ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/342247154

Fundamentals of Statistics in Education

Book · January 2020

CITATIONS
0

READS
9,759

1 author:

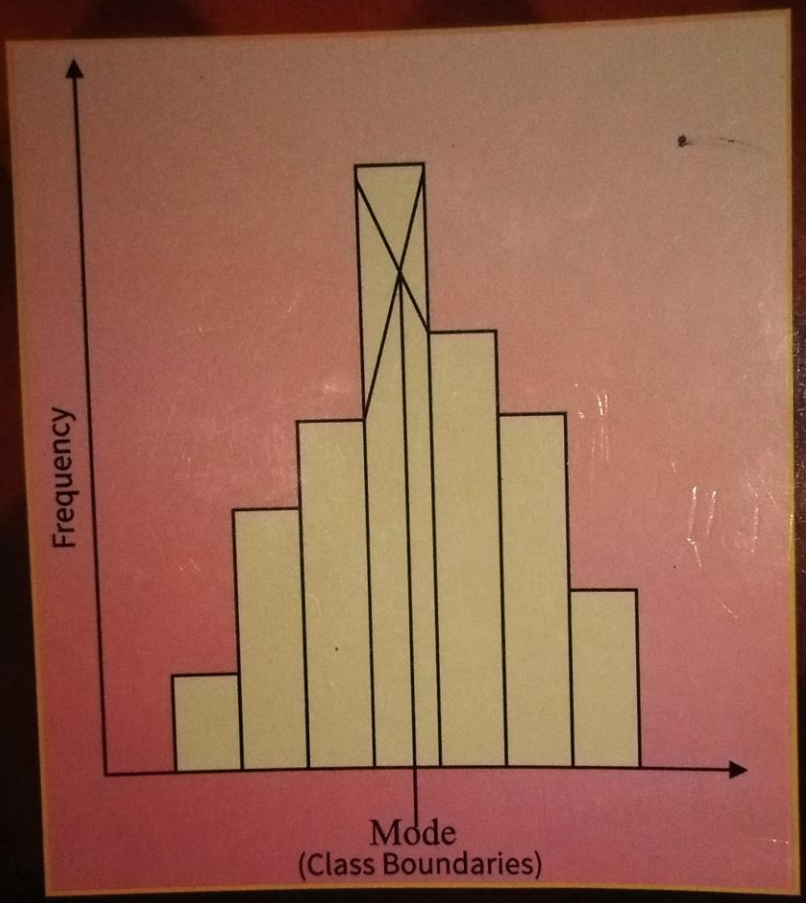
Isaac Ofem Ubi
University of Calabar
44 PUBLICATIONS **105** CITATIONS

SEE PROFILE

All content following this page was uploaded by Isaac Ofem Ubi on 16 January 2022.

The user has requested enhancement of the downloaded file.

FUNDAMENTALS OF STATISTICS IN EDUCATION



ISAAC OFEM UBI

FUNDAMENTALS OF STATISTICS IN EDUCATION

ISAAC OFEM UBI

Copyright © Isaac Ofem Ubi 2017. *Fundamentals of Statistics in Education*

Published by:
University of Calabar Press
Calabar – Nigeria.

ISBN:

All Rights Reserved:

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

Typesetting was by
Metubas Computers, Calabar
+234 806 894 2306

DEDICATED TO

My Wife

PREFACE

This book has been designed to provide good understanding of basic and advanced statistics in education and is particularly relevant to undergraduate and graduate students running degrees in Education. The book can also be relevant to NCE students of Colleges of Education. Examples used in the book are local and relevant to Nigeria and other English speaking West African Countries.

The book has been written with numbered sub-headings, examples and with review questions at the end of each chapter. This approach has been seen as going to be effective among a significant number of users, as the book can be used for independent study or in conjunction with classroom instruction.

For ease of study, users should read chapters of the book sequentially, study the examples of each chapter and practice the questions in the exercise. The questions are a collection of past semester examination questions.

I recommend the book to all my students and students of my colleagues offering courses in Educational Statistics in the University of Calabar and beyond. It will serve as a prerequisite for courses in Educational Research Methods. Make it a Companion.

Isaac O. Ubi
October, 2017.

ACKNOWLEDGEMENTS

First and foremost, I acknowledge with reservation the Almighty God in Heaven for giving me the grace to write this text from conception to delivery. I am indebted to all my Mentors, to wit; Prof Abang J. Isangedighi, Prof Monday T. Joshua, Prof Alice E. Asim, Prof Imo E. Umoinyang and Dr Samuel M. Akpan.

I am also indebted to Dr Idaka Idaka, who had, about a year ago, asked me to co-author a text on statistics with him. His request encouraged me to dust the manuscript I had completed then and send it to editors. Also remembered is my only classmate at the M.Ed and Ph.D (Measurement & Evaluation) programmes many years back, Dr Bassey A. Bassey, for the ideas we have shared and still share in Educational Statistics.

I acknowledge, with all sincerity, all my other contemporary colleagues and academic brothers like Dr Emmanuel Ekuri, Dr German E. Anagbogu, Dr Clement O. Ukpor, Dr Idoreyin Akubuiro, Dr Fidelis Obo, Dr Delight Idika, Dr Ojini, Dr Julius Egbai Mrs Sylvia Ovat, Dr Peter U. Bassey and Dr Maureen Okon, for all the academic ideas we share.

This acknowledgement will not be complete without reference to Chief Ernest I. Anani for painstakingly editing the manuscript, Colonel Jubi J. Iferi (Rtd), and General Enang M. Ukagu (Rtd.) for their useful suggestions each time an issue on the book was raised.

TABLE OF CONTENTS

Chapter One

MEANING AND SCOPE OF STATISTICS

1.1 ORIGIN OF STATISTICS

The concept of statistics evolved with man. From time immemorial, individuals, governments and segments of the society have used statistics to take various forms of decisions. Our fore-fathers, for instance, used tally system to count and keep their money in bundles for easy accountability. Governments, even in the old ancient Egypt, used statistics from censuses in taking various decisions like the construction of pyramids and to know the population of males as distinct from those of females.

Besides all of these facts about tracing statistics to have originated at the beginning of man, there are different comments by different authors as to the specific time statistics, as a discipline, can be said to have started. Some people feel statistics started in the 17^{th} century with John Graunt who observed bills of mortality in Europe by attempting to give rates of deaths per year and working out ratios of such deaths to their possible causes. At that time the most possible causes of death were, by his statistical collection, any ailment and plague. Somebody was believed to either die of an

ailment or of a plague. Such statistics were kept by John Graunt and today some authorities feel that was the origin of statistics.

Some other people accredit the real beginning of statistics, as a discipline, to the 18th century when the meaning of statistics was restricted to information about a state or country. During that period, the term statistics meant the systematic collection of demographic and economic data by different states in Europe and the United States of America. Going by this school of thought, statistics broadened in the 19th century when its meaning began to include collection, summarizing, and analyzing of data. Statistics as a discipline kept expanding to the extent that statistics are now being computed and widely used by governments than they were used years back.

In Nigeria, the first attempt at statistical record keeping was in 1866 when the colonial masters attempted a numerical estimate of the Nigerian population. This attempt was regionally based and no information is available as to whether the census covered all of what is now called Nigeria. The second attempt was in 1952 and since then the country has used census figures to produce statistical data that are useful for economic planning. Various governments in the country have used the discipline of statistics to promote economic activities since after independence. Today, as it is in several other countries the world over, the National Bureau of Statistics (NBS) has been established in Nigeria to develop statistics in all spheres of endeavours including economic growth, health, gender matters, GDP capital importation, education, agriculture, oil production, public administration, the private sector, electronic payments of banks; you can keep naming them.

1.2 MEANING OF STATISTICS

Statistics has been defined in different ways by different authorities. The definitions are so many that recent writers prefer to list as many different definitions as possible to allow their readers appreciate the meaning, scope and limitations of the subject. Gupta (2013) gave two classifications for such a variety of definitions.

First, the field of utility of statistics has been increasing steadily and thus different people define it differently to meet the development of the subject. For instance, statistics some years back was regarded as the science of statecraft but today statistics has embraced almost all areas of natural and human activity. Based on this, the old definitions which were focused on a limited and narrow field of enquiry have been replaced by new definitions which are wider in scope and approach.

Secondly, the word statistics has been used to convey different meanings in singular and plural sense. When the word is used in plural sense, statistics mean numerical set of data and when it is used in singular sense; statistics mean the science of statistical method involving the theory and techniques used for collecting, analyzing and drawing inferences from the numerical data.

Statistics as defined in Spiegel and Stephens (2011) is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as with drawing valid conclusions and making reasonable decisions on the basis of such analysis. The same text has also narrowed the term statistics to denote the data themselves or numbers derived from the data. This

second definition is similar to Afonja (2001) who defined statistics simply as numerical data.

In a more holistic sense, statistics may be defined as the aggregate of facts affected by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other (Gupta, 2013). This definition is very exhaustive and can be explained under the following details:

i) **Statistics are aggregate of facts:** This means that isolated facts cannot be termed as statistics except if such facts are part of the total facts relating to any particular field. For example, the aggregate of scores of children according to their age brackets can be considered as statistics. Contrary to this, we cannot consider a child's examination score and the demand for a commodity as statistics since the two amounts are unrelated.

ii) **Statistics are affected by a multiplicity of causes:** This means that facts are statistics when the figures are affected by many factors in the field of education and the social science. For instance, the score of a child in Mathematics should be the resultant effect of attitude to Mathematics, the child's intelligence, peer group pressure, home factors, age, gender, the environment, etc. It is normally difficult to say which portion of the score was the result of which factor. This is not the same as in the physical sciences where one can isolate the effect of various factors on one single item.

iii) **Statistics are numerically expressed:** This means that for any fact to be statistics it must be in numerical terms. That is, qualitative expressions which cannot be measured

quantitatively like 'the police crime investigation department; the women wing of a political party, attitude, interest, motivation and intelligence cannot be seen as statistics. Once these expressions are numerically expressed by assigning scores to them they can be termed statistics.

iv) **Statistics are estimated according to reasonable standard of accuracy:** Data for statistics must be obtained from the entire population to ensure good standard of accuracy. If the population is large or infinite, a reasonable sample must be drawn from it to arrive at representative statistical results.

v) **Statistical data are collected in a systematic manner** Systematic procedure must be employed in collecting data. That is, for any research, proper method that fits the system of research should be adopted in collecting data. Trained personnel who understand the survey should be used for collection of statistical data.

vi) **Statistical data should be collected for a predetermined purpose:** Data for statistics must be collected based on the objectives of the enquiry. Only data that are required for any given research should be collected. While trying to collect only essential data, those that are not required should not be collected.

vii) Comparability: Statistical analysis should be done only on data that are comparable. It is not correct, for instance for analysis to be carried out on the comparison between the number of students who passed an examination and the population of domestic animals in the community hosting their school. This comparison will not make any meaning. However, a comparison between the number of students

who passed an examination and the qualifications of their teachers can make so much meaning in statistics.

1.3 FUNCTIONS OF STATISTICS

This text identifies eight functions of statistics:

1. **Statistics helps us to present facts in definite form**

With statistics, information about issues is presented in their true picture instead of in a manner that one will begin to guess the true situation of things. One can give figures to show the difference between the performance of students in physics and chemistry instead of just saying that "the students' performance in physics was different from their performance in chemistry". In the former, one can say that the mean performance of the students in physics (say 53.8) is higher than their mean performance in chemistry (say 50.6) and that will be a definite statement. In the later instance, saying that the students performed better in physics than in chemistry is not definite. It is with statistics that one can be definite in his facts.

2) Statistics helps to make facts precise

Since statistics are presented in a definite form, it helps in condensing the data into important figures. This means that statistics helps in simplifying complex data to make them understandable. As statistical data are presented in graphs, pictograms, averages like the means, or inches, etc, it becomes easy for people to understand them.

3) Statistics can be used for comparisons

Statistics as a means of carrying out comparison among variables is an important function. Mathematical quantities are used to show

the relationships that exist between or among the facts so collected. Facts presented in their absolute sense make no meaning until they are reduced into mathematical qualities like means, standard deviations, coefficients, etc.

4) Statistics are useful in formulation and testing of hypotheses

Statistics techniques can be used in arriving at new theories through the test of hypotheses. A good example is to use statistics in finding out the relationship between provision of corporate social responsibility by a company to its host communities and community development. Another example is to statistically find the difference between male and female students' performances in English language. The findings of such proportions form theoretical bases for inferences to be drawn.

5) Statistics are used for forecasting

Statistics provide information that can be very useful in making policies for the future. We can predict the future course of action based on facts available from statistics. In an ideal situation, for instance, the scores obtained by students in a mock examination can be the basis for predicting scores in the final school certificate examination. We can predict the demand for rice in 2018 if we know the population in 2017 with regards to growth rate.

6) Policy making

Statistics are very important for decision making or policy formation. If government intends to do free registration of students in the West African Senior School Certificate Examination (WASSCE), for instance it will need to first know how many

students that are in the certificate class. Such information is provided by statistical facts. Without these facts, it will be very difficult to make the policy.

7) Knowledge enhancement

This is a very important function of statistics. Constant operation with statistical facts helps in widening the knowledge of a person. It makes one think and reason better than he/she would have without statistical facts.

8) Statistics are used to measure uncertainty

Any issue about the future is an uncertainty situation. Statistics helps in making correct estimate about such issues whether in the present or in the future. If for instance, government wants to find out the amount of jobs needed in five years' time, it is statistics on the number of graduates and their different specialties for the period in question that may constitute the source of such uncertain estimates.

1.4 TYPES OF STATISTICS

The simplest way to understand the types of statistics is to categorize it in two pairs and according to the differential functions those pairs perform. These pairs are descriptive and inferential statistics, and parametric and non-parametric statistics.

Descriptive and non-descriptive statistics

When considered on the basis of major cluster of functions, statistics could be categorized into two types- descriptive and inferential. Descriptive statistics involves organizing, summarizing, tabulating and describing collections of data such as

test scores, income earned, GDP, population demographics, frequency counts, ranks and others in this category which can be represented in numerical data or tables or graphs. Descriptive statistics include frequency distribution, measures of central tendency, measures of variability, measures of relationship and measures local standing.

Inferential statistics are those that involve conclusions, summations and extrapolations arrived at scientifically based on available data. Inferential statistics enable us to draw conclusions or make deductions (inference) on a population based on results obtained from a sample of that population. This means that within inferential statistics, one is trying to reach conclusions that extend beyond the immediate data alone. We may use inferential statistics, for instance, to try to infer from the sampled data what the population might think. Put in some other way, we may use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in the given study. From the explanation, it means, inferential statistics are basically from a representative sample of the population.

As a comparison, therefore, we can say that whereas we use inferential statistics to make inferences from our data to more general conditions, we use descriptive statistics simply to describe what is going on in our data. Examples of inferential statistics are z-test, t-test, analysis of variance, chi-square, Mann-Whitney test and, as in some text, regression analysis.

Parametric and non-parametric statistics

Another way to consider types of statistics is to dichotomize it into parametric and non-parametric statistics. By this dichotomy, parametric statistics deals with data obtained from normal distributions (Anikweze, 2013). A parametric statistical test is a test whose model specifies certain conditions about the parameters of the population from which the samples are drawn (Gupta, 2013). In a parametric statistics, operations such as the mean, the standard deviation and other descriptive statistical operations are possible. Data for parametric statistics must meet assumptions of random sampling, equality of variance, normality of distribution, and addition of total variance.

Non-parametric statistics are those whose data must not be normally distributed. Data for non-parametric statistics are either nominal or ordinal. An example is a survey in which a researcher wants to know the relationship between sex (male and female) and political party rating of students. Let's assume that ratings are excellent, very good, good, poor. This type of statistics can be used without the mean, sample size, standard deviation or estimation of any other descriptive statistics when information about them does not exist.

1.5 LIMITATIONS OF STATISTICS

As important as statistics is in different areas of life endeavour, there are several limitations that make its scope narrow. These limitations include the following:

1. Statistics do not handle qualitative information

Statistics is the science that deals with numerical data. One can even say that statistics are numerical statement. This implies that statistics can apply to only phenomena that are measured in numerical quantity. It cannot be used to express qualitative information like when we say the girls performed better than the boys in their English Language test or the company has a very cordial relationship with people of the host community. To be able to statistically handle such information, one must first and foremost reduce the statements into figures.

2. Statistics cannot be individualistic

A single score or figure is not regarded as statistics unless it is part of the aggregate of other scores or figures relating to any particular field of enquiry. This implies that statistics as methods used in different fields of endeavour do not make use of single objects or persons or events. This is a limitation. Figures like a person's score on mathematics, the price of rice, the profit made by a company, the admission figure of a university are all examples of individual and isolated figures that are not connected or related, thus cannot be termed statistics.

3. Statistical laws are inexact

Statistical laws are based on probability and because of this, inferences drawn from statistical results are considered as approximate and not exact. Laws in statistics are true only on the average. This can be more understood if we use the example given in Gupta (2013).

If the probability of getting a head in a single through of a coin is $^1/_2$, it does not imply that if we toss a coin 10 times, we shall get five heads and five tails; in 10 throws of a coin we may get eight heads, 9 heads or all the 10 heads, or we may not even get a single head (pp 13).

4. It is possible to use statistics wrongly

Liability to wrongful use of statistics is a major limitation of statistics. Since it deals with figures which are innocent and cannot be looked at facially and distinguished by their quantity, statistics can easily be distorted and wrongly used by experts. Politicians, dishonest or unskilled workers according to Gupta can manipulate or mould statistical figures for selfish interest and they can be accepted as real. If for instance, election results are spurious, the statistics that emanate from them cannot tell the user what happened during the elections.

1.6 IMPORTANCE OF STATISTICS TO SOCIAL SCIENCES AND EDUCATION

The importance of statistics to social sciences like economics, political science, geography, sociology, management, accounting, finance, marketing and education cannot be overemphasized. Statistics as the method of judging collective natural or social phenomenon from the results obtained from analysis or enumeration or collection of estimates permeates all facets of the social organism. Every discipline in the social sciences and education has a multiplicity of factors which warrant a lot of observations. It is statistical tools that are used in collecting, organizing and analyzing data that emanate from those observations. Students' scores in various examinations, their

progress over time in school and their predicted performances can only be presented in reasonable summaries through statistics. Records of school inventories, classification of teachers according to departments and records of numerical progress by teachers and different schools can easily be kept by government through statistics. Economic planning, political history, geographical information on weather forecast, the gross domestic product, trends in market prices and records of corporate social responsibility of companies in any given country can be reduced into statistical figures that make meaning for developmental planning.

Chapter Two

SOME BASIC STATISTICAL TERMS AND OPERATIONS

2.1 CONCEPT

The term concept can be defined as an abstraction that closely resembles the item it represents. When a concept is mentioned, it is easy to visualize and conceptualize the object being represented or referred to. For example, if one mentions the concept 'Tree', you can easily think of how a tree looks like and you can visualize it by thinking of branches, leaves and other things that make a tree. Other examples of terms that are concepts include motor, clock, book, house etc.

2.2 CONSTRUCT

A construct is a higher level abstraction which can only be inferred from observable behaviour. Examples include terms like motivation, intelligence, love, honesty, attitude, justice, interest, personality, etc. These are things that cannot easily be observed as we observe, tree, clock and other concepts.

2.3 VARIABLE

A variable is any concept or construct that can take on more than one numerical value when measured, observed or manipulated. An example of a variable is height. It varies from one person or object to another. Similarly, sex, age, attitude, intelligence, motivation are all variables. When a variable is considered in terms of statistical computations, we have two kinds of variables – continuous variable and discrete variable. Continuous variables are those variables which can take all possible values (both integral and fractional) in a given specified range. Such a variable is capable of passing from any given value to the next value by infinitely small gradations. An example of a continuous variable is age of students in a school since it can be measured to the nearest fraction of time, years, month, days, minutes, and even seconds. Age can be measured in a certain range, say from 5 years to 10years, 1 year 6 months to 20years 6months as the case may be. Other examples of continuous variables are wages of employees, height, weight, distances, scores emanating from a performance test, etc. On the other hand, any variable which cannot take all the possible values within a given specified range are termed as discrete variables. Examples of discrete variable include family size, sex, number of houses, number of birds in a poultry farm etc.

When variables are considered in terms of research, there can be three types. These are independent variables, dependent variables and confounding variables. This text would not handle these differences.

2.4 BASIC ARITHEMATIC OPERATIONS IN STATISTICS

There are four basic arithmetic operations performed in statistics. The operations are addition, subtraction, multiplication, and division.

2.4.1 Addition

The arithmetic sign for addition is the plus sign (+). The addition of two numbers like x and y is written as x + y. If x = 20 and y = 10, then x + y = 20 + 10 and the result is 30, meaning that 20 + 10 = 30

There are two basic rules that help a beginner in adding directed numbers in statistics.

i. When adding two numbers having the same sign together, add the absolute values and allow your result retain the sign.
   (-3) + (-5) = -8
   (+3) + (+5) = +8

ii. When adding two numbers, one having a negative sign and the other a positive sign, find their difference, and retain the sign of the bigger number.
   (-3) + (+5) = +2 can be written as -3 + 5 = +2, and
   (+3) + (-5) = -2 can be written as 3 – 5 = -2

2.4.2 Subtraction

The arithmetic sign for subtraction is minus (-). The subtraction of one number say y from another say x is written as x - y.

If x = 20 and y = 10

Then x – y = 20 – 10, and the result is 10 meaning that 20-10 = 10

The same rules for addition apply for subtraction except that the sign of the number being subtracted should change.

(-3) – (-5) = -3 + 5 =2

(-3) – (+5) = -3-5 = -8

2.4.3 Multiplication

When two numbers have the same sign (- or +) their product is positive. When two numbers have different signs (- & +), their product is normally negative.

