the independence or otherwise of event occuring.

Example19.2

Assuming that, a statistic professor argues that the afternoon class does not perform as much as morning classes. While the department chairman argues that between 8am and 5pm it does not make any difference the class meets. The instructor locates records for past semester. The instructor claimed that, the same lecture was delivered to each of the classes, the same homework and examinations were given and all students had met the course prerequisites. The grades are presented thus:

	Letter grade					
Time the class met	A	B	C	D	F	Total
Morning	12	18	13	11	6	60
Afternoon	3	7	17	4	9	40
Total	15	25	30	15	15	100

Let r = raw and k = column, then we have r times k (r x k) contingencies table.

Since r = 2 and k = 5 therefore,

r x k = 2 x 5 = 10 = frequencies of occurrence of events.

H_o *: Event A and B are independent*

H_1 *: Event A and B are not independent*

Or

H_o *: The time class meet and letter grade are independent.*

H_1 *: The time class meet and letter grade are dependant.*

In the case of the test of independence, the null hypothesis does not suggest the theoretical distribution it only suggest the algebraic relationship with *r x k*. Therefore to calculate the theoretical frequency.

$$\frac{\text{(Total for row i) (total for column j)}}{\text{Grand total}} = \text{frequency for cell ij.}$$

Example.

Time	Letter grades A	B	C	D	F	Total
AM	60.15 =9 / 100	60.25 =15 / 100	60.30 =18 / 100	60.15 =9 / 100	60.15 = 9 / 100	60
PM	40.15 =6 / 100	40.25 =10 / 100	40.30 =12 / 100	40.15=6 / 100	40.15 =6 / 100	40
Total	15	25	30	15	15	100

$$X^2 = \frac{\Sigma(f_o - f_t)^2}{f_t} \quad \text{or} \quad = \frac{\Sigma\Sigma(fo_{ij} - ft_{ij})^2}{ft_{ij}}$$

Time	letter groups	fo	ft	(fo-ft)	(fo-ft)²	(fo-ft)²/ft
PM	A	12	9	3	9	1.00
AM	B	18	15	3	9	0.60
AM	C	13	18	-5	25	1.39
AM	D	11	9	2	4	0.44
AM	E	6	9	-3	9	1.00
PM	A	3	6	-3	9	1.50
PM	B	7	10	-3	9	0.90
PM	C	17	12	5	25	2.08
PM	D	4	6	-2	4	0.67
PM	E	9	6	3	9	1.50
Total		100	100	0		11.08

Degree of freedom v = (r - 1)(k-1)

(2-1)(5-1) = 4 degrees of freedom

With v = 4 and α = 0.05, X^2 given by the table = 9.49

Hence *if* X^2 => 9.49 reject Ho (and conclude that time and grade are dependent i.e. the instructors hypothesis.)

If $X^2 < 9.49$ accept Ho (and conclude that time and grade are independent.

(*The chairman's hypothesis*).

Since $X^2 = 11.08 > 9.49$; therefore, we rejected reject the null hypothesis.

CHAPTER TWENTY

20.1 NON-PARAMETRIC TESTS

Generally, statistical methods require specific assumptions about the nature, shape as well as other characteristics of the population (or populations) from which data used were obtained. It is assumed that these populations follow or have a normal distribution with mean $\mu = 0$ and constant variant δ^2 such that one can estimate the parameters involved (i.e. mean, variance etc) and subject them to the standard normal test statistical test.

The standard normal test statistic and other statistics are used to test these parameters because the sampling distribution used on these parameters are based on the specific assumption about the population from which the data used were obtained. Hence they are called parametric tests.

However, there are situations where the conditions of normality cannot be reached, since most practical data are non-normal and sometimes, it is not possible to estimate the parameters of the data or populations. Hence, the tests used in such are called non-parametric tests. The non-parametric method is applied when we are not concerned about the parameter of a given population of any kind.

There is another form of non-parametric method or test which is called the distribution free. In this case, although the researcher is concerned with the parameters of the population but such population does not follow any (assumption) distribution. Therefore, the distribution free test is applied when we do not make any assumption about the population.

Advantages of Non-Parametric test

1) Non parametric test can be applied to wide range of situation because it does not have stringent requirements. That is to say, that it does not require the assumption of normality or any assumption about the population.

2) The method can be applied to qualitative or non-numerical data such as gender, political party, or religion.

3) The test requires simpler computation and hence does not require higher level computational ability or competence.

4) It saves time and provides a kind of convenience to the researcher as the application scope is not elaborate.

5) The test, in most cases requires nominal measurement.

Disadvantages of Non-Parametric test

1) The test tends to ignore some sample-information and sometime wastes information because each numerical data are often reduced to a qualitative form. For instance, in sign test certain information is recorded simply as negative sign. The actual magnitude of such information is not considered.

2) The test is not as efficient as the parametric test and therefore it will require stronger evidence such as large sample before a null hypothesis can be rejected.

20.2 Sign Test

Sign test is the test which is based on the mathematical sign of minus and plus that are used to represent certain level of differences between two observation under study. It can be applied to wide range of issues or situation.

One sample size test

In this case we are dealing with the researcher's claim that the probability of having a sample value less than the mean value as well as the claim that the probability of having a sample value more than the mean value is equal to 0.5. Therefore, the hypothesis may be set based on the population mean when the population is small. However, when the population is large, the hypothesis may be based on the median. Therefore, the test involves the subtraction of the mean value from each element of the sample and obtaining only the signs involved either negative or positive as the case may be.

Note that. in sign test we expect the number of negative sign should be equal to the number of positive sign.