(-3) (-5) = +15

(3) (5) = +15

(-3) (5) = -15

(3) (-5) = -15

2.4.4 Division

The division of two numbers that have the same sign gives a positive result, while the division of two numbers that have different signs gives a negative answer

20/5 = 4

-20/-5 = 4

-20/5 = -4

20/-5 = -4

2.5 THE SIGMA NOTATION

The sigma notation, also referred to as the summation operator is an important arithmetic operator used regularly in statistics. It is important in the simplification of statistical formulas. The sigma notation is written in Greek as $\sum$, meaning the sum of. Whenever we sum up some statistical figures we use the notation to explain that the aggregate is the given figure.

Suppose a set of data was made up of the scores:

$X_1 = 5, X_2 = 1, X_3 = 8, X_4 = 9, X_5 = 1, X_6 = 3$

Then,

$$\sum_{i=1}^{6} X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$$

$$= \quad 5 + 1 + 8 + 9 + 1 + 3$$

$$= \quad 27$$

The meaning of $\sum_{i=1}^{6} X_i$ is that the summation begins with the first number and ends with the 6th number. In general, the notation is given as:

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

This means that one should begin the summation from the first number and end with the nth term.

If the nth term is unknown then:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

Consequently:

$$\sum_{i=2}^{n} X_i = X_2 + X_3 + X_4 + \cdots + X_n, \text{ and}$$

$$\sum_{i=3}^{6} X_i = X_3 + X_4 + X_5 + X_6$$

If $X_1 = 5, X_2 = 1, X_3 = 8, X_4 = 9, X_5 = 1, X_6 = 3$

Then, $\sum_{i=3}^{5} X_i = 8 + 9 + 1$

= 18

2.6 BASIC RULES GOVERNING THE SIGMA NOTATION

2.6.1 Rule 1:

Multiplying a constant by each number in a set of data is the same as multiplying the constant with the sum of all the numbers in the set of data.

i.e. $\sum_{i=1}^{n} CX_i = C\sum_{i=1}^{n} X_i$

Given $X_1 = 3$, $X_2 = 8$, $X_3 = 5$, $X_4 = 1$ and the constant (C) is 10

Then, $\sum_{i=1}^{4} CX_i = 10(3) + 10(8) + 10(5) + 10(1)$

=30+80+50+10

= 170

Or

$C\sum_{i=1}^{4} X_i = 10(3 + 8 + 5 + 1)$

$= 10 \times 17$

= 170

Again suppose the population of a country in its six geo-political zones is given as:

Zone	1	2	3	4	5	6
Population	10000	12000	8000	5000	9000	11000

What will be the estimated population of the country in 10 years' time if it is estimated that the population triples every 10 years.

Let each zone be X

Then $X_1 = 10000$, $X_2 = 12000$, $X_3 = 8000$, $X_4 = 5000$, $X_5 = 9000$,

$X_6 = 11000$ and $C = 3$

i.e

$$\sum_{i=1}^{6} CX_i = 3(10000) + 3(12000) + 3(8000) + 3(5000) + 3(9000) + 3(11000)$$

= 30000+36000+24000+15000+27000+33000

= 165,000

Or

$$C\sum_{i=1}^{6} X_i = 3(10000+12000+8000+5000+9000+11000)$$

$$3\times 55000$$

= 165,000

2.6.2 Rule 2

When a series of constant scores is summed, the result is the same as multiplying n by the constant score.

$$\sum_{i=1}^{n} C = nC$$

Given that $X_1 = 10$, $X_2 = 10$, $X_3 = 10$, $X_4 = 10$, then

$$\sum_{i=1}^{4} X_i = 10+10+10+10 = 40$$

Or

$$NC = 4 \times 10 = 40$$

2.6.3 Rule 3

When we sum two or more scores for a single individual and then sum their sums for a number of individuals in a sample, it is the same thing as summing the two or more scores separately for the individuals in the sample then summing the scores.

$$\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

Let's consider the scores of eight students in say Physics (X) and Chemistry (Y) as follows:

Students	X	Y
1	6	8
2	5	5
3	3	4
4	2	3
5	6	5
6	7	6
7	1	2
8	4	5

Then,

$$\sum_{i=1}^{8}(X_i + Y_i) = (6+8)+(5+5)+(3+4)+(2+3)+(6+5)+(7+6)+(1+2)+(4+5)$$

14+10+7+5+11+13+3+9

= 72

Or

$$\sum_{i=1}^{8} X_i + \sum_{i=1}^{8} Y_i = (6+5+3+2+6+7+1+4)+(8+5+4+3+5+6+2+5)$$

= 34 + 38

= 72

Many other operations can be carried out from the data in rule 3.

Example (i): $\sum_{i=1}^{2}(X_i + Y_i) = (6+8)+(5+5)$

= 14+10

= 24

Or

$$\sum_{i=1}^{2} X_i + \sum_{i=1}^{2} Y_i = (6+5)+(8+5)$$

= 11 + 13

= 24

(ii)

$$\sum_{i=2}^{8}(X_i + Y_i) = (5+5)+(3+4)+(2+3)+(6+5)+(7+6)+(1+2)+(4+5)$$

= 10+7+5+11+13+3+9

= 58

Or

$$\sum_{i=1}^{8} X_i + \sum_{i=1}^{8} Y_i = (5+3+2+6+7+1+4)+(5+4+3+5+6+2+5)$$

= 28 + 30

= 58

2.7 EXERCISE ONE

1. With practical examples differentiate between a concept and a construct as may be applied in statistics.
2. (a). Define the term "Variable"
   (b) Write short notes on the following pair of variables
   i. Continuous and discrete variables.
   ii. Independent and dependent variables
   iii. Extraneous variable
3. From the data below carryout the operations with the summation sign.

	Population in thousands		
Year	Obioko	Kaboli	Kapil
1	3	4	6
2	5	6	7
3	6	8	8
4	9	11	10
5	10	12	12

Assume that Obioko is X, Kaboli is Y and Kapil is Z, find

i. $\sum_{i=1}^{n} X_i$ ii. $\sum_{i=1}^{n} Y_i$ iii. $\sum_{i=1}^{5} Z_i$

iv. $\sum_{i=1}^{5}(X_i + Y_i)$ v. $\sum_{i=1}^{n}(X_i + Y_i + Z_i)$ vi.

$$\sum_{i=3}^{5} X_i + \sum_{i=3}^{5} Y_i$$

vii. $\sum_{i=4}^{4}(X_i + Y_i + Z_i)$

4. The constant in a distribution is 4 and the scores are

$X_1 = 6$, $X_2 = 1$, $X_3 = 4$, $X_4 = 5$, $X_5 = 2$, $X_6 = 6$ and $X_7 = 2$.

Estimate

i. $\sum_{i=1}^{7} CX_i$ ii $\sum_{i=1}^{7} X_i$ iii $\sum_{i=3}^{5} CX_i$

Chapter Three

FREQUENCY DISTRIBUTION

Frequency distribution can be defined as all computational effort made statistically by somebody in summarizing data before they are analyzed. A frequency distribution is a tallying of a number of times each score value or intervals of score values occur in a group of scores (Adegoke, 2014). There are two major methods of carrying out such summaries namely:

1. Frequency tables and
2. Frequency diagrams

3.1 FREQUENCY TABLES

A frequency table, according to Afonja (2001) is one in which the variable of interest forms the basis for classification and the entries are in frequencies. Simply put, it deals with classes of distribution and their frequencies.

Two types of frequency tables can be identified:

i. Frequency table without class interval (ungrouped frequency tables).
ii. Frequency table with class interval and boundaries (Grouped frequency tables).

Frequency table without class intervals is one in which the frequencies of individual values are presented, while frequency table with class intervals and boundaries is the one in which intervals instead of individual values are used as classes (See table 3.1 and 3.2). The basic components of a frequency table are individual scores (for ungrouped data) or classes (for grouped data), frequencies, cumulative frequencies, relative frequencies, and relative cumulative frequencies.

Relative frequencies are expressions of each frequency as a ratio of the total frequency. It can be interpreted to mean the percentage of the total number of subjects in a distribution who had a certain score. For example, the relative frequency of .075 in Table 3.1 means that 7.5% of the subjects in the distribution scored 11 points. Relative cumulative frequencies are expressions of each cumulative frequency as a ratio of the total frequency. The relative cumulative frequency for the highest score in a distribution is usually 1.0 since the cumulative frequency for that highest score is usually equal to the total number of subjects in the distribution (N).

Example 3.1

Consider a class of 40 students with the following scores in a test and prepare a frequency table for the class.

11	14	11	12	12	14	14	13	12	13
15	13	12	13	13	13	13	12	16	15
14	14	13	15	14	11	12	14	13	14
12	15	14	16	14	14	14	15	14	15

Table 3.1 Frequency table for ungrouped data

Scores X	Tally	Frequency	Relative	Cumulative Frequency	Relative Cumulative Frequency
16	II	2	2/40 = .050	40	40/40 = 1.000
15	~~IIII~~ 1	6	6/40 = .150	38	38/40 = .950
14	~~IIII IIII~~ III	13	13/40 = .325	32	32/40 = .800
13	~~IIII~~ IIII	9	9/40 = .225	19	19/40 = .475
12	~~IIII~~ II	7	7/40 = .175	10	10/40 = .250
11	III	3	3/40 = .075	3	3/40 = .075
		N = 40			

Example 3.2

Consider the following scores obtained by 26 first degree students of a certain university and use them to prepare a grouped frequency table

61 60 40 50 51 92 75 82 85 65 86 90 70
71 60 47 55 64 70 79 45 85 32 50 65 60

Use a class interval of 10.

An important thing to note is that you are required to use a class interval of 10. We can know the number of classes required by finding the range and dividing it by 10.

Range = the highest score minus the lowest score
= 92 – 32 = 60

Number of classes will be = $\frac{60}{10}$ = 6+1 = 7

The actual number of classes will be 6 + 1 = 7. If the number of classes is known, we can get the class internal thus:

$$\frac{\text{Range}}{\text{Actual number of class minus 1}}$$

i.e. $\frac{60}{7-1} = \frac{60}{6} = 10$

Another important thing to note is where to begin the lower class limit of the first class. The best principle is to start from any number divisible by the class interval size. If the lowest score is not divisible by the class interval size then start the lowest class from the next lower number to the lowest score that is divisible by the class interval size. In example 3.2, the lowest score is 32, but the next lower score divisible by 10 is 30. So the first class will be 30-39.

Table 3.2: Frequency table for grouped data

Class interval (i)	Class mid point (x)	Tally	Frequency	Relative frequency	Cumulative Frequency	Relative Cumulative Frequency
90-99	94.5	II	2	.077	26	1.000
80-89	84.5	IIII	4	.154	24	.923
70-79	74.5	~~IIII~~	5	.192	20	.769
60-69	64.5	~~IIII~~ II	7	.269	15	.577
50-59	54.5	IIII	4	.154	8	.308
40-49	44.5	III	3	.115	4	.154
30-39	34.5	I	1	.038	1	.038
			N = 26			

3.1.1 Class limits and class boundaries

Class limits are the extreme scores of class intervals. Consequently, 30 and 39, 40 and 49, etc are class limits. Accordingly, 30 and 40 are lower class limits while 39 and 49 are upper class limits. Class boundaries can be defined as real class limits that connect the extreme score of one class with the extreme score of the next class. The class boundaries for say 40-49 are 39.5

and 49.5. Thus 39.5 is the lower class boundary for 40-49 while 49.5 is its upper class boundary.

3.2 FREQUENCY DIAGRAMS

Frequency diagrams are important means of presenting data such that outstanding features can readily be seen. Diagrams in statistics do not prove anything as they are no substitutes for tests, as may be applied to the data. Diagrams can only be useful in suggesting such test and in explaining the conclusions found about the test.

There are two categories of diagrams in distribution summaries; these are:

i) Ordinary diagrams comprising line diagrams, bar diagrams (bar charts), pie charts, pictograms, scatter diagrams and maps.

ii) Frequency diagrams comprising histograms, polygons and cumulative frequency curves (Ogive).

3.2.1 Line diagrams

Line diagrams are graphs that show relationships between two or more variables.

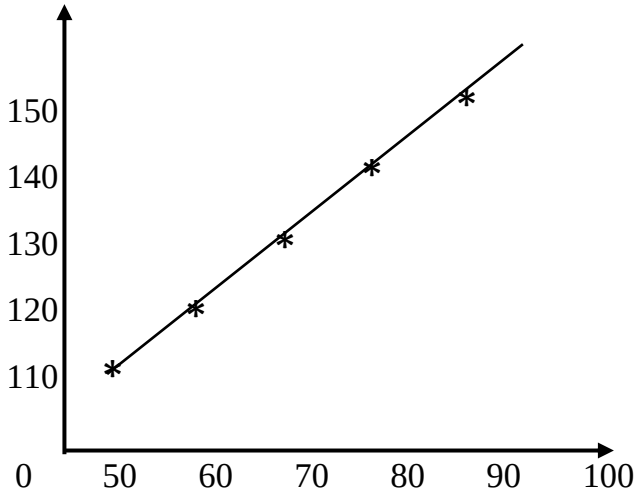


Figure 3.1: A hypothetical relationship between intelligence and achievement

3.2.2 Bar chart

A bar chart is a diagram made up of rectangles, which have the same width but with lengths proportional to the frequencies of the items in the data. The rectangles (bars) are usually drawn separately and at equal distances apart. Figure 3.2 is the bar chart of table 3.1.

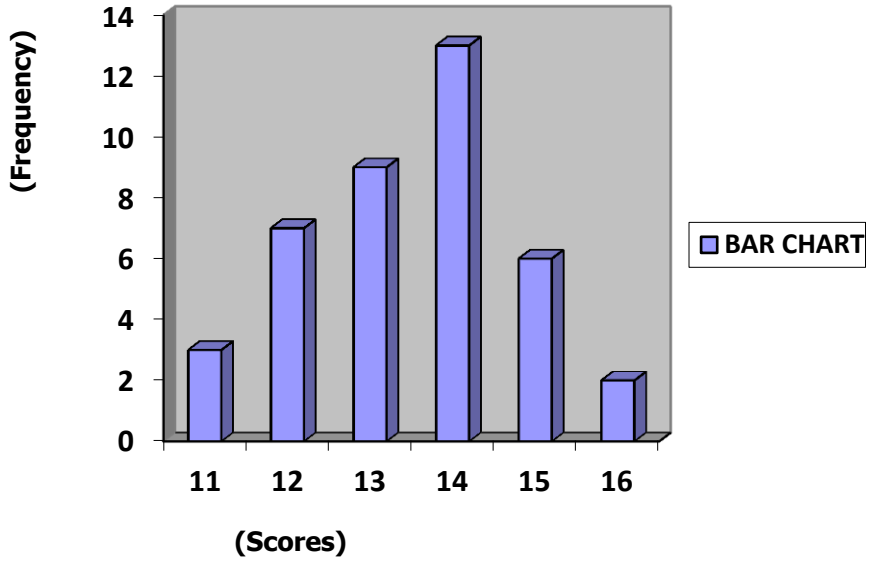


Figure 3.2: Bar chart of the summaries in Table 3.1

3.2.3 Pie chart

A pie chart is a circular diagram partitioned in sectors such that the sectors are proportional to the frequencies of the items. Let's consider a community of 60 in which 25 are members of PDP, 20 are members of APC, 10 are members of labour party, and 5 are members of APGA. To prepare a pie chart for this summary, we first convert the frequencies into degrees as in table 3.3.

Table 3.3: Conversion of frequencies into degrees

S/No	Party affiliation	Frequency	Degrees
1	PDP	25	150
2	APC	20	120
3	Labour	10	60
4	APGA	5	30
	Total	**60**	**360**

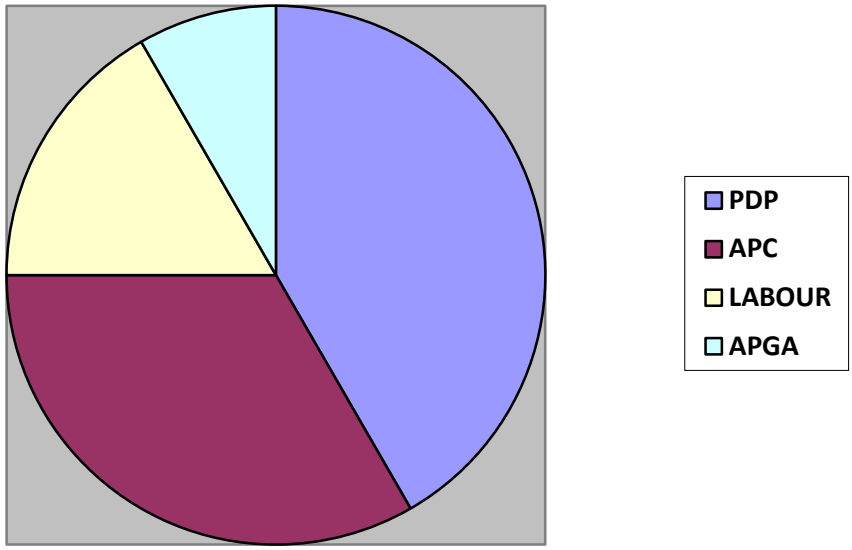


Fig. 3.3: Pie chart of data in table 3.3

3.2.4 Pictogram

Adequate pictures or symbols of equal sizes can be used in presenting statistical data. Such representations are called pictograms. The data in Table 3.3 can be represented as a pictogram.

PDP:
APC:

LABOUR:
APGA:

Fig. 3.4: Pictogram of data in Table 3.3

3.2.5 Histogram

A histogram is a chart that shows the information of a frequency table. To draw a histogram the values of the variable are scaled along the X-axis and the frequencies along the y-axis. A histogram is most appropriate when drawn from a grouped data. Figure 3.5 presents the histogram of table 3.2.

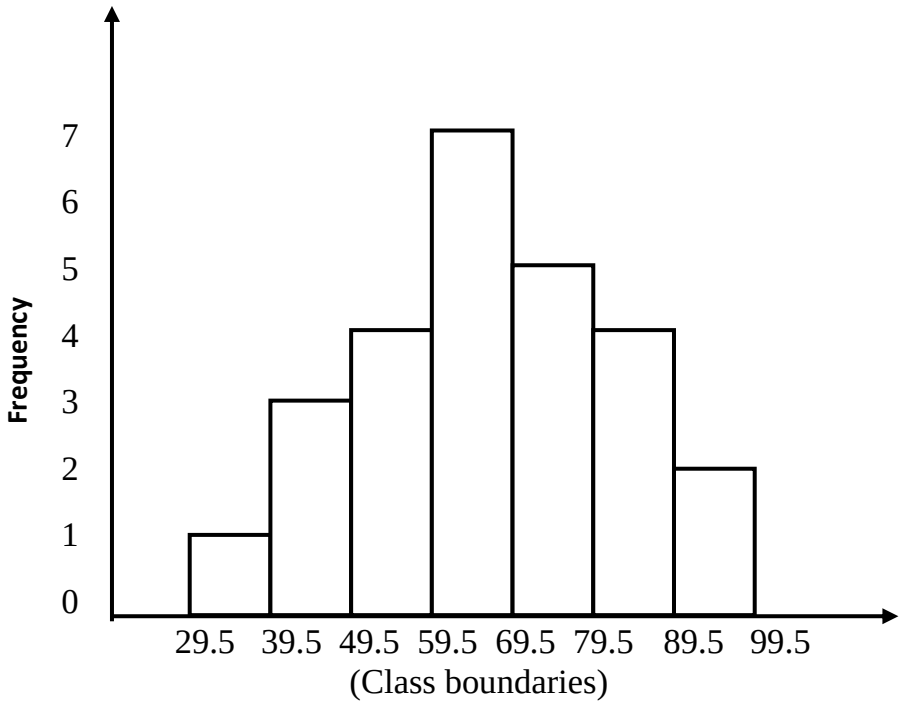


Figure 3.5: Histogram of data in Table 3.2

3.2.6 Frequency Polygon

Frequency polygon is another important method of presenting statistical summaries in diagram form. It is defined as the graph of a frequency table. The graph is normally plotted from the

midpoints of the class intervals. Figure 3.6 is the frequency polygon of Table 3.2.

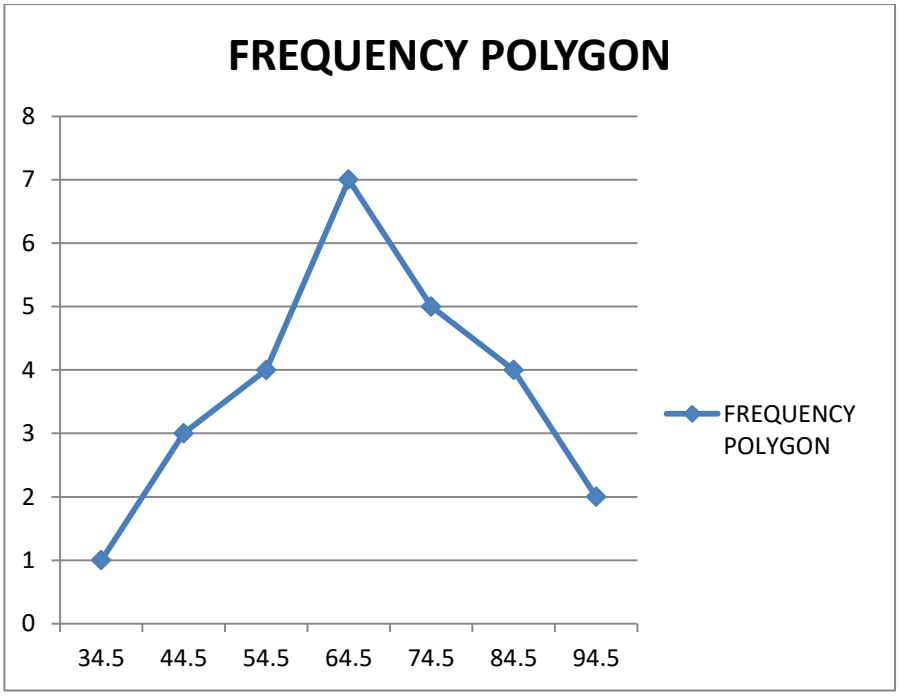


Fig. 3.6: Frequency polygon of data summaries in table 3.2

3.2.7 Cumulative frequency

This is the graph of cumulative frequencies in distribution summaries. Another name for cumulative frequency curve is Ogive. An Ogive can be 'less-than Ogive' or 'more-than Ogive'. When the graph slopes upward from left to right (as in Figure 3.7), the curve is less-than Ogive. When the graph slopes downward from left to right (as in Figure 3.8), the curve is more-than Ogive. The same data set can give either a 'less than' or 'more than' Ogive depending on how the frequencies are cumulated. If the frequencies are cumulated in ascending order, from the lowest

class or score to the highest class or score, the resulting curve will produce a less-than Ogive (as in figure 3.7). Conversely, if the frequencies are cumulated in descending order, from the highest class or score to the lowest class or score, the resulting curve will produce a more-than Ogive (as in Figure 3.8).