Hence $Y = -$, *if* $X < \mu$
$Y = +$, *if* $X > \mu$
$Y = 0$, *if* $X = \mu$

In the case of zero sign. it will make no meaning and therefore be excluded from the sample:

The one sample test can be one tailed or two tailed.

If one tailed. the hypothesis will be:

Ho: $P = 0.5$
Hi: $P > 0.5$ *or*

Ho: $P = 0.5$
Hi: $P < 0.5$
But in the case of two tailed test
Ho: $P = 0.5$
Hi: $P \neq 0.5$
In this type of test small or large samples can be used: however there is difference in the way large samples can be handled.

Small sample sign test
In the case where a small sample of size n is taken then (np and n(1-p)) is less than 5 is taken and therefore the null hypothesis is tested using the Binomial distribution with $p = 0.5$ or $\frac{1}{2}$
Hence we accept the null hypothesis Ho if the computed binomial value p(x ≥ number of plus or minus as the case may be) is greater than α otherwise we reject Ho.

Example 20.1 *Consider the following score of students in statistics exams 26, 40, 27, 32, 41, 21, 29, 28, 18. Check whether the mean score is 25 (Ho = 25) against (Hi>25) at 0.05 level of significance.*

Solution
First step $n = 9$

$p = 0.5$
$Np = n(1-p) = 4.5$
$a = 0.05$
Second step

Subtract 25 from each element of the sample and retain the sign

26	40	27	32	41	21	28	18
+	+	+	+	+	-	+	-

$(+) = 7$
$(-) = 2$

Let x be a random variable representing the plus (+) sign then we state the hypothesis as follows:
Ho: $\mu = 15$ *or* $p = 0.5$
Hi: $\mu > 15$ *or* $p > 0.5$

Therefore the binomial probability that the number of plus sign is more than or equal to seven is
$P(x \geq 7, n = 9, p = 0.5) = 1 - p(x \leq 6, n = 9, p = 0.5)$
$= 1 - 0.9102$
$= 0.0898$
Since the binomial value is greater than the significance. we accept Ho

Two sample sign test

In the case of two dependent samples: the test can be conducted by subtracting each element or value of the second sample from the corresponding value in the first sample thereby recording only the sign of the difference between them.

Example 20.2 In Nasarawa State University Keffi. science courses are often considered before other courses in terms of admission. consider the following number of admitted students in 14 science courses and 14 art and social science courses (i.e. to the right and to the left according to priority and determine or test whether there is a significant difference between the number of admitted students at 0.05 level of significance.

Sc.	92	7	66	65	16	29	71	55	12	02	88	58	21	33
Arts/SS	124	71	91	107	96	65	77	65	40	88	55	119	77	74
	-	-	-	-	-	-	-	-	-	-	+	-	-	-

Solution

$N = 14$

$(-) = 13$

$(+) = 1$

$\alpha = 0.05$

$p = 0.5$

$np = n(1-p) = 7$

Ho: $\mu = 0$ there is no difference (or the median of the difference is equal to zero).

Hi: $\mu \neq 0$ there is a difference (or the median of the difference is not equal to zero).

Let x be a random variable representing the less frequent sign i.e. x = 1 (since the plus is less frequent and equal to one.

Now x = 1

n = 14 and the two tailed test – $\alpha = 0.05$

Hence, the critical value on the sign test table is 2

Decision Rule

Reject Ho if the number of less frequent sign is less than or equal to the value on the sign test table, otherwise accept Ho.

Since x = 1 and is less than 2, we reject Ho and conclude that there is a difference between the number of students admitted in science courses and those of the arts and social science.

Large sample sign test

In the case of large sample (i.e. n > 30) the binomial distribution can be approximated into normal distribution. This means that given a random sample of size $n > 30$ where np and $n(1-p)$ *is at least 5*, then the hypothesis can be tested using the normal distribution with the probability $p = 0.5$. Therefore, if $x \sim B(X, n,p)$ then the larger the value of n (n>30), the researcher can approximate the binomial distribution to normal distribution as follows:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

where x = random variable representing the number of plus sign

np = mean of the distribution

p = probability

or

$$Z = \frac{(x + 0.5) - (n/2)}{\sqrt{n/2}}$$

where x = the number of times less frequent sign occur

n = sample size

Procedure for sign test

Small Sample	*Large Sample*
1) *Assign positive and negative signs and discard any zeros*	1) *Same*
2) *Let n equal the total number of signs*	2) *Same*
3) *Let n equal the number of frequent or less frequent sign as the case may be*	3) *Same*
4) *Set Ho and Hi and find out whether the sample data support Ho.*	4) *Same*
5) *If sample size $n \leq 30$, use binomial distribution*	5) *If sample size $n > 30$, then approximate binomial to normal distribution. In other words, convert test statistic to* $Z = \frac{X - np}{\sqrt{np(1-p)}}$ *or* $Z = \frac{(x + 0.5 - (n/2)}{\sqrt{n/2}}$
6) *Find the critical value from sign test table.*	(6) *Find the critical z value from the normal distribution table*
7) *If the test statistic is less than or equal to the critical value*	(7) *Same* *Reject Ho*

20.3 *Kruska Wallis test*

Kruska Wallis test is a non-parametric alternative to the one way analysis of variance (the F test). The H test assesses the significance of difference among three or more independent samples. but the H test requires data.

The procedure for H-test is to rank the samples from the highest to the lowest or vice versa as though they constitute a single sample. Tied ranks are assigned the average of their positions.