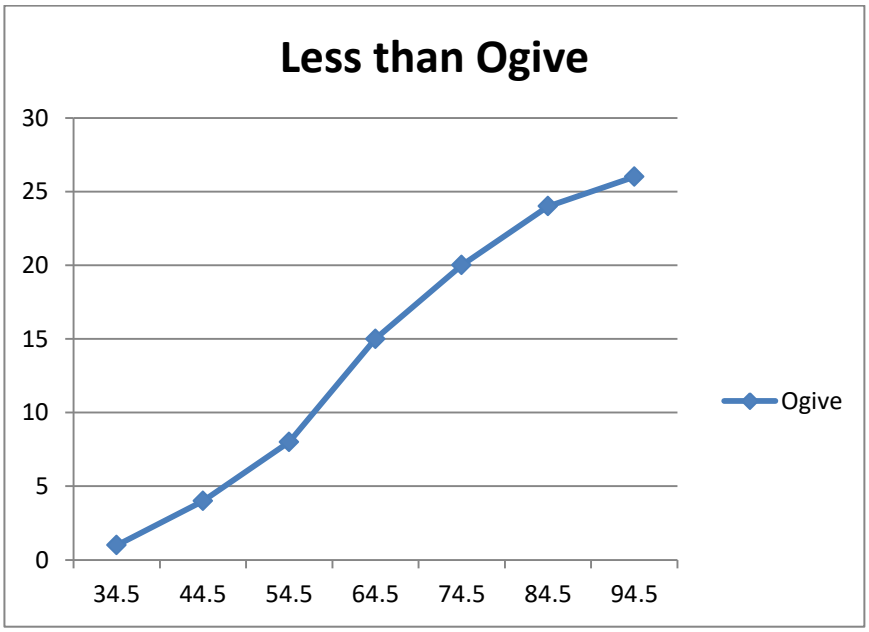


Fig. 3.7: Less-than Ogive of data summaries in Table 3.2

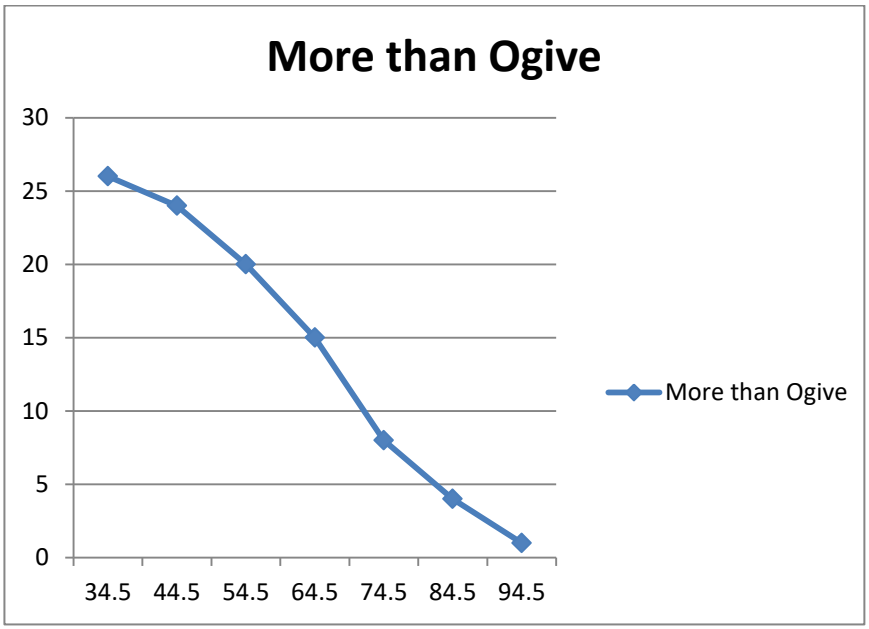


Fig. 3.8: More-than Ogive of data summaries in Table 3.2

3.3 WORKED EXAMPLE ONE

Consider the following data sets for 40 students in physics(x) and Chemistry (y)

Student	Sex	Parents Educ.	X	Y
1	M	N	45	48
2	M	P	53	60
3	F	T	48	52
4	M	S	60	66
5	F	T	58	60
6	F	S	55	57
7	F	N	50	56
8	M	P	43	54
9	M	P	15	43
10	M	T	83	90

11	M	S	52	65
12	F	T	50	59
13	F	P	32	47
14	F	N	28	41
15	M	T	22	37
16	F	N	57	82
17	F	P	70	77
18	F	T	57	65
19	M	T	48	66
20	F	T	44	63
21	M	P	71	88
22	F	S	47	63
23	F	T	50	59
24	M	P	56	64
25	M	S	68	72
26	M	S	64	70
27	M	T	55	63
28	F	S	42	57
29	F	T	38	41
30	F	T	24	48
31	F	S	55	61
32	M	T	63	74
33	F	S	47	62
34	M	S	50	60
35	F	T	55	60
36	M	S	18	25
37	M	T	32	47
38	F	T	50	60
39	M	T	56	65
40	M	S	50	72

Key:
Sex: M = Male
F = Female

Parents Education: N = No Education
P = Primary Education
S = Secondary Education
T = Tertiary Education

i. Draw Bar and Pie Charts for Parents' Education.
ii. Prepare a frequency table for performance in each subject.
iii. Draw histograms for student's performance in Physics and their performance in Chemistry.
iv. Draw the polygon for each distribution.
v. Draw the cumulative frequency curves for each subject.

i. Table 3.4 Frequency table for Parents' Education

Parents' Education	**Frequency**	**Degrees**
No education	4	36
Primary Education	7	63
Secondary Education	12	108
Tertiary Education	17	153
Total	**40**	**360**

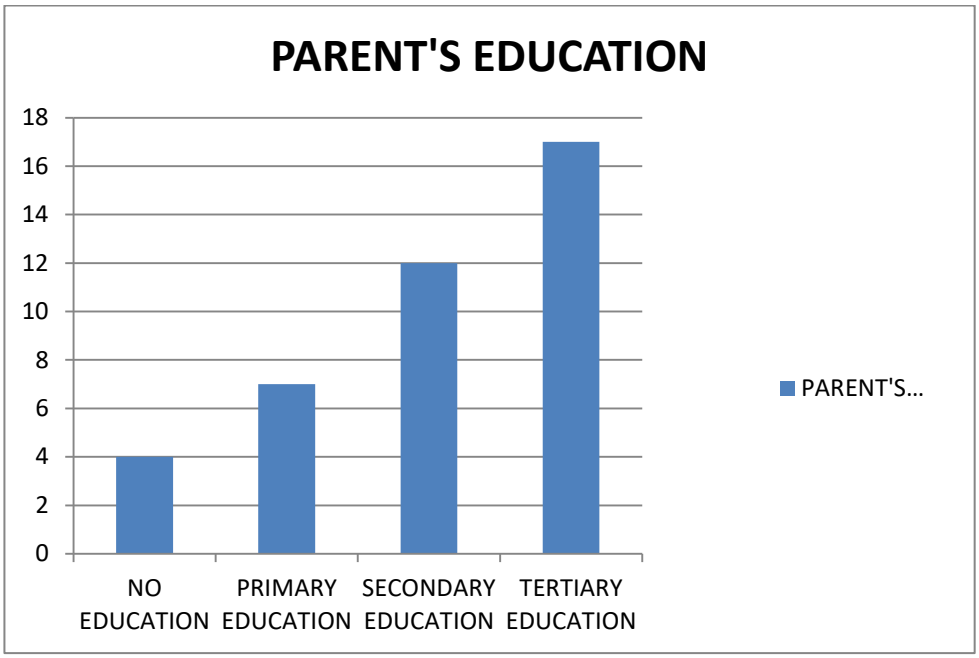


Fig. 3.9 Bar chart for data summaries in Table 3.4

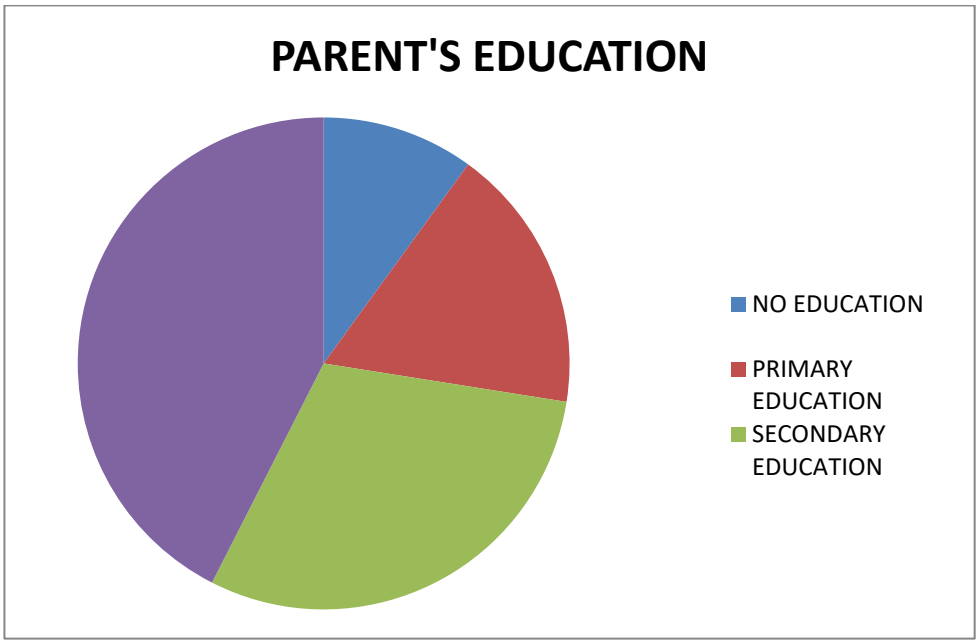


Fig. 3.10: Pie Chart for data summaries in Table 3.4

ii) Table 3.5 Frequency Table for data in worked example 3.3 (physics)

Class Interval	Class Midpoints	Tally	f	rf	cf	rcf
80-89	84.5	I	1	.03	40	1.00
70-79	74.5	II	2	.05	39	.98
60-69	64.5	IIII	4	.10	37	.93
50-59	54.5	~~IIII~~ ~~IIII~~ ~~IIII~~ II	17	.43	33	.83
40-49	44.5	~~IIII~~ III	8	.20	16	.40
30-39	34.5	III	3	.08	8	.20
20-29	24.5	III	3	.08	5	.13
10-19	14.5	II	2	.05	2	.05

Table 3.6: Frequency table for data in worked example 3.3 (Chemistry)

Class Interval	Class Midpoints	Tally	f	rf	cf	rcf
90-99	94.5	I	1	.03	40	1.00
80-89	84.5	II	2	.05	39	.98
70-79	74.5		5	.13	37	.93
60-69	64.5	~~IIII~~ ~~IIII~~ ~~IIII~~ I	16	.40	32	.80
50-59	54.5	~~IIII~~ II	7	.18	16	.40
40-49	44.5	~~IIII~~ II	7	.18	9	.23
30-39	34.5	I	1	.03	2	.05
20-29	24.5	I	1	.03	1	.03

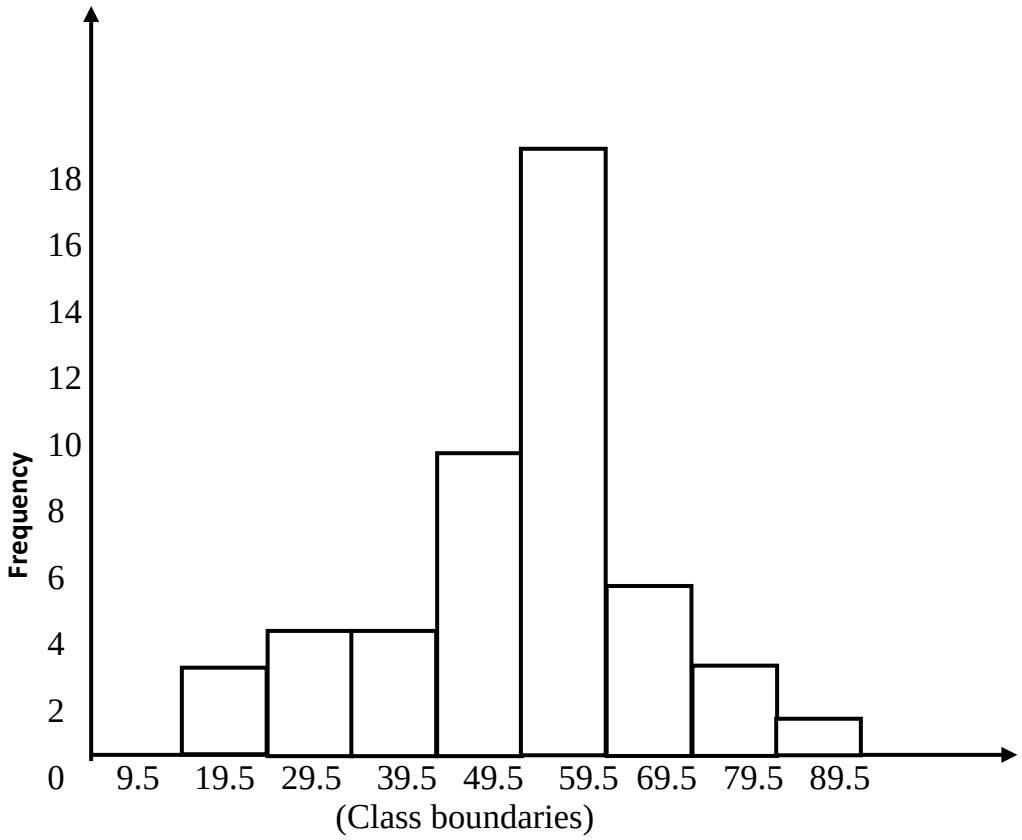


Fig. 3.11 Histogram of data in worked example 3.3 (Physics)

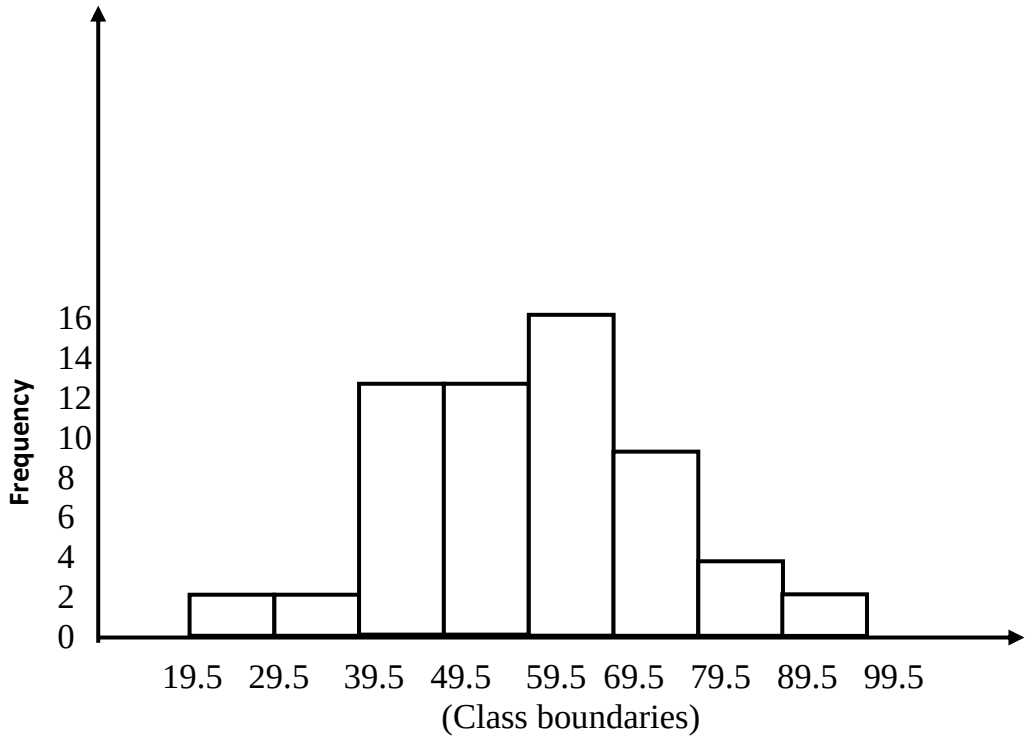


Fig. 3.12: Histogram of data in worked example 3.3 (Chemistry)

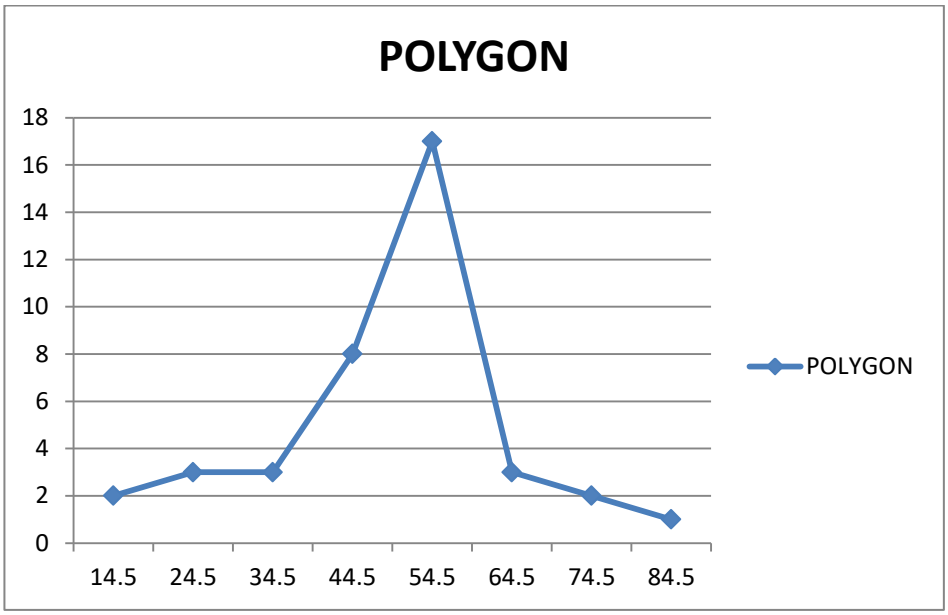


Fig. 3.13: Polygon of data on worked example 3.3 (Physics)

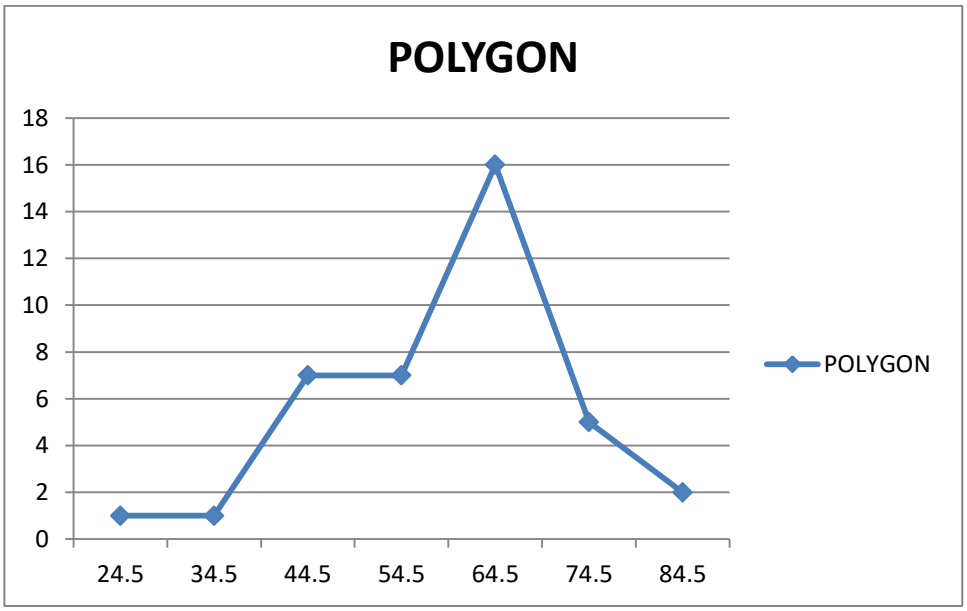


Fig. 3.14: Polygon of data on worked example 3.3 (Chemistry)

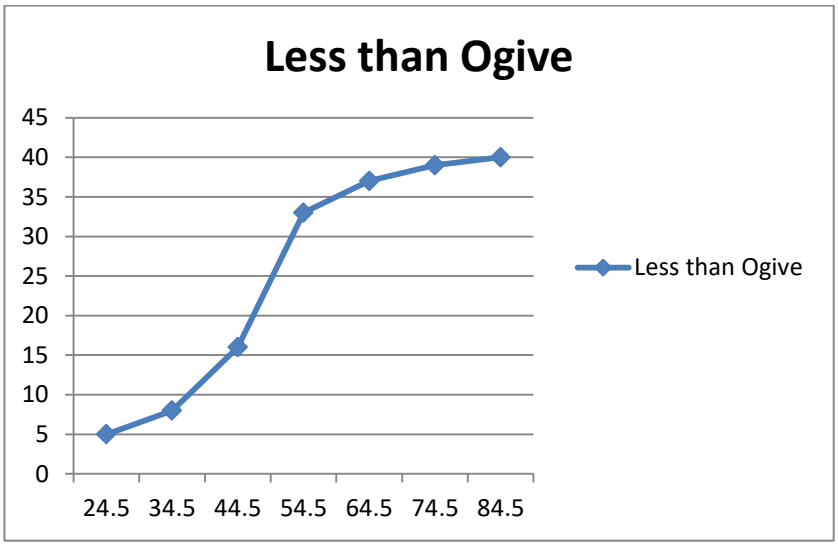


Fig. 3.15: Less than Ogive of data on worked example 3.3 (Physics)

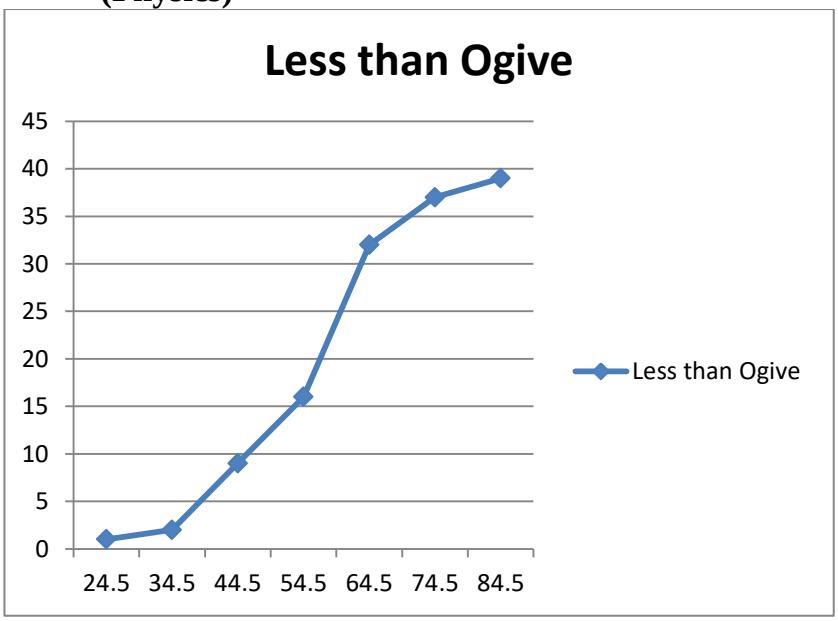


Fig. 3.16: Less-than Ogive of data on worked example 3.3 (Chemistry)

3.4 EXERCISE THREE

1a. Mention two different methods by which one can present data summaries.

b. Differentiate between grouped and ungrouped data.

2. Write short notes on the following:

i. Relative frequency and relative cumulative frequency.

ii Frequency and cumulative frequency

iii Class interval and class midpoint

iv Less-than Ogive and more-than Ogive.

3. Consider the following scores obtained by 40 students in a test and use the scores to prepare a frequency table, a histogram, a frequency polygon and cumulative frequency curve.

4	96	6	3	6	3	2	4	5
7	64	5	5	5	4	5	6	5
5	86	6	4	5	3	6	2	6
5	61	7	8	10	6	5	8	2

Chapter Four

THE MEASURES OF CENTRAL TENDENCY

The measures of central tendency are those single values that attempt to describe a set of data by identifying the central position within that set of data. They can also be described as measures that serve to locate the centre at which most of the data are concentrated or clustered. Based on this definition, the measures of central tendency are sometimes referred to as measures of central location, averages or representative values.

Basically, there are three measures of central tendency; the mean, the median and the mode.

4.1 THE MEAN

The mean is the most frequently used measure of central tendency (Adegoke, 2014). It is the arithmetic average of the scores in a distribution. It can also be defined mathematically as the sum of the values in a dataset divided by the number of observations.

The mean is given as:

1. $\overline{X} = \frac{\sum X}{N}$; where the scores in the data set have not been distributed in frequencies.

2. $\bar{X} = \frac{\sum fx}{N}$ when the scores in the data set have been distributed in frequencies.

Where $\bar{X}$ = Mean when sample is involved. It is μ when population rather than sample is involved.

X = Observations (scores) for ungrouped data, class midpoints for grouped data

N = Number of observations (total of frequencies)

Σ = The sigma notation referring to the summation of.

EXAMPLE 4.1

Let's consider the ages of six workers in the marketing department of a firm given as follows:

23 21 33 30 26 23

then, $\bar{X} = \frac{\sum X}{N}$

$$= \frac{23+21+33+30+26+23}{6}$$

$$= \frac{156}{6}$$

$$= 26$$

EXAMPLE 4.2

Let's assume the data summaries in Table 3.1 and estimate the mean

Table 4.1

Score (X)	f	Fx
16	2	32
15	6	90
14	13	182
13	9	117
12	7	84
11	3	33
Total	40	538

$$\bar{X} = \frac{\sum fx}{N} = \frac{538}{40} = 13.45$$

EXAMPLE 4.3

Let's also consider the data summaries in Table 3.2 and estimate the mean

TABLE 4.2

Class interval (i)	Class Mid-point	f	fx
90-99	94.5	2	189
80-89	84.5	4	338
70-79	74.5	5	372.5
60-69	64.5	7	451.5
50-59	54.5	4	218
40-49	44.5	3	133.5
30-39	34.5	1	34.5
Total		26	1737

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

$$= \frac{1737}{26}$$

$$= 66.8$$

4.2 THE MEDIAN

The median is the most centrally located score in a distribution when the scores are arranged in ascending or descending order of magnitude. Another way to define the median is that, it is the score in a distribution that divides the distribution into two parts such that an equal number of scores fall below and above the particular point of division. When the scores in a distribution are few, the median can be easily identified at the centre after arranging the scores in order of magnitude or through the cumulative frequency. When the scores in the distribution are summarized in class intervals, the median is obtained using a formula. The median formula is given as:

$$\text{Median} = L_1 + \frac{\left(\frac{1}{2}N - Cf_1\right)C}{f}$$

Where: L_1 = The real lower limit (class boundary) of the class interval containing the median

N = Number of scores in the distribution

Cf_1 = Cumulative frequency of the class interval immediately before the class interval containing the median.

f = frequency of the class interval containing the median

C = class interval size.

EXAMPLE 4.4

Let's consider the ages of those six workers in example 4.1

23 21 33 30 26 23

Arranged in ascending order we have

21 23 23 26 30 33

$$\text{Median} = \frac{23+26}{2} \quad \frac{49}{2} = 24.5$$

EXAMPLE 4.5

Let's consider the distribution in Example 4.2 and estimate the median

Score (X)	f	cf
16	2	40
15	6	38
14	13	32
13	9	19
12	7	10
11	3	3
Total	40	

From the cumulative frequencies, midway the distribution is at 32, so the median is the score at the cumulative frequency of 32. That score is 14

i.e, Median = 14

EXAMPLE 4.5

From the data summaries in example 4.3, let's calculate the median

TABLE 4.4

Class interval (i)	Class Mid-point	f	Cf
90-99	94.5	2	26
80-89	84.5	4	24
70-79	74.5	5	20
60-69	64.5	7	15
50-59	54.5	4	8
40-49	44.5	3	4
30-39	34.5	1	1
Total		26	

$$\text{Median} = 59.5 + \frac{\left(\frac{26}{2} - 8\right)10}{7}$$

$$= 59.5 + \frac{(13-8)10}{7}$$

$$= 59.5 + \frac{5 \times 10}{7}$$

$$= 59.5 + 7.14$$

$$= 66.64$$

4.3 THE MODE

The mode is the most frequently occurring score in a distribution. It is the score which has the highest frequency.

The mode can be determined by inspection or by using the histogram or by using a formula. The formula for mode is given as:

$$\text{Mode} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$$

Where:

L = the lower class limit of the class containing the mode

Δ_1 = difference between the frequency of the class containing the mode and the frequency of the class immediately before it.

Δ_2 = difference between the frequency of the class containing the mode and the frequency of the class immediately after it.

C = class interval size.

EXAMPLE 4.6: Estimating Mode by inspection

From our data on the ages of six workers given as:

23 21 33 20 26 23

The mode is 23 because it is the score that occurred most in the distribution.

Also considering the distribution in example 4.5, the mode is simply 14 because 14 had the highest frequency.

EXAMPLE 4.7: Bimodal distribution

When a distribution has two observations occurring more frequently than other observations, the distribution is said to be bimodal. The simple interpretation is that the distribution has two modes.

Consider the following set of scores:

2 1 3 2 2 4 5 3 1 3

Two scores '2' and '3' occurred three times each so the two are modes of the distribution. This distribution is called a bimodal distribution.

EXAMPLE 4.8: Mode by Histogram

When the histogram is drawn, the point of intersection at the model bar of the diagonals from the succeeding and preceding bars gives the modal score.

Let's consider the histogram on Figure 3.5 and estimate the mode.

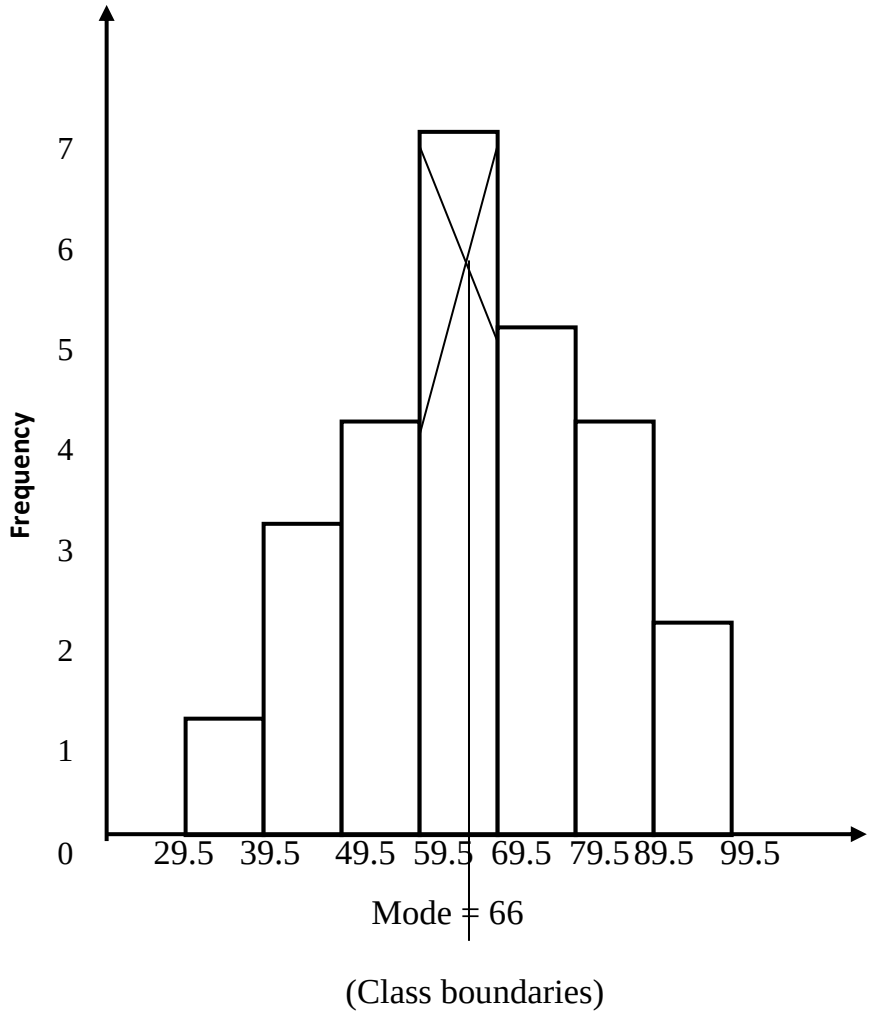


Fig. 4.1: Histogram showing the mode

The Mode = 66 because the imaginary line created on the tallest bar passes through score 66 on the horizontal axis.

EXAMPLE 4.9 Mode by formula

From summaries in Table 4.4 let's calculate the mode.

If L = 60

$\Delta_1 = 7 - 4 = 3$

$\Delta_2 = 7 - 5 = 2$

C = 10

Then, Mode = $60 + \left(\frac{3}{3+2}\right)10$

$= 60 + \frac{3}{5} \times 10$

$= 60 + 6$

= 66

4.4 SOME IMPORTANT ATTRIBUTES OF THE MEASURE OF CENTRAL TENDENCY

4.4.1 Advantage of the Mean

The major advantage of the mean is that it can be used for both continuous and discrete data.

4.4.2 Limitations of the Mean

(i) The mean cannot be calculated for categorical data since the values for categorical data like sex (male and female) cannot be summed.

(ii) The mean can easily be influenced by outliers and skewed distributions. This is because the mean includes all values of a distribution. An outlier is a score that is either too low or too high to fit into the normal pattern of score size in the distribution. Skewed distributions are explained later in this text.

4.4.3 Advantages of the Median

(i) It is the better measure of central tendency, when a distribution is not symmetrical.

(ii) The median is normally less affected by outliers and skewed data.

4.4.4 Limitation of the median

The median cannot be estimated or identified when categorical or nominal data is involved. This is because such data cannot be logically ordered in either ascending or descending order.

4.4.5 Advantage of the mode

The major advantage of the mode is its superiority over the mean and the median as it can be found for both numerical and categorical data.

4.4.6 Limitations of the mode

(i) In some distributions, the mode may not be found at the centre of the distribution.

(ii) There can be more than one mode in a distribution, thus making the mode a non-typical value of the distribution.

(iii) In some distributions especially those that are continuous, it is possible not to have a mode if all values are different.

4.5 SKEWNESS OF A DISTRIBUTION

The skewness of any distribution is an explanation of the nature of the data explained by the histogram. A symmetric distribution is one in which the two halves of the histogram appear as mirror-images of one another (see figure 4.2). In such a distribution the mean and the median are normally the same figure and the curve so formed is normal as in Figure 4.2.

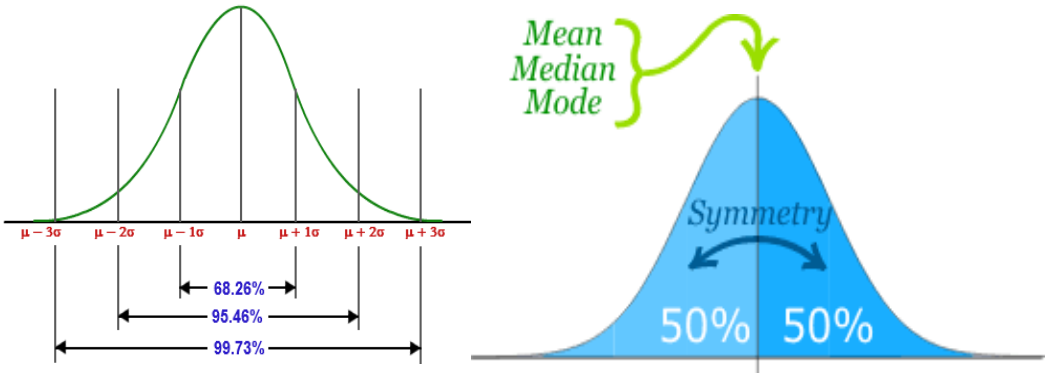


Fig. 4.2: Normal curve showing a symmetrical division of the distribution.

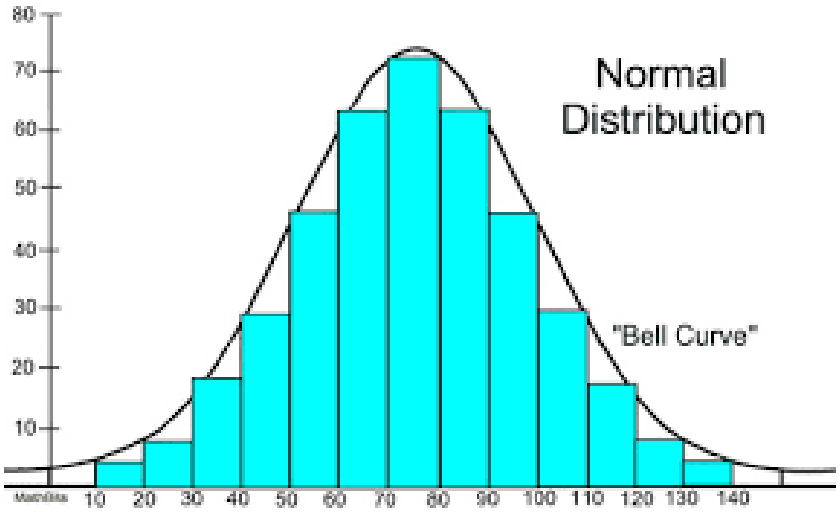


Fig. 4.3: A hypothetical example of an histogram whose distribution is normal with the mean, the median and the mode being equal.

When a distribution is skewed, the median remains the middle value, the mode the most-occurring value, but the mean is pulled in the direction of the tail. This means that the mean cannot be at the centre of the distribution once a distribution is skewed.

4.5.1 Negative and positive skewness

A distribution is said to be negatively skewed or skewed to the left when the tail on the left side of the distribution is longer than the right side. In this kind of distribution, the mean is pulled toward the left tail of the distribution and in most cases the median tends to be greater than the mean.

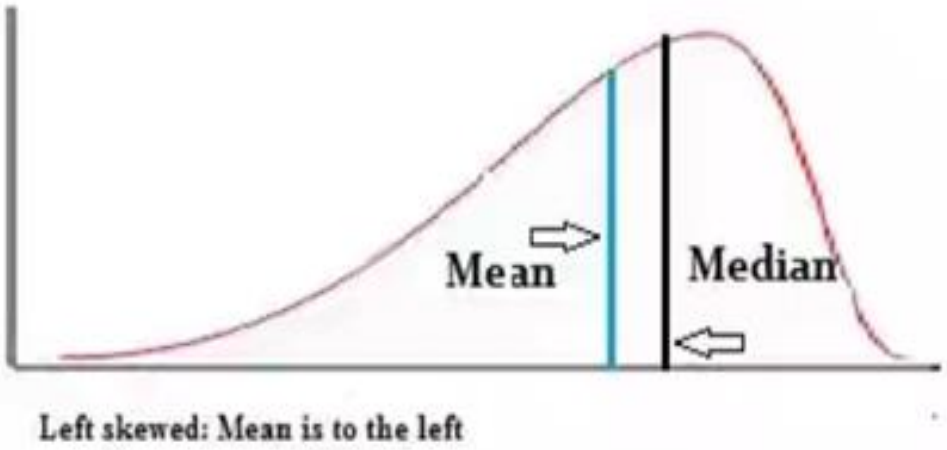


Fig. 4.4: Negatively skewed (left-skewed) distribution

A distribution is said to be positively skewed or skewed to the right when the tail on the right side of the distribution is longer than the left side. In this kind of distribution, the mean is pulled towards the right tail of the distribution and in most cases the mean is greater than the median and most other values.

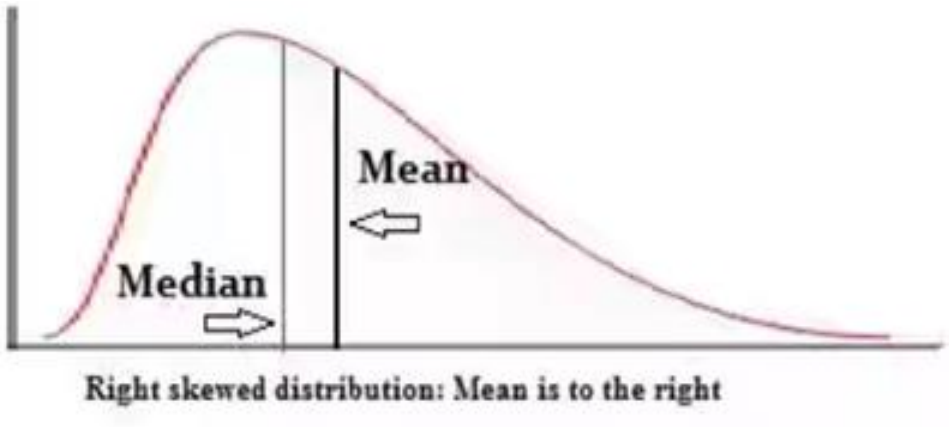


Table 4.5

Class interval (i)	Class Mid-point	F	Fx	Cf
80-89	84.5	1	84.5	40
70-79	74.5	2	149.0	39
60-69	64.5	4	258.0	37
50-59	54.5	17	926.5	33
40-49	44.5	8	356.0	16
30-39	34.5	3	103.5	8
20-29	24.5	3	73.5	5
10-19	14.5	2	29.0	2
Total		40	1980	

(i) $$\bar{X} = \frac{\sum fx}{N} = \frac{1980}{40} = 49.5$$

(ii) $$\text{Median} = L_1 + \frac{\left(\frac{1}{2}N - Cf_1\right)c}{f}$$

$$= 49.5 + \frac{\left(\frac{40}{2} - 16\right)10}{17}$$

$$= 49.5 + \frac{(20-16)10}{17}$$

$$= 49.5 + \frac{4}{17} \times 10$$

$$= 49.5 + 2.353$$

= 51.85

(iii) Mode = $L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c$

$= 50 + \left(\frac{9}{9+13} \right) 10$

$= 50 + \left(\frac{9}{22} \right) \times 10$

$= 50 + 4.091$

= 54.091

Table 4.6

Class interval (i)	**Class Mid-point**	**F**	**fx**	**Cf**
90-99	94.5	1	94.5	40
80-89	84.5	2	169	39
70-79	74.5	5	372.5	37
60-69	64.5	16	1032	32
50-59	54.5	7	381.5	16
40-49	44.5	7	311.5	9
30-39	34.5	1	34.5	2
20-29	24.5	1	24.5	1
Total		40	2420	

(i) $\bar{X} = \frac{\sum fx}{N} = \frac{2420}{40} = 60.5$

(ii) $\text{Median} = L_1 + \dfrac{\left(\frac{1}{2}N - Cf_1\right)c}{f}$

$= 59.5 + \dfrac{\left(\frac{40}{2} - 16\right)10}{16}$

$= 59.5 + \dfrac{(20-16)10}{16}$

$= 59.5 + \dfrac{4}{16} \times 10$

$= 59.5 + 2.5$

$= 62.0$

(iii) $\text{Mode} = L + \left(\dfrac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$

$= 60 + \left(\dfrac{9}{9+11}\right)10$

$= 60 + \dfrac{9}{20} \times 10$

$= 60 + 4.5$

$= 64.5$

4.6 EXERCISE FOUR

1. Mention the three measures of central tendency and discuss the computational method of each using a practical example.
2. Explain the functions and limitations of each of the measures of central tendency.
3. Exhaustively differentiate between negatively and positively skewed distributions using appropriate diagrams.
4. The following were scores obtained by 20 students. Find the mean, the median and the mode of the distribution.