The next step is to obtain sums of ranks of each sample separately. Their sums of ranks are used in calculating the H-value. The value is given by

$$H = \frac{12}{N(N+1)}\sum\frac{(\sum R)^2}{n} - 3(N+1)$$

Where N = Total sample size
n = sample size of each group
$\sum R$ = sum of ranks for each group

Note that

$$\sum\frac{(\sum R)^2}{n} = \frac{\sum(R1)^2}{n} + \frac{\sum(R2)^2}{n} \ldots \frac{\sum(Rn)^2}{n}$$

Example 20.3

Prior to the introduction of social science statistics as a required course in the faculty of social sciences, lecturers were asked to indicate whether they wanted it taught as a faculty course, or as individual department course. The degree to which individual lecturers were predisposed towards a pooled system of teaching the course was measured on a scale of zero-to thirty (0-30). It was suspected that opinion will differ on the basis of academic departments within the faculty. Therefore, a stratified probability sampling was used to select respondents from each department within the faculty. The sample plan and measurement procedure produced the information below.

Table: Attitude of lecturers towards pooled teaching of faculty statistics by academic departments

Eco (n_1 = 5)		*Geo (n_2 = 5)*		*S sc (n_3 = 6)*		*Psy (n_4 = 4)*		*Pol. Sc (n_5 = 5)*	
Sci$_1$	*Rank$_1$*	*Sci$_2$*	*Rank$_2$*	*Sci$_3$*	*Rank$_3$*	*Sci$_4$*	*Rank$_4$*	*Sci$_5$*	*Rank$_5$*
24	4	25	2	22	6	23	5	21	7.5
12	15	15	12	16	11	8	20	10	19
17	10	9	18.5	13	14	3	24	6	21
18	9	9	18.5	28	1	2	25	4	23
21	7.5	11	16	14	13			5	22
				25	3				
$\sum R_1 = 45.5$		$\sum R_2 = 67.0$		$\sum R_3 = 48.0$		$\sum R_4 = 74.0$		$\sum R_5 = 90.5$	

To conduct the test we set the following hypothesis

Ho: (Ho =0) The attitude of lecturers in favour of pooled teaching of faculty statistics is not likely to differ on the basis of academic departments.
Hi: (Hi ≠0) The attitude of lecturers in favour of pooled teaching of faculty statistics is likely to differ on the basis of academic departments.

Given $p = 0.01$ (2 tailed) we substitute as follows:

$$H^* = \frac{12}{25(25+1)}\left[\frac{(45.5)^2}{5} + \frac{(67.0)^2}{5} + \frac{(48.0)^2}{6} + \frac{(74.0)^2}{4} + \frac{(90.5)^2}{5}\right] - 3(25+1)$$

$$= \frac{12}{650}\left[\frac{12070.25}{5} + \frac{4489}{5} + \frac{2304}{6} + \frac{5476}{4} + \frac{8190.25}{5}\right] - 78$$

$$= \frac{12}{650}\left[1414.05 + 897.8 + 384 + 1369 + 1638.05\right] - 78$$

$$= 0.01846154\,[4702.9] - 78$$

$$= 86.82 - 78$$

$$H^* = 8.82$$

To test the significance of the H- statistics we need to determine the degree of freedom (df) which is given by

$K - 1$

Where k = the number of sample groups. Based on the data above, degree of freedom df = $5 - 1 = 4$

In most cases the Kruskal Wallis H-test uses the chi-square (χ^2) table, therefore the table value of H at d.f. = 4, and p = 0.01 for two tailed test is 13.277

Decision Rule
If $H^* > H$, reject Ho
If $H^* < H$, accept Ho

Recall that *Ho: H = 0* = no difference
Hi: h ≠0 = they differ

Recall also that $H^* = 8.82$ and
$H_t = 13.277$

Since H* is less than H-tabulated, we accept Ho which states that the attitude of lecturers in favour of pooled teaching of faculty statistics is not likely to differ on the basis of academic departments.

Therefore, we can conclude that there is no significant departmental difference in attitude regarding the pooled teaching of faculty statistics course among the lecturers.

REFERENCES

Agburu J.I (2001). Modern Research Methodology. Makurdi. Solid Printing and Publishing Company Limited.

Allen R.G.D. (1996). Statistics for economists. London Hutchison and co. Publishing Limited.

Allen. R.G.O. (1966). Mathematical Economics. 2ns ed. Macmillan. London.

Almon. S. (1965). "The Distributed Lag between Capital Appropriations and Expenditures". Econometrical. Vol.30. pp. 178-96.

Arise Spanos (1968) Probability theory and statistical inference: Econometries modeling with observational data.

Amemiya. T. (1985). Advanced Econometries. Basil Blackwell. Oxford..

Anderson. R.L. (1952) and Bancroft. T.A.. Statistical Theory in Research. New York: McGraw Hill.

Barnett. R. A. (1979). Finite Mathematic, 2nd edition: Califonia: Dellen publishing company.

Basmann. R.L.. (1957). "A Generalized Classical Method for Linear Estimation of Coefficients in Structural Equation. Econometric. Vol.25. pp.77-83.

Damodar. Gujarati. (1970) Basic econometrics 2nd edition London.

Dominick Salvatore (1982) Theory and problems of statistics and Economics:

Damodar Gujarati N. (2005): Basic Econometrics: New delhi: Tata McGraw-Hill publishing company limited.

Desai. M. (1976). Applied Econometries. London: Philip – Allan..

Dhrymes. P.J. (1970). Econometrics: Statistical Foundation and Application New York: Harper and Row..

Dhrymes. P.J. (1978). Introductory Econometrics. New York: Springer – Verlag..

Durbin J. (1953). "A note on Regression when there is Extraneous Information on one of the Coefficient". Journal of the American Statistical Association. Vol.48. pp. 799-808.

Durbin. J.. and Watson. G.S.. "Testing for Serial Correlation in Least Squared Regression." Biometrica. Vol.37. pp. 409-28.

Epstein. R.J.. (1989) A History of Econometries. Amsterdam: North Holland Publishers.

Frank. C.R.J. (1971). Statistics and Econometrics. Holt. Rinehan and Winston.

Golberger Acthus S. (1998): introductory econometrics. Harvard University Press.

Gujarati Damdar N. (1999). New York: Essentials of Econometrics second edition McGraw-Hills.

Hill. Carter. William Griffiths and George Judge (2001): Undergraduate Econometrics New York. John Willis and sons.

Hu. The-Wei. (1973): Econometrics: An Introductory Analysis: Baltimore. University Park Press.

Katz. David A (1982): Econometric theory and Applications. Prentice Hall. Engle Wood,Cliffs. NJ.

Klein. Lawrence R: (1962): An Introduction to Econometrics. Prentice Hall. Engle-Wood Cliffs N.J.

Koop. Gary. (2000): Analysis of Economic data. New York: John Willis and Sons

Kothari C.R.(2009): Quantitative Techniques. New Dechi. Vikas Publishing House PVT Limited.

Shanim's outline series New York: MaCGraw Hill-INC.

Haires. Brain (1978) Introduction to Quantitative Economic: London: George Allen & Unwin.

Johnson J. (1963) Econometrics methods. 2nd Edition. Aukland International Students edition: MaCGraw Hill-INC.

Johnson S. (1991) Estimation and Hypothesis Testing of Intergration Vector: Econometrics London..

Oche E.S. (2011). Fundamentals of Research Methods and Statistics. Makurdi. Selfers Academic Press Limited.

Obama J.S. (1982). Glossary of Statistical Terms and formulae. Benin City. Koda Publishing Limited.

Omotosho M.Y (2000) Practical Econometrics Ibadan.

Oyeniyi T.A. (1997): Fundamental Principles of Economics: Lagos: Cedar Group Nigeria Limited.

Pannier Selvam (2010). Research Methodology. New Delhi PHI Learning Private Limited.

Speigel M.R et al (2009). Probability and statistics. New York. Mc.Graw-Hill. Companies Inc.

Tyriola M.F. (1998). elementary statistics (eleventh edition). Philadelphia. Addison Wesley Longman Inc.

Upender M. (1977) Applied Econometrics 1st Edition: London..

Panneerselvan R. (2000): Research methodology: New Delhi-[illegible] PHI Learning private limited.

Walters A.A. (1968): An introduction of Economics: London. Macmillan.

Farrar. D.E.. and Glauber. R.R.. "Multicollinearity in Regression Analysis. The problem Resisted. Review of Economics and Statistics. Vol. 49. pp. 92-107.

Freund. J.E.. and Walpole. R.E. (1987.). Mathematical Statistics. 4th ed.. New York: Prentice Hall International.

Fox K.A., (1958) Econometric Analysis for public policy. Iowa State College Press.

Goldheger. A.S. (1964.). Econometric Theory. New York: John Wisley & Sons.

Goldberger. A. S., (1991) A Course in Econometrics. Cambridge. Mass: Harward University Press..

Gujarati D., (1978) Basic Econometrics. 2nd edition. London: McGraw-Hill Publishing Company..

Haavelmo. Tryge. "The Role of the Econometrics in the Advancement of Economic Theory". Econometrica. Vol.26. No.3 pp.351-357.

Haavelmo. T., "The Probability Approach in Econometrics". Supplement to Econometrica. Vol.12. pp. 1-118.

Hadley. G.,(1965) Liner Algebra. Mass: Addition - Wesley..

Harvey. A. C., The Econometric Analysis of Time Series. 2nd ed., Philip. Allan Hasel Hempstead.

Johnston. J., (1984.) Econometric Methods. 3rd ed. New York: Mc Grill Hill.

Kadiyala. K.R. "A Transformation Used to Circumvent the Problem of Autocorrelation". Econometrica. Vol.36. No.1. 1968. pp. 93-96.

Kendall. M., and Ord. J. K., (1990) Time Series. 3rd edition. London: Editorial Arnold.

Kelejian. H.H., (1989) and Oats. W.E., Introduction to Econometrics – Principles and Applications. 3rd edition. New York: Harper & Row Publishers..

Klein. L.R. (1953). A Textbook of Econometrics. Row Peterson. Evanston. Ill..

Klein. L.R.,. Whither Econometrics. Journal of the American Statistical Association. Vol.66. 1971 pp. 415-421.

Kmenta. J.,(1986) Elements of Economics. 2nd edition: New York Macmillan..

Koopmans. T.C., Measurement Without Theory. Review of Economics and Statistics. Vol.29. 1974 pp.161-172.

Koutsoyiannis. A. (1977). Theory of Economics. 2nd edition. London: Macmillian Publishing Company..

Koyck. L.M.(1954). Distributed Lags and Investment Analysis. Amsterd: North Holland Publishing Company..

Kramer. W., Ploberger. W., and Alt. R., "Testing for Structural Change in Dynamic Models". Econometrical. Vo. 56. pp. 1355-1369.

Leamer. E.E.,(1983) "Let's take the con out of Econometrics". American Economic Review. Vol.23. pp.31-43. reprinted in Granger. C.W.J. (ed.) Modeling Economic Series: Readings in Econometric Methodology. Oxford University Press. Oxford.

Liu. T.C., "Underindentification. Structural Estimation and F... Econometrical. Vol.28. pp.835-865.

Maddala. G.S.(1981). (ed) "Model Selection". Journal of Econometrics.

Maddala. G.S. (1988.). Introduction to Econometrics. London Macmillan.