Score	1	2	3	4	5	6	7	8	9	10
No of students	1	3	3	5	10	11	2	3	1	1

5. From the following data summaries, find the mean, the median and the mode.

Class interval	10-14	15-19	20-24	25-29	30-34	35-39	40-44
frequency	2	5	12	15	11	6	4

Chapter Five

MEASURES OF VARIABILITY OR DISPERSION

5.1 Definition

Different authors have defined the concept of variability in different ways. Gupta (2013) summarized various such definitions and said the term variability is used to indicate the fact that within a given group, the scores differ from one another in size, or in other words, there is lack of uniformity in their sizes.

This definition highlights a universal tendency which is that at all times there will normally be variations in the sizes of scores in a distribution. This makes it unjustifiable for the measures of central tendency to speak completely well of any given normal distribution. That is there should be some other measures that will describe extent of scatter of data in a set. Those measures are the measures of variability most often called the measures of dispersion.

Measures of variability, according to Joshua (2008) are those indices that express quantitatively the extent to which scores in a given distribution cluster together or scatter, or put simply the

extent to which such scores vary from one another. The measures of variability explain what the measures of central tendency (mean, median and mode) cannot explain.

Let's examine the scores obtained by two students in eight subjects, each set totaling 60 and having a mean of 7.5.

Student	English	Maths	Biology	Chemitry	Physics	Econs	CRK phy	History	Total	Mean
Fulljack	3	2	1	12	13	15	10	4	60	7.5
Special	8	7	9	9	8	7	5	7	60	7.5

The two students obtained the same total scores from eight subjects and the scores have the same mean of 7.5. Consequently, if we are told that the mean for the scores obtained by a student is 7.5, we cannot know if the student is Fulljack or Special. By this explanation, it is therefore possible to have several other cases where the same mean could be obtained from eight scores. This shows how inadequate the measures of central tendency are in completely describing a distribution. This showcases the importance of supporting the measures of central tendency with the measures of variability.

The variability in the scores obtained by Fulljack can be described as relatively more dispersed while those obtained by Special are slightly dispersed. The meaning is that scores obtained by Fulljack are heterogeneous while those obtained by Special are homogenous.

5.2 TYPES OF MEASURES OF VARIABILITY

There are five measures of variability considered in this text. These are:

i. The Range
ii. Quartile deviation or semi inter-quartile range
iii. Mean deviation
iv. The variance
v. The standard deviation

5.3 RANGE

The range is simply the difference between the highest and the lowest scores in a distribution. The range is given as.

$$R = X_h - X_l$$

Where: R = the range

X_h = the highest value in a distribution

X_l = the lowest value in a distribution

EXAMPLE 5.1

Find the range in our hypothetical example where scores were given as:

11	14	11	12	12	14	14	13	12	13
15	13	12	13	13	13	13	12	16	15
14	14	13	15	14	11	12	14	13	14
12	15	14	16	14	14	14	15	14	15

X_h = 16

X_l = 11

Range = 16 – 11

= 5

5.3.1 Advantages of the Range as a Measure of Variability

The range is the simplest measure of variability to compute. Its definition is quite rigid (never having any other way to define it) and readily comprehensible.

5.3.2 Disadvantages of the Range as a Measure of Variability

Some of the limitations of the range are as follows:

i. The range is based on only the two extreme scores of a distribution. This means that other scores of the distribution are not involved in the estimations of the range. This is a limiting factor that makes the range not very reliable.

ii. The range will always change when extreme scores in a sample change. This means that, it is possible for the range to change each time a sample is either reduced or increased.

iii. The range does not change except the smallest and/or the largest score change. No matter the fluctuation in the scores, as far as the smallest or the largest does not change, the range is unaffected. This is limiting for a measure of variability.

iv. Most often the range increases as sample size increases and reduces as sample size reduces.

5.4 QUARTILE RANGE AND INTER QUARTILE RANGE

Quartiles are the values that divide a list of numbers into quarters. When we put the list of scores in ascending order, then cut the list into four equal parts, the quartiles are at the points of cuts. Each

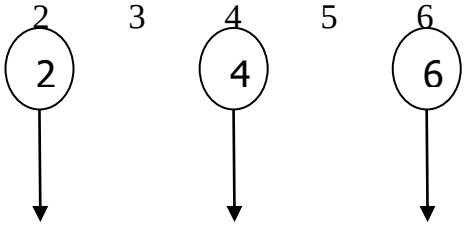
quartile is denoted by the letter 'Q'. Based on this, we have on any set of data Q_1, Q_2, Q_3.

Where: Q_1 = lower quartile
Q_2 = middle quartile (the median)
Q_3 = upper Quartile

EXAMPLE 5.2

Consider the score: 2 1 4 3 8 5 6

1 2 3 4 5 6 8



Q_1 Lower Quartile

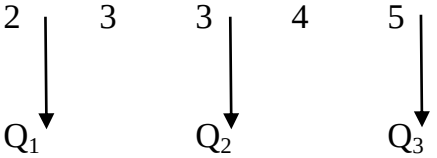
Q_2 Middle Quartile (Median)

Q_3 Upper Quartile

If the scores were rather: 2 1 4 3 3 8 5 6

The quartile are as demonstrated here

1 2 3 3 4 5 6 8



Q_1 Q_2 (Median) Q_3

In the first consideration, the scores are seven, so Q_2 is exactly score 4, but in the second consideration the scores are eight, so

$$Q2 = \frac{3+4}{2} = 3.5$$

The idea about quartile is that the scores are divided into four such that each quartile has 25% of the scores.
The inter-quartile range is the difference between the third quartile (Q_3) and the first quartile (Q_1) given as:

$Q_3 - Q_1$.

This implies that the inter-quartile range is the difference between 75% of the scores and 25% of the scores. That is, the inter-quartile range is only 50% of the distribution.

Semi inter-quartile range also called quartile deviation is obtained by dividing the interquartile range by 2.

i.e, Semi inter-quartile Range = $\frac{Q_3 - Q_1}{2}$

From the example with 1 2 3 3 4 5 6 8

Inter-quartile Range = 5.5 – 2.5 = 3.0

and, semi inter-quartile Range $\frac{5.5 - 2.5}{2}$ = 1.5

5.4.1 ADVANTAGES OF THE QUARTILE DEVIATION AS A MEASURE OF VARIABILITY

i. It is quite easy to understand and calculate.
ii. It is not based on only two extreme scores of the distribution as it is the case with the range. It makes use of 50% of the scores in a distribution.

iii. It is not affected by the size of the extreme scores in the distribution. That is, even if the scores change, the quartile deviation is not affected.

iv. It is the only measure of variability that can be obtained with a distribution having open end classes.

5.4.2 DISADVANTAGES OF THE QUARTILE DEVIATION

i. It ignores 50% of the observations in a distribution. That is 25% at the lower and 25% at the upper end of the distribution.

ii. If the sample size changes, the quartile deviation also changes.

iii. Apart from just to show the differences between the third quartile (Q_3) and the first quartile (Q_1), the quartile deviation has no further mathematical implications.

5.5 VARIANCE AND STANDARD DEVIATION

These two measures are hardly treated separately. The variance leads to the estimation of the standard deviation and the two are mathematically related to each other. The standard deviation is mathematically the square root of the variance, while the variance is the square of the standard deviation.

The calculation and notation of the variance and standard deviation depends on whether it is the entire population or a sample of the population that is considered. The notations are given as:

σ for population standard deviation

σ^2 for population variance

S for standard deviation based on the sample

S^2 for variance based on the sample.

The formula for estimating variance is given as:

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

For population variance, and

$$S^2 = \frac{\sum(X-\bar{X})^2}{n-1} \text{ or } \frac{n\sum X^2 - \left(\sum X\right)^2}{n^2-1}$$ for variance based on the sample.

For the measures of central tendency, the mean that is obtained from the sample of a population is not a biased administration of the population mean. This is not usually true for the sample variance if it is calculated in the same manner as the population variance. If some samples of 'n' members are used to calculate sample variance for each combination using 'n' as denominator, the average of the results, would not be equal to the true value of the population variance. The result would be biased. This bias can be corrected by using 'n-1' as the denominator.

The formula for estimating standard deviation is given as:

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$ for population standard deviation, and

$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$ or $\sqrt{\frac{n\sum X^2 - \left(\sum X\right)^2}{n^2 - 1}}$ for sample standard deviation.

When the measures are in frequencies, the formulae would be given as:

$$\sigma^2 = \frac{\sum f(X - \mu)^2}{N}$$

$$S^2 = \frac{\sum f(X - \bar{X})^2}{n-1} \text{ or } \frac{n\sum fX^2 - \left(\sum fX\right)^2}{n^2 - 1}$$

$$\sigma = \sqrt{\frac{\sum f(X - \mu)^2}{N}}$$

$$S = \sqrt{\frac{\sum f(X - \bar{X})^2}{n-1}} \text{ or } \sqrt{\frac{n\sum fX^2 - \left(\sum fX\right)^2}{n^2 - 1}}$$

EXAMPLE 5.3

Consider the observation in Example 4.1 with the mean of 26

To estimate the variance, we will first deviate the scores from the mean.

Thus:

$$S = \sqrt{\frac{(23-26)^2 + (21-26)^2 + (33-26)^2 + (30-26)^2 + (26-26)^2 + (23-26)^2}{6-1}}$$

$$= \sqrt{\frac{9+25+49+16+0+9}{5}}$$

$$= \sqrt{\frac{108}{5}}$$

$$= \sqrt{21.6}$$

$$= 4.6475800$$

$$\approx 4.65$$

Note that the variance in this computation is 21.6 since standard deviation is the square root of the variance and variance is the square of the standard deviation.

i.e. $4.65^2 \approx 21.6$.

EXAMPLE 5.4

Let's us also consider the data in Table 5.1 and compute the standard deviation.

Table 5.1

X	f	fx	x^2	fx^2
16	2	32	256	512
15	6	90	225	1350
14	13	182	196	2548
13	9	117	169	1521
12	7	84	144	1008
11	3	33	121	363
Total	40	538		7302

$$S = \sqrt{\frac{40 \times 7302 - 538^2}{40^2 - 1}}$$

$$= \sqrt{\frac{292080 - 289444}{1600 - 1}}$$

$$= \sqrt{\frac{2636}{1599}}$$

$$= \sqrt{1.6485}$$

$$= 1.283939 \cong 1.3$$

Alternatively the definitional formula can be used based on 13.45 as mean

Table 5.2

X	f	$(X-\bar{X})$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
16	2	2.55	6.5025	13.0050
15	6	1.55	2.4025	14.4150
14	13	0.55	0.3025	3.9325
13	9	-0.45	0.2025	1.8225
12	7	-1.45	2.1025	14.7175
11	3	-2.45	6.0025	18.0075
Total	40			65.9

$$S = \sqrt{\frac{65.9}{40-1}}$$

$$= \sqrt{\frac{65.9}{39}}$$

$$= \sqrt{1.6897}$$

$$= 1.299884$$

$$\approx 1.3$$

EXAMPLE 5.5

Consider the distribution below and estimate

(i) The standard deviation

(ii) The inter-quartile range

(iii) The semi inter-quartile range

5	8	6	4	5	3	6	2	6	6
5	6	7	8	10	6	5	8	2	1
4	9	6	3	6	3	2	4	5	6
7	6	5	5	5	4	5	6	5	4

(i) Standard deviation: computational method

Table 5.3

X	f	fx	x^2	fx^2
10	1	10	100	100
9	1	9	81	81
8	3	24	64	192
7	2	14	49	98
6	11	66	36	396
5	10	50	25	250
4	5	20	16	80
3	3	9	9	27
2	3	6	4	12
1	1	1	1	1
Total	40	209	385	1237

$$S = \sqrt{\frac{n\sum fX^2 - \left(\sum fX\right)^2}{n^2 - 1}}$$

$$S = \sqrt{\frac{40 \times 1237 - 209^2}{40^2 - 1}}$$

$$= \sqrt{\frac{49480 - 43681}{1600 - 1}}$$

$$= \sqrt{\frac{5799}{1599}}$$

$$= \sqrt{3.6266}$$

$$= 1.90436 \cong 1.9$$

Standard deviation: definitional method

First determine the mean

$$\bar{X} = \frac{209}{40}$$

$$= 5.23$$

Then deviate the scores from the mean

Table 5.4

X	F	$(X - \bar{X})$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
10	1	4.77	22.7527	22.7529
9	1	3.77	14.2129	14.2129
8	3	2.77	7.6729	23.0187
7	2	1.77	3.1329	6.2658
6	11	0.77	0.5929	6.5219
5	10	-0.23	0.0529	0.5290
4	5	-1.23	1.5129	7.5645
3	3	-2.23	4.9729	14.9187
2	3	-3.23	10.4329	31.2987
1	1	-4.23	17.8929	17.8929
Total				144.9760

$$S = \sqrt{\frac{\sum f(X - \bar{X})^2}{N - 1}}$$

$$= \sqrt{\frac{144.9760}{40 - 1}}$$

$$= \sqrt{3.717}$$

$$= 1.928$$

$$\cong 1.9$$

EXAMPLE 5.6

Find the inter-quartile and semi inter-quartile ranges for the following data.

11	14	11	12	12	14	14	13	12	13
15	13	12	13	13	13	13	12	16	15
14	14	13	15	14	11	12	14	13	14
12	15	14	16	14	14	14	15	14	15

To estimate the quartile range, we must find the cumulative frequencies as shown in Table 5.5

Table 5.5

X	*f*	*Cf*
16	2	40
15	6	38
14	13	32
13	9	19
12	7	10
11	3	3
Total	40	

Inter-quartile Range = 14 – 12.5 = 1.5

Semi inter-quartile Range = $\frac{14-12.5}{2} = \frac{1.5}{2} = 0.75$

Let's attempt an explanation.

25% has exactly 10 cases ending on it, so Q_1 is 12.5. 75% of the 40 scores is 30 and 30 cases end at 14 and more cases up to 32 are still at 14, so Q_3 is exactly 14.

5.6 PERCENTILES

Percentile is a figure that describes the rank of any given score in a distribution. It is used to describe the percentage of scores that fall at or below a given score in a distribution. Percentile system is widely used in educational measurement to report the standing of an individual relative to the performance of known group (Joshua, 2005).

Suppose a student scored 25 points in a forty-item test, and by ranking she is 5^{th} in that test. This information as being 5^{th} in rank cannot tell his parents what percentage of the students in his class has performed below 25 points. It is only the percentile rank that makes his parents visualize more easily their child's attainment in relation to the rest of the pupils in his class.

The calculation of percentile point and percentile rank is the same with the median. The formulas are given as:

$$P_x = L_1 + \left(\frac{Np - Cf_1}{f} \right) C \text{ for percentile point}$$

Where: P_x = Percentile point can be given as P_{75} for 75^{th} percentile.
L_1 = Lower boundary of the score or class interval containing the percentile point.
N = number of scores in the distribution
P = proportion of the desired percentile (for 75^{th} percentile, P = .75)
cf_1 = cumulative frequency of the score or class interval before the percentile point score or class interval.
f = frequency of the percentile point score or class interval.
c = class size; c = 1 for ungrouped data.
Or

$$\Pr = \left\{ \frac{f(X - L) + cf_1}{Nc} \right\} \bullet 100 \qquad \text{for percentile rank.}$$

Where: Pr = Percentile Rank

F = frequency of the score or class interval containing the score for which percentile rank is required.

X = The score for which the percentile rank is required.

L = lower boundary of the score or class interval of the score for which percentile rank is required.

Cf_1 = cumulative frequency of the score or class interval before the score or class interval containing the score for which percentile rank is required.

C = class size: C = 1 for ungrouped data

N = Number of scores in the distribution.

EXAMPLE 5.7 (UNGROUPED DATA)

Find the percentile point for the 70^{th} percentile in the distribution in Table 5.6 that follows:

Table 5.6

X	*f*	*Cf*
16	2	40
15	6	38
14	13	32
13	9	19
12	7	10
11	3	3
Total	40	

First, let's determine the number of scores below which 70 percent of the scores fall.

In which case $\frac{70}{100} \times 40 = 28$

This means that the score containing the percentile point is 14 since 28 falls at the cumulative frequency of 32.

Then, $P_{70} = 13.5 + \left[\frac{28-19}{13}\right]1$

$$= 13.5 + \left[\frac{28-19}{13}\right]1$$

$$= 13.5 + 0.6923$$

$$= 14.1923$$

$$\cong 14.19$$

We can find the percentile rank for score 13, for instance, as follows:

$$\Pr = \left[\frac{9(13-12.5)+10}{40 \times 1}\right]100$$

$$= \left[\frac{9(0.5)+10}{40}\right]100$$

$$= \left[\frac{4.5+10}{40}\right]100$$

$$= \frac{14.5}{40} \times 100$$

$$= 36.25$$

$$\cong 36$$

This means that score 13 is at the 36^{th} percentile.

That is, score 13 in that distribution can be ranked 36^{th} in terms of percentage.

EXAMPLE 5.8 (GROUPED DATA)

(i) Find the percentile point for the 65^{th} percentile for the distribution in Table 5.7.

(ii) Estimate the percentile rank for score 45 in the distribution.

Table 5.7

Class interval	Class mid-point	f	Cf
80-89	84.5	1	40
70-79	74.5	2	39
60-69	64.5	4	37
50-59	54.5	17	33
40-49	45.5	8	16
30-39	34.5	3	8
20-29	24.5	3	5
10-19	14.5	2	2
Total		40	

(i) 65 percent of the scores fall at cumulative frequency 26 (i.e,

$0.65 \times 40 = 26$)

Then $P_{65} = 49.5 + \left[\frac{(40 \times 0.65) - 16}{17}\right]10$

$= 49.5 + \left[\frac{26 - 16}{17}\right]10$

$= 49.5 + \frac{10}{17} \times 10$

$= 49.5 + 5.88$

$$= 55.38$$

(ii) $$\text{Pr} = \left[\frac{8(45 - 39.5) + 16}{40 \times 10}\right]100$$

$$= \left[\frac{8(5.5) + 16}{400}\right]100$$

$$= \left[\frac{44 + 16}{400}\right]100$$

$$= 15$$

Chapter Six

MEASURES OF RELATIONSHIP

6.1 INTRODUCTION

In all we have done from chapter three to five, we have based our discussion on univariate distributions. That is, distributions involving only are variable. We used such distributions in estimating the measures of central tendency, variability and local standing. There are other distributions that are bivariate. In such distributions the individuals of the sample or population take on a pair of scores each. An example is when we compare the scores obtained by members of a class on Mathematics and Physics. In such a distribution each member of the group will obtain a pair of scores that can be compared to obtain a relationship. This would not have been possible with the distributions we have so far examined. In a bivariate distribution, we may want to find if there is any relationship between the two variables under study. The correlation is a statistical tool with which the relationship between such two variables can be estimated. By this explanation, measures of relationship involve various techniques used for measuring the extent of the correlation between the two variables.

6.2 METHODS OF ESTIMATING CORRELATION

This book is confined to three frequently used methods of estimating correlation. These methods are:

(i) Scatter diagram method
(ii) Pearson's Product Moment Coefficient of Correlation method
(iii) Spearman rank order method.

6.2.1 Scatter diagram

Scatter diagram is a diagram of dots obtained from pairs of relationships of two variables possibly *x* and *y*. In education, for instance, there is bound to be a relationship between students' attitude towards school and their academic performance in school subjects. In such case, attitude towards school can be x and academic performance *y*. The relationship between *x* and *y* can be considered as x_1, y_1, x_2, y_2, x_3, y_3,…, x_n, y_n for the n students under consideration. A plot of these pairs of relationship on a graph forms a scatter diagram.

A scatter diagram may show one of the following relationships:

(i) Perfectly correlated (moving in perfect unison)
(ii) Partly correlated (some inter-relationship but not exact).
(iii) Uncorrelated (no relationship between their movements).