Mills. T.C.,(1990) Time Series Techniques for Economists. Cambridge University Press..

Mogan. M.S,(1990). The History of Economic Ideas. Cambridge: Cambridge University Press..

Nerlove. M.,(1958) Distributed Lags and Demand Analysis. Agricultural handbook. 141. US Department of Agriculture. Washington..

Nerlove. M.,(1963) Returns to Scale in Electricity Supply. in Christ. C., (ed) Measurement in Economics. Standford University Press..

Orcutt. G.H. et al.(1981.) Micro analysis of Socioeconomic Systems: A Simulation Study New York: Harper & Row. Publishers.

Pagan, A.R. (1987), "Three Econometric Methodologies: A critical Appraisal", Journal of Economic Surveys Vol.8, pp. 3-24. , reprinted in Granger, C.W.J., (ed) Modeling Economic series: Reading in Econometric Methodology. Oxford: Oxford University Press. 1990.

Quandt, R.E., "On Certain Small Sample Properties of K-class Estimators".

Roos, C.F., "Survey of Econometric Forecasting Techniques", Econometrica, Vol.23, 1955 pp.363-395.

Schumpeter, J.A. (1833), "The Common Sence of Econometrics", Econometrica, Vol I, No.1, pp.5-12.

Stewart, J., (1984) Understanding Economics, 2nd ed. Hutchinson Publishing Company.

Stone, J.(1951) and Richard, N., The Role of Measurement in Economics, Cambridge University Press,.

Theil, Henri (1953) Timation and Simultaneous Correlation in Complete Equation System. (The Hague: Central Planning Bureau,.

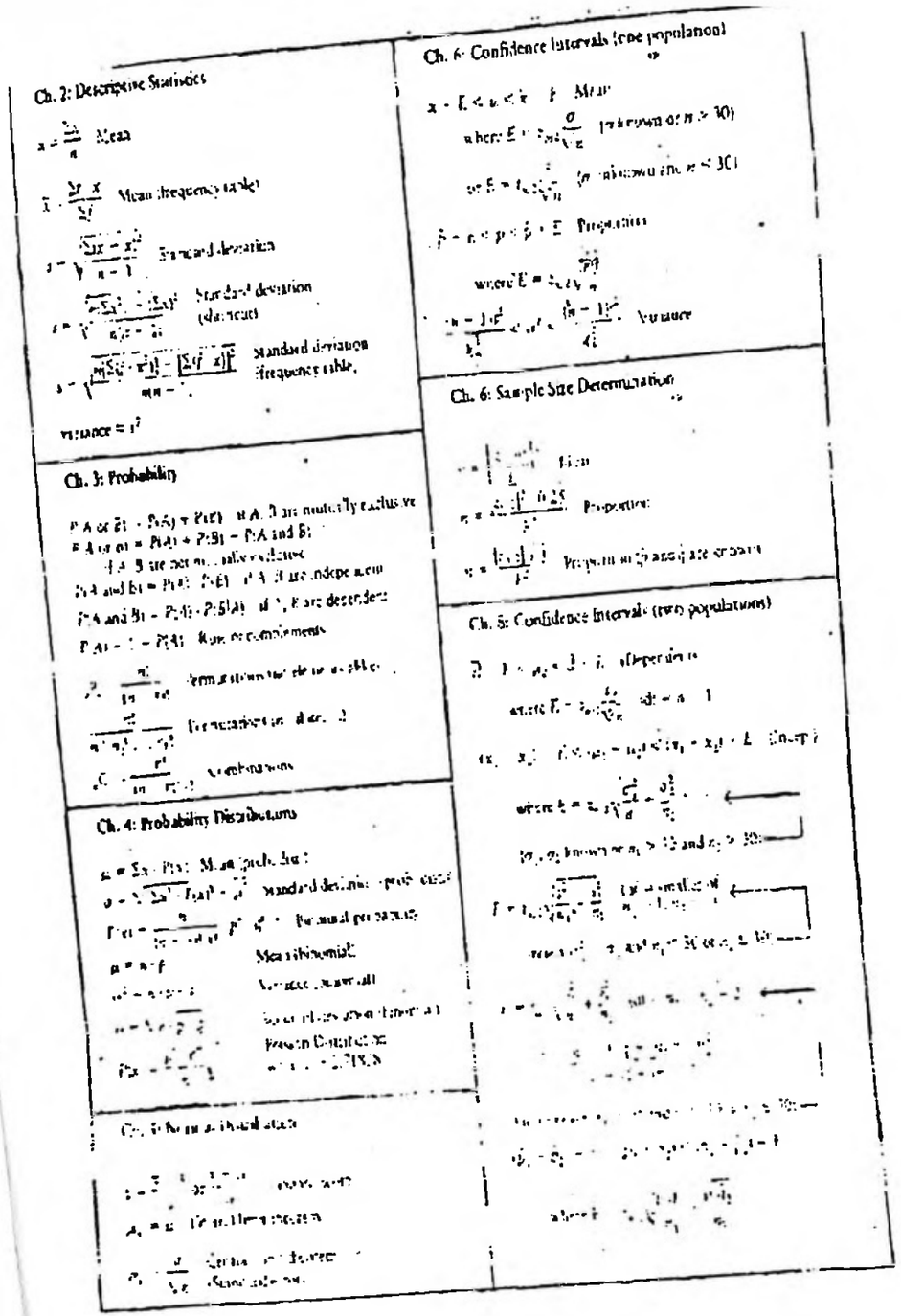
Theil, Henri, (1961) Economic Forecast and Policy 2nd ed., North – Holland Publishing Company,.

Thomas, R.L.,(1985) Introductory Economics: Theory and Application, New York, Longman Inc.

Walters, A.A.,(1970) An Introduction to Econometrics, 2nd ed., Macmillan Publishing Company,

Zellner, A., Huang, D.S and Chau, L.C., "Further Analysis of the Short-run Consumption Functions with Emphasis on the Role of Liquid Assets", Econometrical, Vol. 33, 1965 pp. 571-81.

Zellner, A., and Geisel, M.K., Analysis of Distributed Lag models with Applications to Consumption Function Estimation, Econometrical, Volume 38,, 1970 pp.865-88.



Ch. 7: Test Statistics (one population)

$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ Mean—one population (σ known or $n > 30$)

$t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ Mean—one population (σ unknown and $n \le 30$)

$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$ Proportion—one population

$\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$ Standard deviation or variance—one population

Ch. 8: Test Statistics (two populations)

$t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$ Two means—dependent: $df = n - 1$

$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ Two means—independent (σ's known); $n_1 > 30$ and $n_2 > 30$

$F = \dfrac{s_1^2}{s_2^2}$ Standard deviation or variance—two populations (where $s_1^2 \ge s_2^2$)

$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ (df = smaller of $n_1 - 1, n_2 - 1$)

Two means—independent (reject $\sigma_1^2 = \sigma_2^2$ and $n_1 \le 30$ or $n_2 \le 30$)

$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$ ($df = n_1 + n_2 - 2$)

where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Two means—independent (fail to reject $\sigma_1^2 = \sigma_2^2$ and $n_1 \le 30$ or $n_2 \le 30$)

$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\bar{p}\bar{q}}{n_1} + \dfrac{\bar{p}\bar{q}}{n_2}}}$ Two proportions

Ch. 10: Multinomial and Contingency Tables

$\chi^2 = \sum \dfrac{(f_o - f_e)^2}{f_e}$ Multinomial ($df = k - 1$)

$\chi^2 = \sum \dfrac{(f_o - f_e)^2}{f_e}$ Contingency table ($df = r - 1)(c - 1$)

where $f_e = \dfrac{\text{row total} \times \text{column total}}{\text{grand total}}$

Ch. 9: Linear Correlation/Regression

Correlation $r = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

$b_1 = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

$b_0 = \bar{y} - b_1\bar{x}$ or $b_0 = \dfrac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n\Sigma x^2 - (\Sigma x)^2}$

$\hat{y} = b_0 + b_1 x$ Estimated eq. of regression line

$r^2 = \dfrac{\text{explained variation}}{\text{total variation}}$

$s_e = \sqrt{\dfrac{\Sigma(y - \hat{y})^2}{n-2}}$ or $\sqrt{\dfrac{\Sigma y^2 - b_0\Sigma y - b_1\Sigma xy}{n-2}}$

$\hat{y} - E < y < \hat{y} + E$

where $E = t_{\alpha/2}\, s_e\sqrt{1 + \dfrac{1}{n} + \dfrac{n(x_0 - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$

Ch. 11: One-Way Analysis of Variance

$F = \dfrac{ns_{\bar{x}}^2}{s_p^2}$ k samples each of size n (num. $df = k - 1$, den. $df = k(n-1)$)

$F = \dfrac{MS(\text{treatment})}{MS(\text{error})}$ ← num. $df = k - 1$; ← den. $df = N - k$

$MS(\text{treatment}) = \dfrac{SS(\text{treatment})}{k - 1}$

$MS(\text{error}) = \dfrac{SS(\text{error})}{N - k}$ $MS(\text{total}) = \dfrac{SS(\text{total})}{N - 1}$

$SS(\text{treatment}) = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2$

$SS(\text{error}) = (n_1 - 1)s_1^2 + \cdots + (n_k - 1)s_k^2$

$SS(\text{total}) = \Sigma(x - \bar{\bar{x}})^2$

$SS(\text{total}) = SS(\text{treatment}) + SS(\text{error})$

Ch. 11: Two-Way Analysis of Variance

Interaction: $F = \dfrac{MS(\text{interaction})}{MS(\text{error})}$

Row Factor: $F = \dfrac{MS(\text{row factor})}{MS(\text{error})}$

Column Factor: $F = \dfrac{MS(\text{column factor})}{MS(\text{error})}$

Ch. 13: Nonparametric Tests

$z = \dfrac{(x - 0.5) - (n/2)}{\sqrt{n}/2}$ Sign test for $n > 25$

$z = \dfrac{T - n(n+1)/4}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$ Wilcoxon signed ranks (two dependent samples for $n > 30$)

$z = \dfrac{R - \mu_R}{\sigma_R} = \dfrac{R - \dfrac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$ Wilcoxon rank-sum (two independent samples)

$H = \dfrac{12}{n(n+1)}\left(\dfrac{R_1^2}{n_1} + \dfrac{R_2^2}{n_2} + \cdots + \dfrac{R_k^2}{n_k}\right) - 3(n+1)$

Kruskal-Wallis (chi-square $df = k - 1$)

$r_s = 1 - \dfrac{6\Sigma d^2}{n(n^2 - 1)}$ Rank correlation

(critical value for $n > 30$: $\dfrac{\pm z}{\sqrt{n-1}}$)

$z = \dfrac{G - \mu_G}{\sigma_G}$ $r = \dfrac{G - \dfrac{2n_1n_2}{n_1 + n_2} + 1}{\sqrt{\dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}}$ Runs test for $n > 20$

Ch. 12: Control Charts

R chart: Plot sample ranges

UCL: $D_4\bar{R}$

Centerline: $\bar{R}$

LCL: $D_3\bar{R}$

x chart: Plot sample means

UCL: $\bar{\bar{x}} + A_2\bar{R}$

Centerline: $\bar{\bar{x}}$

LCL: $\bar{\bar{x}} - A_2\bar{R}$

p chart: Plot sample proportions

UCL: $\bar{p} + 3\sqrt{\dfrac{\bar{p}\bar{q}}{n}}$

Centerline: $\bar{p}$

LCL: $\bar{p} - 3\sqrt{\dfrac{\bar{p}\bar{q}}{n}}$

TABLE A-6
Critical Values of the Pearson Correlation Coefficient r

n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	330
70	.236	.305
80	.220	.286
90	.207	269
100	195	.256

NOTE: To test H_0: $\rho = 0$ against H_1: $\rho \ne 0$, reject H_0 if the absolute value of r is greater than the critical value in the table.