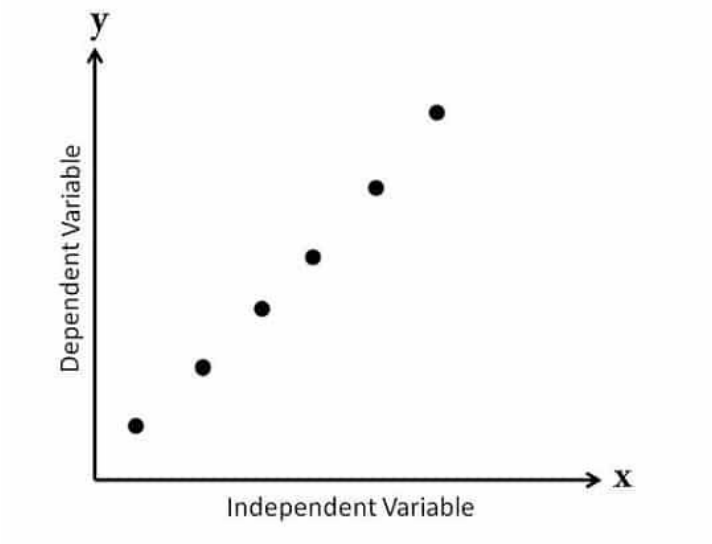


Figure 6.1: Perfect positive relationship

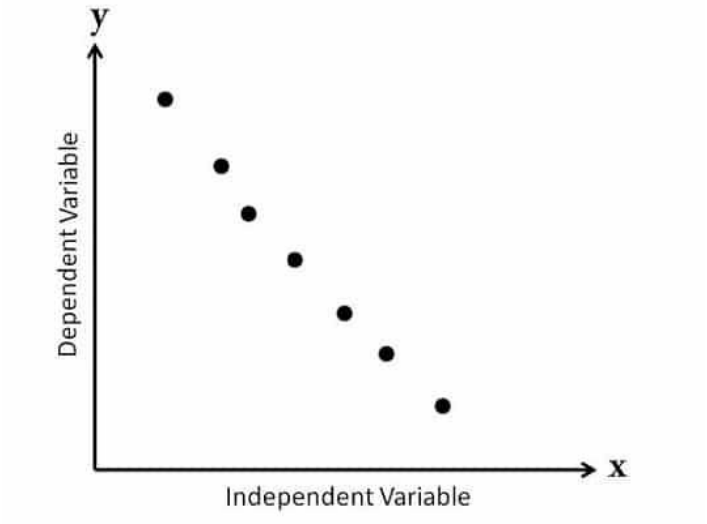


Figure 6.2: Perfect negative relationship

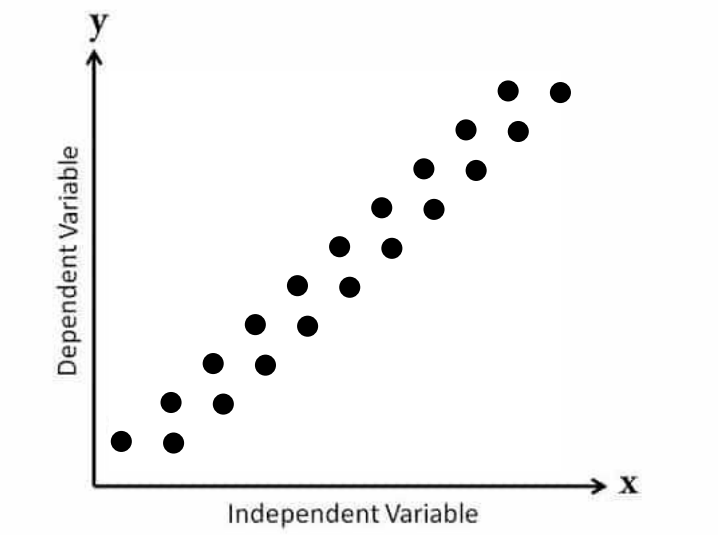


Figure 6.3: High degree of positive relationship

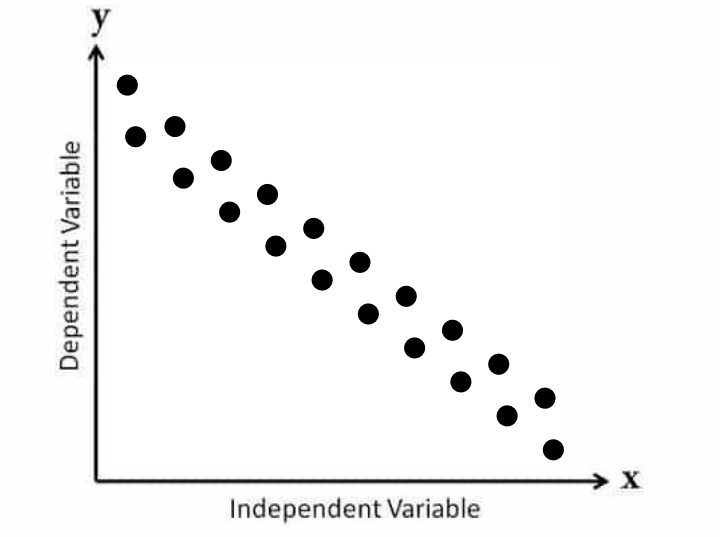
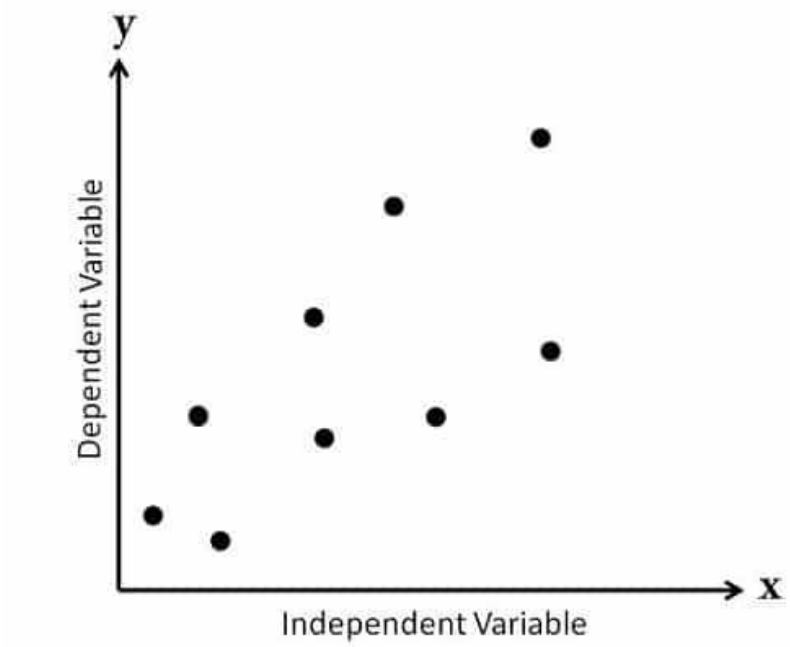


Figure 6.4: High degree of negative relationship


y
Dependent Variable
x
Independent Variable

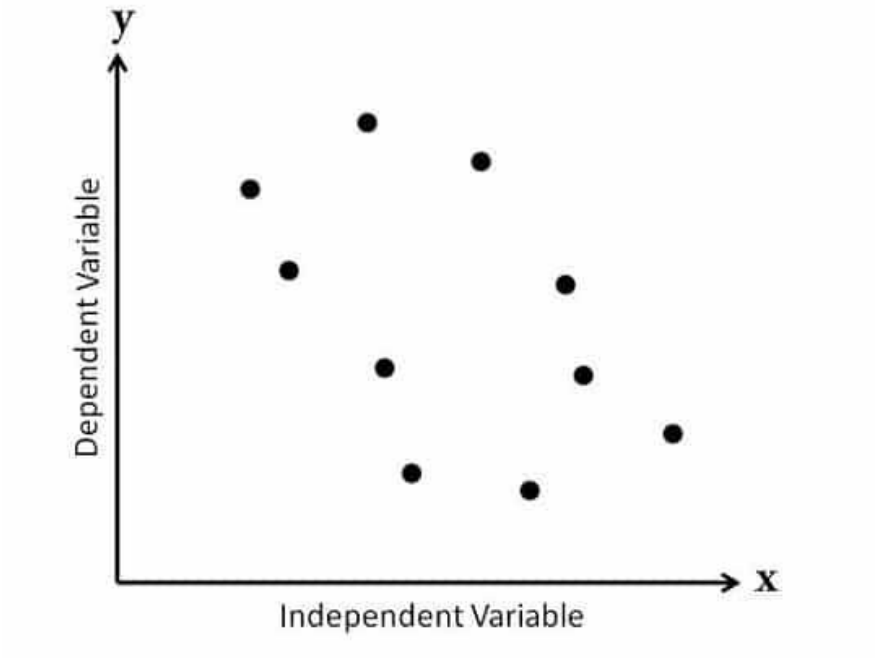

y
Dependent Variable
x
Independent Variable

Figure 6.6: Low degree of negative relationship

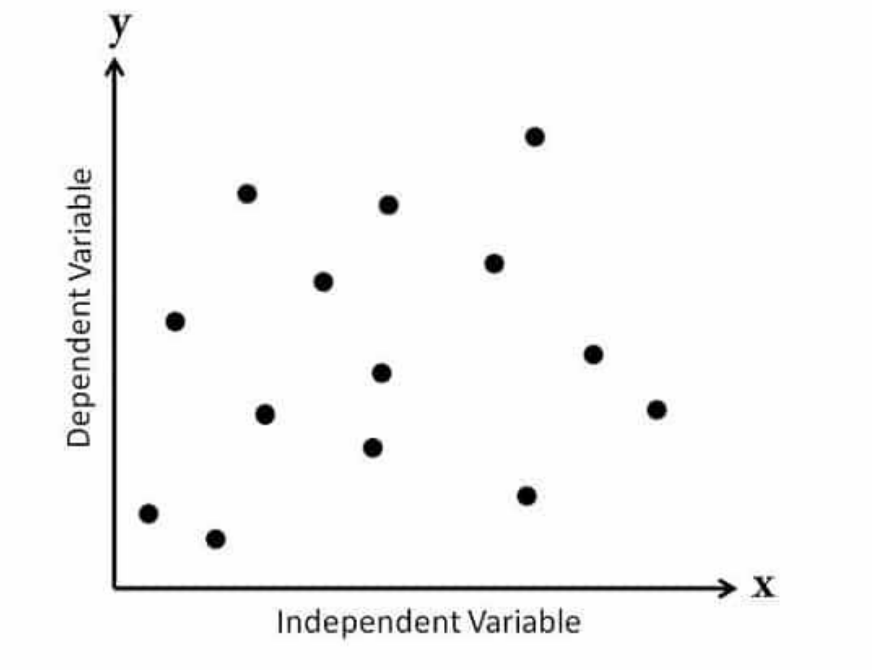


Figure 6.7: No particular relationship

6.4 PEARSON'S PRODUCT MOMENT COEFFICIENT OF CORRELATION

This is a mathematical method which provides a measure of the intensity or strength of association between two variables, one being the *x* variable, the other the *y* variable. This coefficient of correlation, widely used and denoted by r (small letter), was suggested by a renowned British Bio-Metrician and Statistician Karl Pearson; 1867-1936. The coefficient of correlation r can range from +1 (i.e perfect positive correlation where the values change in the same direction) to -1 (i.e. perfect negative correlation where y decreases linearly as x increases or vice versa) (Lucey, 2002).

The formula for Pearson's coefficient of correlation (r) is given as:

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

The formula can also be given as:

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

Where x and y are small letters of X and Y denoting the deviations of x and y from their arithmetic mean $\overline{X}$ and $\overline{Y}$ respectively.

EXAMPLE 6.1

Consider the scores on Table 6.1 obtained by ten students in Mathematics and Physics and use the two methods of Pearson Correlation to estimate the coefficient of association between performance in Mathematics and performance in Physics. What kind of relationship does your obtained r denote?

X	**Y**	**X^2**	**Y^2**	**XY**
60	58	3600	3364	3480
45	55	2025	3025	2475
72	63	5184	3969	4536
55	60	3025	3600	3300
45	43	2025	1849	1935
58	55	3364	3025	3190
41	35	1681	1225	1435
60	61	3600	3721	3660
48	45	2304	2025	2160

50	48	2500	2304	2400
534	523	29308	28107	28571

Using: $r = \dfrac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2} \bullet \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$

Then $r = \dfrac{10 \times 28571 - 534 \times 523}{\sqrt{10 \times 29308 - 534^2} \times \sqrt{10 \times 28107 - 523^2}}$

$= \dfrac{285710 - 279282}{\sqrt{293080 - 285156} \times \sqrt{281070 - 273529}}$

$= \dfrac{6428}{\sqrt{7924} \times \sqrt{7541}}$

$= \dfrac{6428}{89.02 \times 86.84}$

$= \dfrac{6428}{7730.4968}$

$= 0.8315$

$\cong 0.83$

Using $r = \dfrac{\sum xy}{\sqrt{\sum x^2} \bullet \sqrt{\sum y^2}}$ entails preparing another table to deviate $\bar{X}$ from X and $\bar{Y}$ from Y. $\bar{X} = 53.4$, $\bar{Y} = 52.3$.

Table 6.2

X	Y	x	Y	x^2	y^2	xy
60	58	6.6	5.7	43.56	32.49	37.62
45	55	-8.4	2.7	70.56	7.29	-22.68
72	63	18.6	10.7	345.96	114.49	199.02
55	60	1.6	7.7	2.56	59.29	12.32
45	43	-8.4	-9.3	70.56	86.49	78.12
58	55	4.6	2.7	21.16	7.29	12.42
41	35	-12.4	-17.3	153.76	299.29	214.52
60	61	6.6	8.7	43.56	75.69	57.42
48	45	-5.4	-7.3	29.16	53.29	39.42
50	48	-3.4	-4.3	11.56	18.49	14.62
				792.4	**754.1**	**642.8**

Then, $r = \dfrac{642.8}{\sqrt{792.4} \times \sqrt{754.1}}$

$$= \frac{642.8}{28.15 \times 27.46}$$

$$= \frac{642.8}{772.999}$$

$$= 0.8315$$

$$\cong 0.83$$

Note that using either of the formulae gives the same or approximately the same coefficient. The correlation coefficient of 0.83 for 10 subjects can be regarded as High Degree of positive relationship. It is however important to note that when a sample is low, high correlation coefficients may not mean significant relationship.

6.5 SPEARMAN RANK ORDER

The spearman Rank Order Correlation is also a Mathematical method which provides a measure of the association between two sets of ranked or ordered data. It is most useful when data involves qualitative attributes that can only be qualified through ranks. Attributes like punctuality, honesty, beauty, morality are good examples. This method of estimating correlation was developed by a British Psychologist, Charles Edward Spearman in 1904. The formula is given as;

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

Where: rho = Spearman Rank order

d = difference between the pair of ranks of same individual.

n = number of individuals in the distribution.

EXAMPLE 6.2

Consider the scores in Table 6.2 and compute the coefficient of correlation using Spearman rho method.

Table 6.3

X	Y	Rank x	Rank y	d	d^2
60	58	2.5	4	-1.5	2.25
45	55	8.5	5.5	3	9
72	63	1	1	0	0
55	60	5	3	2	4
45	43	8.5	9	-0.5	0.25
58	55	4	5.5	-1.5	2.25
41	35	10	10	0	0
60	61	2.5	2	0.5	0.25
48	45	7	8	-1	1
50	48	6	7	-1	1
					20

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

Then $rho = 1 - \frac{6 \times 20}{10(10^2-1)}$

$= 1 - \frac{120}{10 \times 99}$

$= 1 - \frac{120}{990}$

$= 1 - 0.1212$

$= 0.879$

$\cong 0.88$

Notice that when we use Karl Pearson's method the coefficient is 0.83 and when we use Spearman rho method the coefficient is 0.89. Both coefficients are above 0.80 despite the different methods applied, and both show high degree of positive correlation.

EXAMPLE 6.3

Use the data in Table 6.4 to:

(a) Calculate the Pearson's Correlation coefficient using raw score method.

(b) Calculate the Pearson's correlation coefficient using deviation method.

(c) Calculate the Spearman rank order correlation coefficient

Table 6.4

Student	1	2	3	4	5	6	7	8
Scores in Maths	6	8	5	4	7	3	5	6
Scores in History	4	2	5	6	3	7	3	3

Table 6.5

X	Y	X^2	Y^2	XY
6	4	36	16	24
8	2	64	4	16
5	5	25	25	25
4	6	16	36	24
7	3	49	9	21
3	7	9	49	21
5	3	25	9	15
6	3	36	9	18
44	33	260	157	164

Since: $r = \dfrac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2} \bullet \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$

$$r = \frac{8 \times 164 - 44 \times 33}{\sqrt{8 \times 260 - 44^2} \times \sqrt{8 \times 157 - 33^2}}$$

$$r = \frac{1312 - 1452}{\sqrt{2080 - 1936} \times \sqrt{1256 - 1089}}$$

$$= \frac{-140}{\sqrt{144} \times \sqrt{167}}$$

$$= \frac{-140}{12 \times 12.92}$$

$$= \frac{-140}{155.04}$$

$$= -0.903$$

$$\cong -0.90$$

(b) $\bar{X} = 5.5$, $\bar{Y} = 4.13$

Table 6.6

X	Y	x	y	x^2	y^2	xy
6	4	0.5	-0.13	0.25	0.02	-0.065
8	2	2.5	-2.13	6.25	4.54	-5.325
5	5	-0.5	0.87	0.25	0.76	-0.435
4	6	-1.5	1.87	2.25	3.50	-2.805
7	3	1.5	-1.13	2.25	1.28	-1.695
3	7	-2.5	2.87	6.25	8.24	-7.175
5	3	-0.5	-1.13	0.25	1.28	0.565
6	3	0.5	-1.13	0.25	1.28	-0.565
				18.00	20.8752	-17.5

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

Then, $r = \frac{-17.5}{\sqrt{18} \times \sqrt{20.8752}}$

$$= \frac{-17.5}{4.243 \times 4.569}$$

$$= \frac{-17.5}{19.386}$$

$$= -0.903$$

Table 6.7

X	Y	Rank x	Rank y	d	d^2
6	4	3.5	4	-0.5	0.25
8	2	1	8	-7	49
5	5	5.5	3	2.5	6.25
4	6	7	2	5	25
7	3	2	6	-4	16
3	7	8	1	7	49
5	3	5.5	6	-0.5	0.25
6	3	3.5	6	-2.5	6.25
					152

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

Then $rho = 1 - \frac{6 \times 152}{8(8^2 - 1)}$

$= 1 - \frac{912}{504}$

$= 1 - 1.809$

$= 0.809$

$\cong 0.81$

Chapter Seven

THE z-TEST AND t-TEST STATISTICS

7.1 INTRODUCTION

The z-test and t-test statistics are two test statistics that confuse many students, both at the undergraduate and at the post graduate levels. The reason for the confusion is their very close operational methods. There seem to be very tiny difference outlining their conditions of application. This text attempts to observe the tiny difference.

A z-test by definition is any statistic test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution and for which the population is known. The most basic condition for the use of the z-test statistic is that it is used when the population is large ($N \geq 30$) given in Kothari (2011).

A t-test is an analysis of two population means through the use of statistical examination which require the use of a sample of the population. A t-test is mostly used with small samples drawn from the population. When the sample

required gets large the use of the two statistics is possible, but as the sample gets so large and becomes approximately the population, the z-test statistics become more appropriate.

7.2 ASSUMPTIONS OF THE z-TEST AND t-TEST STATISTICS

1. A z-test assumes that the population standard deviation (σ_P) is known, a t-test assumes that the sample standard deviation (σ_S) is what is rather known.
2. To carry out a t-test an estimate of the sample standard deviation must be carried out. This is not required for a z-test since it is σ_P that is known.
3. The samples from each population must be independent of one another.
4. The population from which the samples are taken must be large and normally distributed (i.e N≥30)
5. Based on the central limit theorem, most test statistics are normally distributed for large samples. This implies that each time we carry out a z-test the critical value is 1.96 because the entire population or a large sample is used.

7.3 COMPUTATION OF z-TEST

Two computational formulas are popular for determination of z-test statistic. The first formula is used when the population is infinite. In such a case the z-test statistic is given as:

$$z = \frac{\bar{x} - \mu}{\sigma_p / \sqrt{n}}$$

The second formula is used when the population is finite and it is given as:

$$z = \frac{\bar{x} - \mu}{\left(\sigma p / \sqrt{n}\right) \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

EXAMPLE 7.1

The mean performance of a sample of 250 candidates in a state common entrance was given as 56.81 while the population mean was given as 57.01 with a standard deviation of 4.11. Test at 0.05 level of significance to say whether or not the sample mean can be regarded as coming from that population.

Solution

To solve this problem we apply the formula

$$z = \frac{\bar{x} - \mu}{\sigma_p / \sqrt{n}}$$

where $\bar{x}$ = 56.81

μ = 57.01

σ_p = 4.11

n = 250

i.e $z = \frac{56.81 - 57.01}{4.11/\sqrt{250}}$

$$= \frac{-0.2}{4.11/15.8114}$$

$$= \frac{-0.2}{0.25994}$$

$$= -0.769$$

Decision:

The calculated z-value in absolute sense is 0.769. This value is less than the critical z-value of 1.96 (it has been explained that each time we carry out a z-test, the critical value is 1.96 since the entire population or a large sample is involved). Since 0.769 is less than the critical value of 1.96 the null hypothesis is accepted. This means that the sample can be regarded as being drawn from that population. In other words, the sample mean is not significantly different from the population mean.

EXAMPLE 7.2

An unknown population has a mean of 36.14 and standard deviation of 1.33. If a sample of 500 drawn from it has a mean of 39.36, test at 5% level of significance to say if the sample is a true representative of the population.

Solution:

If, $z = \frac{\bar{x} - \mu}{\sigma_p / \sqrt{n}}$

Then, $z = \frac{39.36 - 36.14}{1.33/\sqrt{500}}$

$$= \frac{3.22}{1.33/22.361}$$

$$= \frac{3.22}{0.059}$$

$$= 54.576$$

Decision:

The calculated value in this example is greater than the critical z-value of 1.96. This implies that there is a significant difference between the sample mean and the population mean. The sample cannot be regarded as a reasonable representative of the population.