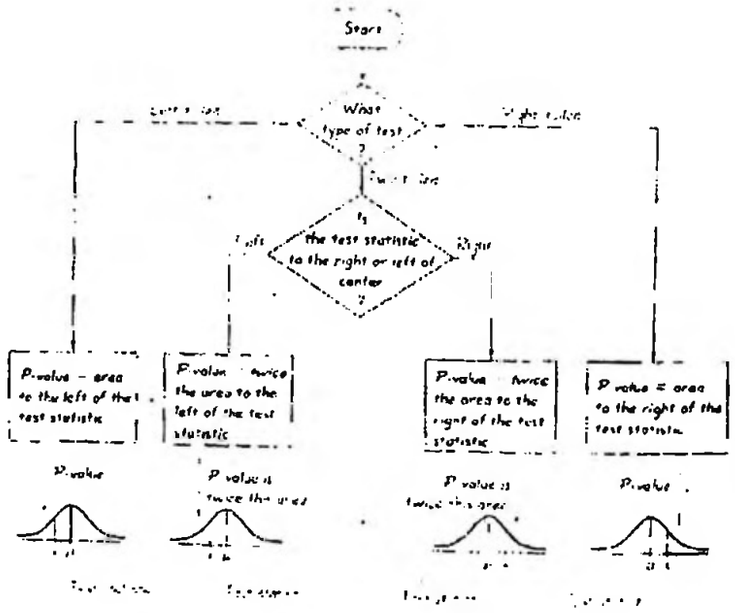
Control Chart Constants

subgroup Size n	A_2	D_3	D_4
2	1.880	0.000	3.267
3	1.023	0.000	2.574
4	0.729	0.000	2.282
5	0.577	0.000	2.114
6	0.483	0.000	2.004
7	0.419	0.076	1.924

TRADITIONAL METHOD OF HYPOTHESIS TESTING

1. Identify the specific claim or hypothesis to be tested and put it in symbolic form.
2. Give the symbolic form that must be true when the original claim is false.
3. Of the two symbolic expressions obtained so far, let the null hypothesis H_0 be the one that contains the condition of equality; H_1 is the other statement.
4. Select the significance level α based on the seriousness of a type I error. Make α small if the consequences of rejecting a true H_0 are severe. The values of 0.05 and 0.01 are very common.
5. Identify the statistic that is relevant to this test, and identify its sampling distribution.
6. Determine the test statistic, the critical value, and the critical region. Draw a graph and include the test statistic, critical value(s), and critical region.
7. Reject H_0 if the test statistic is in the critical region. Fail to reject H_0 if the test statistic is not in the critical region.
8. Restate this previous conclusion in simple, nontechnical terms.

FINDING P-VALUES



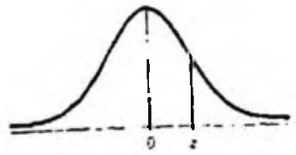


TABLE A-2 Standard Normal (z) Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	
1.6	.4452	.4463	.4474	.4484	.4495 *	.4505	.4515	.4525	.4535	
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949 *	.4951	
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	
3.1										
3.2										
3.3	.4995									
higher										

Note: For values of z above 3.09, use 0.4999 for the area.

*Use these common values that result from interpolation:

z score	Area
1.645	0.4950
2.575	0.4995

From Frederick C. Mosteller and Robert E. K. Rourke, Sturdy Statistics, 1973, Addison-Wesley Publishing Co., Reading, MA. Reproduced with permission of Frederick Mosteller.