EXAMPLE 7.3

From a student population of 1,200 a sample of 300 students yielded a mean height of 4.23 feet and population standard deviation of 0.86. Given the population mean as 4.19 feet, test at 0.05 level of significance, and find if the sample mean is significantly different from the population mean.

Solution:

Since the population is finite, the appropriate formula to test for z is given as:

$$z = \frac{\bar{x} - \mu}{\sigma_p / \sqrt{n} \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

Where, $\bar{x}$ = 4.23

$\mu = 4.19$

$\sigma_p = 0.86$

N = 1,200

n = 300

$$= \frac{4.23 - 4.19}{0.86/\sqrt{300} \times \left[\sqrt{(1200-300)/(1200-1)}\right]}$$

$$= \frac{0.04}{0.04965 \times \sqrt{900/1199}}$$

$$= \frac{0.04}{0.04965 \times 0.86639}$$

$$= \frac{0.04}{0.04301}$$

$$= 0.93$$

Decision:

The calculated value of 0.93 is less than the critical z-value of 1.96 at 0.05 level of significance. This means that there is no significant difference between the sample mean and the population mean. The decision here is to accept the null hypothesis.

It is important to note that the z-test statistic can also apply when two samples are drawn from a population or two

different populations provided that the samples are large ($n_1+n_2 \geq 30$). Let's consider the following example.

EXAMPLE 7.4

A state government was interested in finding out if the academic achievement of students in English Language was significantly different between Urban and Rural students. A sample of 200 Urban students and 200 Rural students was drawn and their achievement scores were summarized as following:

Urban students	Rural students
$\bar{x} = 68.27$	$\bar{x} = 63.94$
$\sigma_s = 11.93$	$\sigma_s = 10.61$

Solution:

In this kind of study, the σ_p for the two groups may not be necessary if σ_{s_1} and σ_{s_2} are known. The z-test can be applied because of the high number in the sample.

i.e $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_{s_1}^2}{n_1} + \frac{\sigma_{s_2}^2}{n_2}}}$$

$$= \frac{68.27 - 63.94}{\sqrt{\frac{11.93^2}{200} + \frac{10.61^2}{200}}}$$

$$= \frac{4.33}{\sqrt{0.7116 + 0.56286}}$$

$$= \frac{4.33}{\sqrt{1.27446}}$$

$$= \frac{4.33}{1.1289}$$

$$= 3.836$$

Decision:

The calculated z-value of 3.836 is greater than the critical z-value of 1.96 at 0.05 level of significance at infinite degrees of freedom (one does not need to calculate the degrees of freedom for a z-test because the number of subjects is normally high, thus making the degrees of freedom to be infinite. Note that the critical value for infinite degrees of freedom, as mentioned earlier, is 1.96). Based on this result the null hypothesis is rejected since there is a significant difference in academic achievement between Urban and Rural students. From the mean achievement scores, it means that the Urban students ($\bar{x}$=68.27), achieved significantly better than the Rural student ($\bar{x}$=63.94) in English Language.

7.4 COMPUTATION OF t-TEST

There are three computational methods for the t-test. It is however important to note that the t-test statistic is used each time the sample or population is less than 30. The three computational methods are:

i. The independent t-test statistic
ii. The population t-test statistic

iii. The dependent t-test or Correlational t-test statistic

7.4.1 INDEPENDENT t-TEST: Used when two independent variables, measured discretely are compared on a dependent variable that is measured continuously.

The formula for the independent t-test is give as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Degrees of freedom is given as $n_1 + n_2 - 2$

where:

$\bar{X}_1$ = Mean for group one

$\bar{X}_2$ = Mean for group two

σ_1 = Standard deviation for group one

σ_2 = Standard deviation for group two

n_1 = number of subjects in group one (sample size for group one)

n_2 = number of subjects in group two (sample size for group two).

7.4.2 POPULATION t-TEST

The population t-test is used when the mean observation of a sample is to be compared with the mean observation of an entire population. An example is when the mean performance of a given school in Mathematics is compared with the known state performance in that same Mathematics (Joshua, 2008). The formula for population t-test is given as:

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

Where,

$\bar{x}$ = Mean for the sample

μ = Mean for the entire population

σ_s = Standard deviation for the sample

n = Number of subjects in the sample

Degrees of freedom is given as n-1

7.4.3 DEPENDENT t-TEST

Another name for dependent t-test is related t-test since it is the suitable statistic for related samples. When a pair of responses is coming from a group, it is most appropriate to consider them as being dependent and compare them using the related t-test statistic. An example is when a pre-test and a post-test are given to the same set of students or when determining whether a sample of students performs significantly different in two school subjects.

In very simplistic terms, the formula for dependent t-test is given as:

$$t = \frac{\bar{D}}{\sigma / \sqrt{n}}$$

where: $\bar{D}$ = Mean of difference in pairs of the two samples given as $\sum D / n$

σ = Standard deviation of difference in pairs of scores given as:

$$\sigma = \sqrt{\frac{\Sigma D^2 - \left(\Sigma D\right)^2 / n}{n-1}}$$

n = number of pairs in the sample

Degrees of freedom are given as n-1.

EXAMPLE 7.5: INDEPENDENT t-TEST

Two independent samples are drawn from populations with the same variance. The sample results are as given below: $\bar{x}_1$ = 21.64, $\bar{x}_2$ = 18.81, σ_1 = 2.60, σ_2 = 2.52, n_1 = 12, n_2 = 12. Estimate the students t-value and report your result at 0.05 level of significance.

Solution:

The solution to this problem is straight forward. We will put down the appropriate formula and then substitute the information given with appropriate components of the formula.

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

$$= \frac{21.64 - 18.81}{\sqrt{\frac{(2.60)^2}{12} + \frac{(2.52)^2}{12}}}$$

$$= \frac{2.83}{\sqrt{\frac{6.76 + 6.35}{12}}}$$

$$= \frac{2.83}{\sqrt{1.093}}$$

$$= \frac{2.83}{1.045}$$

$$= 2.708$$

The calculated t-value $\cong 2.708$

Degrees of freedom = $n_1 + n_2 - 2$

$= 12 + 12 - 2$

$= 22$

Critical t-value (from the table of values) = 2.074

Decision: The calculated t-value of 2.708 is greater than the critical t-value of 2.074 at 0.05 level of significance using 22 degrees of freedom. This means that there is a significant difference between the sample means. Based on the result, reject the null hypothesis

EXAMPLE 7.6: INDEPENDENT t-Test

The following are the raw scores obtained by some male and female students in an English Language achievement test:

Male	55	61	50	60	55	73	68	65	45	57
Female	53	55	60	60	67	68	51	48	50	51

Test and explain the nature of difference between male and female students in English Language.

Solution:

To solve this problem we will first estimate the means and standard deviations for each set of data.

Let Male be X_1 and Female X_2

then $\bar{x}_1 = \frac{55+61+50+60+55+73+68+65+45+57}{10} = 58.9$

$\bar{x}_2 = \frac{53+55+60+60+67+68+51+48+50+51}{10} = 56.3$

Male

X	$x-\bar{x}_1$	$(x-\bar{x})^2$
55	-3.9	15.21
61	2.1	4.41
50	-8.9	79.21
60	1.1	1.21
55	-3.9	15.21
73	14.1	198.81
68	9.1	82.81
65	6.1	37.21
45	-13.9	193.21
57	-1.9	3.61
		630.90

Female

X	$x-\bar{x}_2$	$(x-\bar{x})^2$
53	-3.3	10.89
55	-1.3	1.69
60	3.7	13.69
60	3.7	13.69
67	10.7	114.49
68	11.7	136.89
51	-5.3	28.09
48	-8.3	68.89
50	-6.3	39.69
51	-5.3	28.09
		456.10

$\sigma_1 = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ $\sigma_2 = \sqrt{\frac{456.10}{9}}$

$\sigma_1 = \sqrt{\frac{630.90}{9}}$ $= \sqrt{50.68}$

$= \sqrt{70.1}$ $= 7.12$

$= 8.37$

We can now estimate the t-value as follows:

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

$$= \frac{58.9 - 56.3}{\sqrt{\frac{(8.37)^2 + (7.12)^2}{10}}}$$

$$= \frac{2.6}{\sqrt{\frac{70.1 + 50.7}{10}}}$$

$$= \frac{2.6}{\sqrt{12.08}}$$

$$= \frac{2.6}{3.48}$$

$$= 0.747$$

The calculated $t = 0.747$
Degree of freedom = $10 + 10 - 2$
= 18
Critical t = 2.101
Report: The calculated t-value of 0.747 is less than the critical t-value of 2.101 at 0.05 level of significance using

18 degrees of freedom. This means that the males do not differ from the females in their mean achievements in English Language.

Decision: Accept the null hypothesis.

EXAMPLE 7.7: INDEPENDENT t-Test

Calculate the t-value for the following data summaries emanating from chronological ages of students.

Urban	Rural
$\bar{x}_1 = 17.25$	$\bar{x}_2 = 18.90$
$\sigma_1 = 3.16$	$\sigma_2 = 3.23$
$n_1 = 15$	$n_2 = 12$

Solution:

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

$$= \frac{17.25 - 18.90}{\sqrt{\frac{(3.16)^2}{15} + \frac{(3.23)^2}{12}}}$$

$$= \frac{-1.65}{\sqrt{\frac{9.9856}{15} + \frac{10.4329}{12}}}$$

$$= \frac{-1.65}{\sqrt{0.6657 - 0.8694}}$$

$$= \frac{-1.65}{\sqrt{1.5351}}$$

$$= \frac{-1.65}{1.23899}$$

$$\cong -1.332$$

The results should be discussed on the absolute value or 1.332. The negative sign is discountenanced for an independent t-test.

EXAMPLE 7.8: POPULATION t-Test

A school principal interested in rating his schools performance on Mathematics in NECO examination for a given year got the mean performance of his school as 63.55 with standard deviation of 10.17 and mean performance in the entire state as 65.27. If there were 250 candidates in his school, test at 0.05 level of significance, the hypothesis that, the mean performance of the school is not significantly different from the mean performance of the entire state.

Solution:

In solving this problem we will consider the school's number of candidates as the sample. As far as the sample is clearly defined and it is to be compared with the entire population of candidates in the state, then the one sample t-test (population t-test) will be employed. It is important to note that population t-test is applicable for small samples ($n<30$), but other requirement is that it is the sample variance that must be used so the z-test cannot apply here.

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

$$= \frac{63.55 - 65.27}{10.17 / \sqrt{250}}$$

$$= \frac{-1.72}{10.17/15.81}$$

$$= \frac{-1.72}{0.643}$$

$$= -2.675$$

The calculated t-value is –2.675
The degrees of freedom = 250 – 1
= 249
The critical t-value at 0.05 level of significance = 1.96.

Decision: The calculated t-value of -2.675 is in absolute sense greater than the critical t-value of 1.96 at 0.05 level of significance using 249(∞) degrees of freedom. This means that the mean performance of the school is significantly different from the mean performance of the state. Based on this result, the null hypothesis is rejected.

EXAMPLE 7.9: POPULATION t-Test

In the past years a school's performance in an external examination was 51%. The school draws a sample of 15 students to determine whether the school is performing well now when compared with the past. The mean performance

of the students was 54% with variance of 9.28%. Test the hypothesis that the school is not performing significantly well using significance level of 0.05.

Solution

The first thing to do in solving the problem is to define the parameters involved.

μ = 51%

$\bar{x}$ = 54%

$\sigma = \sqrt{9.28}$ = 3.05 [Standard Deviation (σ) = square root of variance (σ^2)].

n = 15

then $t = \frac{54-51}{3.05/\sqrt{15}}$

$= \frac{3}{3.05/3.87}$

$= \frac{3}{0.788}$

$= 3.807$

The calculated t-value = 3.807
The degrees of freedom = 15-1 = 14
The critical t-value = 2.145

Report: Again in this example, the calculated t-value of 3.807 is greater than the critical t-value of 2.145 at 0.05 level of significance using 14 as the degrees of freedom. This means that the school is performing significantly well.
Decision: Reject the null hypothesis.

EXAMPLE 7.10: DEPENDENT t-TEST

A teacher administered a test to 10 students and went ahead to administer the test after two weeks to the same students. The result of the test was a follows:

Student	Test	Retest
1	6	8
2	5	6
3	6	6
4	7	5
5	8	9
6	5	6
7	4	5
8	3	3
9	6	5
10	9	8

Estimate the t-statistic and test the hypothesis at 0.05 level of significance.

Solution

Student	Test X_1	Retest X_2	D	D^2
1	6	8	-2	4
2	5	6	-1	1
3	6	6	-2	0

4	7	5	-2	4
5	8	9	1	1
6	5	6	-1	1
7	4	5	-1	1
8	3	3	0	0
9	6	5	1	1
10	9	8	1	1
Total			-6	14

$$t = \frac{\bar{D}}{\sigma/\sqrt{n}}$$

First, find $\bar{D}$

$$\bar{D} = \frac{\Sigma D}{n} = \frac{-6}{10} = -0.6$$

$$\sigma = \sqrt{\frac{\Sigma D^2 - (\Sigma D)^2/n}{n-1}}$$

$$= \sqrt{\frac{14-(-6)^2/n}{10-1}}$$

$$= \sqrt{\frac{14-3.6}{9}}$$

$$= \sqrt{1.156}$$

$$\cong 1.08$$

Then, $t = \frac{-0.6}{1.08/\sqrt{10}}$

$$= \frac{-0.6}{0.342}$$

$$= -1.754$$

Calculated t-value in absolute terms $= 1.754$
Degree of freedom = 10 – 1 = 9
Critical t-value = 2.262

The results show that there is no significant difference in the performance of the students in the pretest and post-test.

t-TEST FOR PEARSON r CORRELATION COEFFICIENT

The t-test is sometimes useful for testing hypothesis on Correlation. When pairs of scores for a sample are correlated using Pearson's Product Moment methods or Spearman Rank order method, the results do not mean anything than to report the extent of relationship. If the researcher must, however, draw conclusions about the population using the sample, the correlation coefficient (r) must be converted into a t statistic. This enables the researcher to test the null hypothesis that the population Correlation coefficient is zero and the Correlation (r) observed in the sample is a function of chance.

When the Pearson r is converted into a t-distribution, the calculated t can be compared with the table t (critical t) at N-2 degrees of freedom. This can enable the researcher take a decision as to whether the null hypothesis should be accepted or rejected.

The formula for t is given as:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

Where:

r = the Correlation Coefficient obtained from the sampled pair of scores.

N = The number of elements in the sample

Chapter Eight

ANALYSIS OF VARIANCE

8.1 INTRODUCTION

Developed in 1918 by Ronald Fisher, Analysis of Variance, referred to as ANOVA by acronym, is a collection of statistical techniques used to analyze the difference among group means and their associated procedures like variations among and between groups. Put some other way, ANOVA is used to determine whether there is a significant difference among means of three or more groups, concurrently at a selected probability level. Going by its name ANOVA focuses on variability of several measures. Analysis of Variance procedure is underlined by the fact that total variance of scores can be attributed to two sources (variance between groups caused by treatment and variance within the groups caused by error). This text examines two types of ANOVA, namely: One-way ANOVA and two-way ANOVA. However, before considering the two methods it will be important to consider the assumptions of ANOVA.

8.2 BASIC ASSUMPTIONS OF ANALYSIS OF VARIANCE

1. **Assumption of normality:** This assumption simply means that the population distribution should be normal. That is, the histogram produced from the data should produce a normal curve. Koul (2012) described this assumption as not being especially important. He explained that the F-ratio is insensitive to variations in the shape of the population distribution.
2. **Assumption of randomness:** This means that the sample used for the experiment should be randomly selected. That is, all heterogeneous units in the population should be relatively represented in the sample. For instance if two types of samples are to be drawn one from female and the other from the male, then the researcher must select the sample, strictly in proportion, from the two groups. This assumption is the nucleus of ANOVA technique, so failure to fulfill this assumption gives biased result.
3. **Assumption of homogeneity of Variance:** This assumption means that the sub-groups under study should have the same variance. When the variances across groups are not equal, the usual ANOVA assumptions are not satisfied and the F-ratio that will be obtained will not be valid. This assumption can be tested by applying Barlett's test of homogeneity of variance.

8.3 COMPUTATION OF ONE-WAY ANALYSIS OF VARIANCE

The Analysis of Variance Statistic is denoted by the letter 'F' and called the F-ratio. It is statistically given as:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

i.e $F = \frac{MSB}{MSW}$

Where: MSB = Mean Square Between

MSW = Mean Square Within

but, $MSB = \frac{SSB}{k-1}$

and, $MSW = \frac{SSW}{N-k}$

Where: SSB = Sum of Squares Between

K = number of groups in the independent variable

SSW = Sum of Squares Within

N = number of elements in the sample

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

or,

$$SSB = \left[\frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \cdots + \frac{\left(\sum X_k\right)^2}{n_k} \right] - \frac{\sum \left(X_1, X_2, \cdots X_k\right)^2}{N}$$

Where X_k is the last group of the independent variable

$$SSW = \Sigma\Sigma X^2 - \sum \frac{\left(\Sigma X\right)^2}{n}$$

or,

$$\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \left[\frac{\left(\Sigma X_1\right)^2}{n_1} + \frac{\left(\Sigma X_2\right)^2}{n_2} + \cdots + \frac{\left(\Sigma X_k\right)^2}{n_k} \right]$$

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

or,

$$SST = SSB + SSW$$

where: SST = Sum of Squares Total

EXAMPLE 8.1

A study involving the academic performance of 18 students taught with three different methods produced the following raw scores.

Method A	Method B	Method C
3	4	1
4	3	1
2	3	3
5	5	2
4	7	3
6	2	4

Test at 0.05 level of significance, the hypothesis that method of teaching has no significant influence on students' academic performance.

Solution:

Let Method A = X_1

Method B = X_2

Method C = X_3

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
3	9	4	16	1	1
4	16	3	9	1	1
2	4	3	9	3	9
5	25	5	25	2	4
4	16	7	49	3	9
6	36	2	4	4	16
24	106	24	112	14	40

$\Sigma X_1 = 24$

$\Sigma X_1^2 = 106$

$\Sigma X_2 = 24$

$\Sigma X_2^2 = 112$

$\Sigma X_3 = 14$

$\Sigma X_3^2 = 40$

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

$$= \left[\frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_k\right)^2}{n_3} \right] - \frac{\sum \left(X_1 + X_2 + X_3\right)^2}{N}$$

$$= \left[\frac{(24)^2}{6} + \frac{(24)^2}{6} + \frac{(14)^2}{6} \right] - \frac{(24+24+14)^2}{18}$$

= 96+96+32.67-213.55

= 224.67-213.55

= 11.12

$$SSW = \Sigma\Sigma X^2 - \sum \frac{\left(\Sigma X\right)^2}{n}$$

$$= 106 + 112 + 40 - \frac{(24)^2}{6} + \frac{(24)^2}{6} + \frac{(14)^2}{6}$$

$$= 258 - (96 + 96 + 32.67)$$

$$= 258 - 224.67$$

$$= 33.33$$

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

= 258-213.55

= 44.45

OR

$$SST = SSB + SSW$$

= 11.12+33.33

= 44.45

From the computations above:

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

= 11.12

3-1

= 5.56

$$MSW = \frac{SSW}{N\_K}$$

= 33.33

18-3

= 2.222

From these calculated Mean Square values the F-ratio can be calculated as follows:

$$F = \frac{MSB}{MSW}$$

$$= \frac{5.56}{2.963}$$

$$= 2.502$$

The calculated F-ratio = 2.502

The degrees of freedom are k-1 and N-k = 2 & 18

Critical F-ratio from the table of values using 2 and 18 degrees of freedom at 5% level of significance is 3.55.

Report: The calculated F-ratio of 2.502 is less than the critical F-ratio of 3.55 at 0.05 level of significance. This means that teaching methods used did not significantly influence students' academic performance.

Decision: Accept the null hypothesis (H_0).

EXAMPLE 8.2

A researcher was interested in finding out if there is a significant influence of teachers' years of teaching experience on their job performance. The following were the job performance ratings of the teachers.

Below 10 years	10-15 years	16-20 years	Above 20 years
10	14	18	22
6	10	15	20
12	10	10	18
8	12	16	23
10	11	15	20
9	8	12	

Solution:

From the information let

Below 10 years = X_1

10-15 years = X_2

16-20 years = X_3

Above 20 years = X_4

The table will look like this:

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
10	100	14	196	18	324	22	484
6	36	10	100	15	225	20	400
12	144	10	100	10	100	18	324
8	64	12	144	16	256	23	529
10	100	11	121	15	225	20	400
9	81	8	64	12	144		
55	525	65	725	86	1274	103	2137

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

$$= \frac{55^2}{6} + \frac{65^2}{6} + \frac{86^2}{6} + \frac{103^3}{5} - \frac{(55+65+86+103)^2}{23}$$

$$= 504.17 + 704.17 + 1232.67 + 2121.80 - \frac{(309)^2}{23}$$

$$= 4562.81 - 4151.35$$

$$= 411.46$$

$$SSW = \Sigma\Sigma X^2 - \sum \frac{(\Sigma X)^2}{n}$$

$$= 525 + 725 + 1274 + 2137 - 4562.81$$

$$= 4661 - 4562.81$$

$$= 98.19$$

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

$$= 4661 - 4151.35$$

$$= 509.65$$

OR

$$SST = SSB + SSW$$

$$= 411.46 + 98.19$$

$$= 509.65$$

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

$$= \frac{411.46}{4-1} = \frac{411.46}{3}$$

$$= 137.15$$

$$MSW = \frac{SSW}{N-K}$$

$$= \frac{98.19}{23-4} = \frac{98.19}{19}$$

$$= 5.17$$

$$F = \frac{MSB}{MSW}$$

$$= \frac{137.15}{5.17}$$

$$= 26.53$$

The calculated F-ratio = 26.53

The degrees of freedom are k-1 and N-k = 3 and 19
Critical F-ratio from the table of values using 3 and 19 degrees of freedom at 5% level of significance is 3.15.