6,202

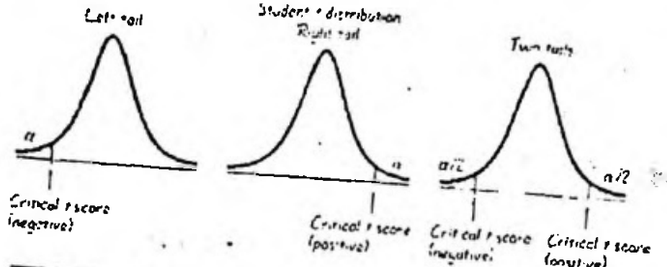


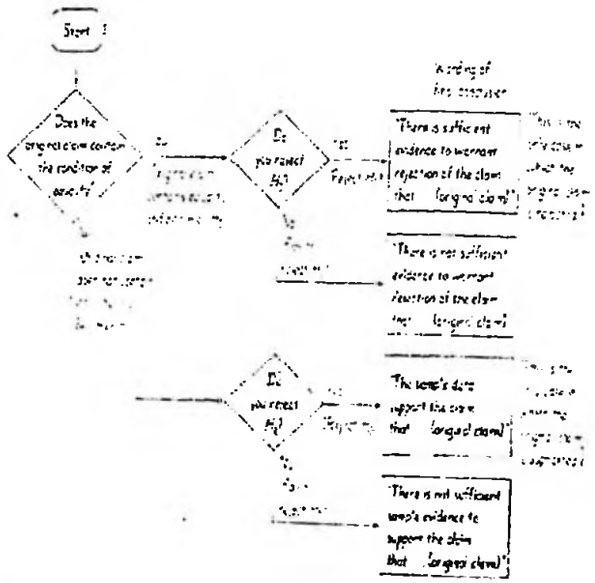
TABLE A-3 t Distribution

Degrees of freedom	.005 (one tail) .01 (two tails)	.01 (one tail) .02 (two tails)	.025 (one tail) .05 (two tails)	.05 (one tail) .10 (two tails)	.10 (one tail) .20 (two tails)	.25 (one tail) .50 (two tails)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.131	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.539	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.787	2.485	2.060	1.708	1.316	.684
26	2.779	2.479	2.056	1.706	1.315	.684
27	2.771	2.473	2.052	1.703	1.314	.684
28	2.763	2.467	2.048	1.701	1.313	.683
29	2.756	2.462	2.045	1.699	1.311	.683
Large (∞)	2.575	2.327	1.960	1.645	1.282	.674

TABLE A-4 Chi-Square (χ^2) Distribution

Degrees of freedom	Area to the Right of the Critical Value 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

HYPOTHESIS TEST: WORDING OF FINAL CONCLUSION



INDEX

A
A priori criteria 16
Abuse of statistic 2
Advantages of personal interview 26
Age interval 44
Analysis of variance 156
Anxious of probability 65
Applied research 7
Approach to probability 62
Appropriate sequence of questions 25
Arithmetic mean 43
Assignment of probability 64

B
Bar chart 35
Bayer's theorem 71
Binomial distribution 85

C
Class intervals 33
Class boundary 33
Conclusive research 8
Criteria for verification of a model 16
Close ended interview 26
Cumulative frequency 32,41
Chart 34
Calculation of probability 64
Central limit theorem 105
Chi-square test 249.
Chow test 243
Coefficient of probability 65
Combination 77
Complement of a sat 59
Component bar chart 37
Compound events 66
Conditional model 176
Conditional probability 70
Confidence in interval 136
Correlation theory 169
Counting 73
Critical value 144
Curve fitting 164

D
Data.gathering and confrontation 15
Definition of research problem 13
Degree of freedom 148
Deli deciding question words 25
Dependent variables 6,167
Descriptive research 8
Desirable properties of a good estimator 135
Desirable properties of a good model 18
Determination of binomial experiment 87

E
Econometric criteria 17
Efficient estimator 136
Element of a set 56
Elements of measurement 4
Empty set 56,59
Error type one 143
Essentials of a good interviewer 27
Estimation 135
Estimation of parameter 12
Experimental research 8
Explained sums of square 193
Exponential research 8
Extensive Litrature survey 10
Interval estimate 136
Point estimate 136

F
Factorial rule 74
Forecasting 17
Formulation of questions 24
Frequency distribution 29
Frequency polygon 40
Fundamental research 8

Fundamentals of probability 61

G
Gauss Markov theorem 205
Guidelines for choice of questions word 25
Guidelines for conducting observation 22

H
Histogram 39
Hypothesis testing 40

I
Independent variable 6.147
Initial survey of literature 10
Inter sect 56
Interval estimation 136
Interval scale 4,6

J
Judgment sampling 126

L
Laboratory observation 21
Less than orgive curve 41
Levels of measurement 3
Lin Long model 223
Linear trend model 215
Log-tin model 222

M
Margine of error 138
Mean deviation 51
Mean test of hypothesis 162
Measure of dispersion 48
Measurement 3
Measurement of central tendency 43
Measurement techniques 5
Median 46
Method of data collection 20
Methodology of economic research 12
Midpoint 44
Minimum variance e 136
Modeling research 8
More than orgive curve 42
Multiple regression 228
Multi-stage sampling 118,125
Mutually exclusive events 65

N
Non laboratory observation 21
Non-normal distribution 102
Non-parametric test 256
Non-probability 80
Non-stochastic assumption 181
Norminal scale distribution 102
Normal OLS equations 281
Normal scale 3
Notations of probability 63
Null set 56

O
Observation method 20
Omission of variable 245
Open ended interview 26
Ordinal scale 3,4
Orgive curve 41

P
Parameters 14
Permutation 74
Personal interview 26
Pie chart 34
Poison distribution 89
Pre-testing of questions 25
Probability distribution 80
Probability sampling 118
Probability theory 56
Properties of normal distribution 95

Q
Quartile 49
Questionnaire design 24

Questionnaire method 20
Quota sampling 120

R
Random variable 80
Range 48
Rank correlation 174
Ratio scale 3,5
Rationale for sampling 110
Reasons for poor performance of a model 19
Relative frequency 33
Research design 12
Research methodology 7
Review of questions for improvement 26
Rule of thumb 203

S
Observation 20
Sample size 112
Sampling design 111
Sampling techniques 116
Sampling theory 110
Scientific and non scientific observation 21
Selection of research problems 10
Set theory 56
Sign test 257
Simple bar chart 36
Simple random sampling 118
Simple regression 176
Sources of data 11
Specification of econometric model 14
Specification of mathematical model 13
Statistical bias 245
Statistical criterion 171
Stochastic assumptions 179
Sufficient estimator 136
Symbolic research/simulation research 8

T
Techniques of data presentation 29
Telephone interview 22
Test of hypothesis 140
Test of significant of estimates 156,202
Theorems of probability 71
Types of research 7
Types of scale of measure

U
Unbias estimator 136
Unconditional probability 47
Union of a set 57
Uses of estimated model 19
Uses of statistics 1

V
Variables 14
Verification of the model 16