Decision: The calculated F-ratio of 26.53 is greater than the critical F-ratio of 3.15 at 0.05 level of significance. This means

that teachers' years of teaching experience significantly influence their job performance ratings.

8.4 POST-HOC COMPARISON TEST

When the F-ratio of ANOVA test is greater than the table value (critical value) it means there is a significant influence or relationship, but it does not tell the researcher the mean difference responsible for the influence. This is because in every ANOVA test, the groups compared are always three or more. The only way to identify the mean group difference responsible for such influence is to carry out a Post-Hoc comparison test. There are many methods apart from the popular Fisher's Least Significant Difference (LSD) method. For all the methods, however, what is important is to find the difference among group means. The report can be given using the group difference by considering the significant group difference as the likely reason for the significant F-ratio.

A Fisher's t-test can be carried out to give room for a more straight forward explanation of the group mean difference. The formula for the t-test is given as:

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{MSW\left(\frac{1}{n_a} - \frac{1}{n_b}\right)}}$$

Where: $\bar{X}_a$ and $\bar{X}_b$ = Mean of the two groups to be compared

MSW = Mean Square Within

n_a and n_b = number of scores in the two groups to be compared.

8.5 FACTORIAL ANALYSIS OF VARIANCE

The analysis of variance (One-way ANOVA) so far presented has been single-factor design. In it, the researcher is interested in using one independent variable, with dimensions, to produce an influence on the dependent variable. It is sometimes possible that the researcher has more than one independent variable (each with dimensions) to be compared with a single dependent variable. When such situation arises, the design likely to be used is called factorial design, so the experimenter will employ a factorial analysis of variance statistic. A good example is two-way Analysis of Variance (Two-Way ANOVA), where two independent variables are used to examine one dependent variable (Adegoke, 2014).

8.6 STEPS FOR COMPUTATION OF TWO-WAY ANOVA (adapted from Ary, Jacob & Razaveith, 1985)

Step 1: Find the total sum of squares, the sum of squares between groups, and the sum of squares within groups. The procedure for calculating the sum of squares for Two-Way ANOVA is the same as for One-Way ANOVA.

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

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

$$SSW = \Sigma\Sigma X^2 - \sum \frac{\left(\Sigma X\right)^2}{n}$$

Step 2: Break down the sum of the squares between groups into three separate sums of squares:
(a) The sum of squares between columns
(b) The sum of squares between rows, and
(c) The sum of squares for interaction between columns and rows.

(a) $$SSBC = \frac{\left(\Sigma X_{c_1}\right)^2}{nc_1} + \frac{\left(\Sigma X_{c_2}\right)^2}{nc_2} + \cdots + \frac{\left(\Sigma\Sigma X\right)^2}{N}$$

(b) $$SSBR = \frac{\left(\Sigma X_{r_1}\right)^2}{nr_1} + \frac{\left(\Sigma X_{r_2}\right)^2}{nr_2} + \cdots + \frac{\left(\Sigma\Sigma X\right)^2}{N}$$

(c) $$SSINT = SSB - (SSBC + SSBR)$$

Step 3: Determine the number of degrees of freedom (df) associated with each source of variation.
df for between-columns sum of squares = C-1
df for between-rows sum of squares = R-1
df for interaction = (C-1)(R-1)
df for between-groups sum of squares = G-1
df for within-groups sum of squares = N-G
df for total sum of squares = N-1
Where:

C = the number of columns
R = the number of rows
G = the number of groups
N = the number of subjects in all groups

Step 4: Find the mean-square values by dividing each sum of squares by its associated number of degrees of freedom.

Step 5: Compute the F-ratios for the main and the interaction effects by dividing the between-groups mean squares by the within-groups mean squares for each of the three components.
Step 6: Present the results in a Table and check table of values for significance of each of the value.

EXAMPLE 8.3
Assume that the following scores are observations from a two-way design showing performance of 30 students (15 males, 15 females) with different levels of intelligence. Find how sex and level of intelligence affect the performance of the students (Ubi & Bassey, 2012).

Sex	Intelligence level		
	High	Average	Low
Male	6	6	5
	8	6	6
	7	5	3
	6	4	4
	7	5	3
Female	6	5	4
	4	4	3
	7	4	3
	5	5	5
	7	6	2

Solution

To answer the question a computational Table will be prepared (Let High = X_1, Average = X_2, and Low = X_3).

Sex			Intelligence level						
	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	ΣX_R	$\bar{X}_R$	ΣX_R^2
Male	6 8 7 6 7	36 64 49 36 49	6 6 5 4 5	36 36 25 16 25	5 6 3 4 3	25 36 9 16 9			
ΣX	34	234	26	138	21	95	81		467
$\bar{X}$	**6.8**		**5.2**		**4.2**			**5.4**	
Female	6 4 7 5 7	36 16 49 25 49	5 4 4 5 6	25 16 16 25 36	4 3 3 5 2	16 9 9 25 4			
ΣX	29	175	24	118	17	63	70		356
$\bar{X}$	**5.8**		**4.8**		**3.4**			**4.7**	
ΣX_C	63		50		38		151		
$\bar{X}_C$	6.3		5.0		3.8			5.03	
ΣX_C^2		409		256		158			823

Where:

C = Column

R = Row

STEP 1:

$$SST = 823 - \frac{151^2}{30}$$

$$= 823 - 760.03$$

$$= 62.97$$

$$SSB = \frac{34^2}{5} + \frac{26^2}{5} + \frac{21^2}{5} + \frac{29^2}{5} + \frac{24^2}{5} + \frac{17^2}{5} - \frac{(151)^2}{30}$$

$$= 231.2 + 135.2 + 88.2 + 168.2 + 115.2 + 57.8 - 760.03$$

$$= 795.8 - 760.03$$

$$= 35.77$$

$$SSW = 823 - 795.8$$

$$= 27.2$$

STEP 2:

$$SSBC = \frac{63^2}{10} + \frac{50^2}{10} + \frac{38^2}{10} - \frac{(151)^2}{30}$$

$$= 396.9 + 250 + 144.4 - 760.03$$

$$= 791.3 - 760.03$$

$$= 31.27$$

$$SSBR = \frac{81^2}{15} + \frac{70^2}{15} - \frac{(151)^2}{30}$$

$$= 437.4 + 326.67 - 760.03$$

$$= 764.07 - 760.03$$

$$= 4.04$$

$$SSINT = 35.77 - (31.27 + 4.04)$$

$= 35.77 - 35.31$

$= 0.46$

Step 3: df between-columns sum of squares = 3-1 = 2

df between-rows sum of squares = 2-1 = 1

df for interaction = $2 \times 1 = 2$

df for between-groups sum of squares = 6-1 = 5

df for within-groups sum of squares = 30-6 = 24

df for total sum of squares = 30-1 = 29

Step 4: Mean Square (MS) values:

MS between-columns (intelligence) = $31.27 \div 2 = 15.64$

MS between-rows (sex) = $4.04 \div 1 = 4.04$

MS Interaction (Intelligence and sex) = $0.46 \div 2 = 0.23$

MS between-groups = $35.77 \div 5 = 7.154$

MS within-groups = $27.2 \div 24 = 1.133$

Step 5: F-values (ratios)

F-ratio for between-columns (intelligence) = $15.64 \div 1.133 = 13.804$

F-ratio for between-rows (sex) = $4.04 \div 1.133 = 3.566$

F-ratio for interaction = $0.23 \div 1.133 = 0.203$

Results:

The results show that the F-ratios for individual effects of intelligence and sex were significant compared to the Table values. The F-ratio for interaction was not significant. The result implies that there is no interactive effect of intelligence and sex on

students' academic performance. To find the significance of each F-ratio, as usual we consult the Table of F-values for the critical value. If the calculated value is greater than the critical value (the value in the Table), using appropriate degrees of freedom for the given F-ratio, the null hypothesis is rejected. Conversely, if the calculated value is less than the critical value, the null hypothesis is accepted.

8.7 EXERCISE EIGHT

1. Write short notes on the following assumptions of ANOVA:
   a. Assumption of normality
   b. Assumption of randomness
   c. Assumption of homogeneity of variance.
2. Consider the following scores for three randomly formed groups and find out whether the sets of scores are significantly different or not:

X_1	X_2	X_3
2	2	3
4	3	3
3	4	3
4	5	5
2	6	7

3. Suppose a researcher is interested in investigating how mode of entry into University (UME or Direct Entry) and students' economic status (High, Average, & Low) affects their academic performance. Use the following scores obtained from the study to compute a Two-Way Analysis of Variance and say whether or not mode of entry interacted

with students' economic status in significantly affecting students' performance.

Mode of Entry	Economic Status		
	High	Average	Low
UME	16 15 14 13 12	15 14 13 12 11	15 14 11 10 10
Direct Entry	14 13 12 11 10	16 15 14 13 12	15 14 13 12 11

Chapter Nine

CHI-SQUARE STATISTIC

9.1 INTRODUCTION

The Chi-square statistic is a non-parametric inferential statistic used to determine the difference in the distributions of one or two categorical variables. Introduced by Karl Pearson in 1900, the Chi-square test can be described as a measure of relationship, association or independence of variables. It is adjudged the best known and most important of all non-parametric statistical methods. The main principle in a Chi-square test is that it is a measure of reliability between observed frequency distribution with theoretically expected frequency.

As earlier mentioned in Chapter two of this text (precisely in 2.1), there are basically two types of random variables; discrete and continuous variables. Discrete variables are categorical and in frequencies. Examples include; sex, number of houses, number of birds in a poultry, etc. Continuous variables are numerical or metric like wages of employees, heights of school children, distances, scores from a performance test, etc. When data are in frequencies or in counts of categorical responses from independent groups, the Chi-square statistic is the most appropriate statistical

tool to use. The Chi-square statistic is denoted with the sign χ^2. The formula is given as:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

χ^2 = Chi-square

O = Observed frequency

E = Expected frequency

9.2 TYPES OF CHI-SQUARE TESTS

There are two types of Chi-square test even when the formula for the two is the same.

i. **One variable Chi-square**: This type of Chi-square, also referred to as Goodness of fit test, is used to determine whether a sample data fits a population. It is applied when you have one categorical variable from a single population and you wish to determine whether the sample data observed are consistent with hypothesized distribution (expected frequency). Degrees of freedom for one variable Chi-square are given as k-1, where k represents number of categories.

ii. **Contingency Chi-square**: This type of Chi-square is also called Chi-square for independence. It compares two variables that are measured categorically to see if the variables differ from one another. An example is when a researcher wants to find the association between sex (male & female) and party affiliations (PDP, APC, Labour, APGA & AD) of a sample. The two variables are categorized, so contingency Chi-square is most appropriate. Degrees of

freedom are (C-1)(R-1), where C = number of columns and R = number of rows.

9.3 COMPUTATION OF CHI-SQUARE

EXAMPLE 9.1: One variable Chi-square

The preferences of 50 students for five political parties were sampled by a researcher to determine if there is significant association between observed and expected frequencies. The following are the data summaries:

Political party	PDP	APC	LABOUR	APGA	AD
Frequencies	18	10	14	6	2

Analyse the data and give results at 5% level of significance.

Solution:

First we will estimate the expected frequencies. If there were no variations in their preferences for the parties then each party will have a frequency of 10.

i.e $\frac{18+10+14+6+2}{5}=10$

So the expected frequency for each political party is 10. We can prepare a new Table as follows:

Political party	PDP	APC	Labour	APGA	AD
Observed Frequencies	18	10	14	6	2
Expected Frequencies	10	10	10	10	10

if $\chi^2 = \sum \frac{(O-E)^2}{E}$

then, $\chi^2 = \frac{(18-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(2-10)^2}{10}$

$$= \frac{64}{10} + \frac{0}{10} + \frac{16}{10} + \frac{16}{10} + \frac{64}{10}$$

$$= 6.4 + 0 + 1.6 + 1.6 + 6.4$$

$$= 16.0$$

The calculated $\chi^2 = 16.0$

Degrees of freedom = K-1 = 5-1 = 4

Then, the critical χ^2 at 5% (0.5) level of significance = 9.49

Result: Since the calculated value of 16.0 is greater than the critical value, we will conclude that the observed frequencies are significantly associated with the expected frequencies.

This problem can be presented in tabular form:

Political party	O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
PDP	18	10	8	64	6.4
APC	10	10	0	0	0
Labour	14	10	4	16	1.6
APGA	6	10	-4	16	1.6
AD	2	10	-8	64	6.4
TOTAL	50	50	0		$\chi^2 = 16.0$

EXAMPLE 9.2: CONTINGENCY CHI-SQUARE

Suppose the 50 students in example 9.1 were from three Faculties of a University and the researcher, this time, wants to find out whether the students affiliations to the parties were significantly because of the faculties they belong to. Use the raw data below to analyse the data at 5% level of significance.

		Political party				
		PDP	APC	Labour	APGA	AD
Faculty	Education	6	3	7	2	0
	Management	8	4	5	3	1
	Science	4	3	2	1	1

Solution

Step 1:

Let's reproduce the Table and include totals and also identify cells:

Variable	Categories	Political party					
		PDP	APC	Labour	APGA	AD	TOTAL
Faculty	Education	Cell 1 6(6.48)	Cell 2 3(3.6)	Cell 3 7(5.04)	Cell 4 2(2.16)	Cell 5 0(0.72)	18
	Management	Cell 6 8(7.56)	Cell 7 4(4.20)	Cell 8 5(5.88)	Cell 9 3(2.52)	Cell 10 1(0.84)	21
	Science	Cell 11 4(3.96)	Cell 12 3(2.20)	Cell 13 2(3.08)	Cell 14 1(1.32)	Cell 15 1(0.44)	11
	TOTAL	18	10	14	6	2	50

Frequencies expected are in parenthesis. Aggregate of expected frequencies for each column and row should equal aggregate of observed frequencies for the column or row.

Step 2:

Now let's estimate 'E' cell by cell. The formula for estimating 'E' is given as:

$$E = \frac{CT \times RT}{GT}$$

where: CT = Column Total

RT = Row Total

GT = Grand Total

Cell 1: $E = \frac{18 \times 18}{50} = 6.48$

Cell 2: $E = \frac{18 \times 10}{50} = 3.60$

Cell 3: $E = \frac{18 \times 14}{50} = 5.04$

Cell 4: $E = \frac{18 \times 6}{50} = 2.16$

Cell 5: $E = \frac{18 \times 2}{50} = 0.72$

} = 18

Cell 6: $E = \frac{21 \times 18}{50} = 7.56$

Cell 7: $E = \frac{21 \times 10}{50} = 4.20$

Cell 8: $E = \frac{21 \times 14}{50} = 5.88$

Cell 9: $E = \frac{21 \times 6}{50} = 2.52$

Cell 10: $E = \frac{21 \times 2}{50} = 0.84$

} = 21

Cell 11: $E = \frac{11 \times 18}{50} = 3.96$

Cell 12: $E = \frac{11 \times 10}{50} = 2.20$

Cell 13: $E = \frac{11 \times 14}{50} = 3.08$ } = 11

Cell 14: $E = \frac{11 \times 6}{50} = 1.32$

Cell 15: $E = \frac{11 \times 2}{50} = 0.44$

Step 3: let's compute the Chi-square

i.e

$$\chi^2 = \frac{(6-6.48)^2}{6.48} + \frac{(3-3.6)^2}{3.6} + \frac{(7-5.04)^2}{5.04} + \frac{(2-2.16)^2}{2.16} + \frac{(0-0.72)^2}{0.72} +$$
$$\frac{(8-7.56)^2}{7.56} + \frac{(4-4.20)^2}{4.20} + \frac{(5-5.88)^2}{5.88} + \frac{(3-2.52)^2}{2.52} + \frac{(1-0.84)^2}{0.84} +$$
$$\frac{(4-3.96)^2}{3.96} + \frac{(3-2.20)^2}{2.20} + \frac{(2-3.08)^2}{3.08} + \frac{(1-1.32)^2}{1.32} + \frac{(1-0.44)^2}{0.44}$$

$$= \frac{0.2304}{6.48} + \frac{0.36}{3.6} + \frac{3.8416}{5.04} + \frac{0.0256}{2.16} + \frac{0.5184}{0.72} + \frac{0.1936}{7.56} + \frac{0.04}{4.20} + \frac{0.7744}{5.88} +$$
$$\frac{0.2304}{2.52} + \frac{0.0256}{0.84} + \frac{0.0016}{3.96} + \frac{0.64}{2.20} + \frac{1.1664}{3.08} + \frac{0.1024}{1.32} + \frac{0.3136}{0.44}$$

$$= 0.0356 + 0.1 + 0.7622 + 0.0119 + 0.72 + 0.0256 + 0.0095 + 0.1317 +$$
$$0.0914 + 0.0305 + 0.0004 + 0.2909 + 0.3787 + 0.07758 + 0.7127$$

= 3.37868

The calculated χ^2 = 3.37868
The Degrees of freedom = (C-1)(R-1) = (5-1)(3-1) = $4 \times 2 = 8$
The critical χ^2 at 5% level of significance = 15.51

Result: Since the calculated χ^2 of 3.37868 is less than the critical χ^2 of 15.51, we will conclude that there is no association between the faculty of study and political party preferences of the students.

These computations can be presented in tabular form:

Cells	O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
1	6	6.48	-0.48	0.2304	0.0356
2	3	3.60	-0.6	0.3600	0.1000
3	7	5.04	1.96	3.8416	0.7622
4	2	2.16	-0.16	0.0256	0.0119
5	0	0.72	-0.72	0.5184	0.72
6	8	7.56	0.44	0.1936	0.0256
7	4	4.20	-0.20	0.0400	0.0095
8	5	5.88	-0.88	0.7744	0.1317
9	3	2.52	0.48	0.2304	0.0914
10	1	0.84	0.16	0.0256	0.0305
11	4	3.96	0.04	0.0016	0.0004
12	3	2.20	0.80	0.6400	0.2909
13	2	3.08	-2.28	1.1664	0.3787
14	1	1.32	-0.32	0.1024	0.07758
15	1	0.44	0.56	0.3136	0.7127
Total					χ^2 =3.37868

9.3 YATE'S CORRECTION

This is a correction made to make up for any contingency Chi-square with less than 10 cells. Yate (1934) suggested this correction that adjusts the formula for Pearson's Chi-square test by subtracting 0.5 from the difference between each observed value and its expected value in a 2 by 2 contingency table. The main reason for Yate's correction is to prevent over estimation of statistical significance for small data. The formula for Yate's correction is given as:

$$\chi^2 = \sum \frac{(O - E - 0.5)^2}{E}$$

EXAMPLE 9.4

Suppose the research situation in example 9.1 involved four instead of five political parties as follows:

Political party	PDP	APC	LABOUR	APGA
Frequencies	18	10	14	8

Use Yate's correction methods to test for significance at 5% level.

Solution:

As usual, we will first estimate the expected frequencies

i.e $E = \frac{18+10+14+8}{4} = \frac{50}{4} = 10.5$

If, $\chi^2 = \sum \frac{(O - E - 0.5)^2}{E}$

then,

$$\chi^2 = \frac{(18-10.5-0.5)^2}{10.5} + \frac{(10-10.5-0.5)^2}{10.5} + \frac{(14-10.5-0.5)^2}{10.5} + \frac{(8-10.5-0.5)^2}{10.5}$$

$$= 4.667 + 0.095 + 0.857 + 0.857$$

$= 6.476$

Calculated $\chi^2 = 6.476$

Degrees of freedom = K-1 = 4-1 = 3

Critical χ^2 = 7.81

Result: The calculated χ^2 value of 6.476 is less than the critical χ^2 value of 7.81 at 5% level of significance using 3 degrees of freedom. This means that observed frequencies do not significantly differ from expected frequencies with regards to students' preference for political parties.

9.5 EXERCISE NINE

1. (a) What is Chi-square test of goodness of fit
   (b) Differentiate between a goodness of fit test and a contingency Chi-square test.
2. In a class of students of different ages, the ages 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, were found to have the following frequencies:

Ages	11	12	13	14	15	16	17	18	19	20
Frequencies	2	3	5	8	8	6	4	2	1	1

Test whether the observed frequencies are significantly different from those expected on the hypothesis of uniform distribution.

3 A survey of 200 'Married', 'Single' and 'Separated' civil servants showed that 87 of them were in the "Head office", 42 in 'zonal offices' and 71 in 'state offices'. Test the hypothesis that there is no significant relationship between marital status and work place location.

4 (a) What is Yate's Correction

(b) Discuss the relevance of Yate's correction in a 2 by 2 contingency study.

(c) The following were the choices of candidates for courses in an Education faculty of a University.

Courses:	EduAdmin	EduArt	EduScience	EduSpecial
Frequencies	125	143	54	35

Calculate the Chi-square with Yate's correction for the data summaries.

Chapter Ten

REGRESSION ANALYSIS

(With some adaptations from Ubi & Bassey, 2012)

10.1 INTRODUCTION

Regression analysis refers to the determination of a statistical relationship between two or more variables. One or two or more of the variable is/are used to predict another single variable through an analysis of relationship. The variable(s) that predicts the relationship is/are the independent variable(s), while the other one being predicted (the criterion) is the dependent variable. The relationship in a regression analysis must interpret things that exist physically. In other words, the independent variable(s) "X" must have affected the dependent variable "Y" by a physically measurable quantity. The relationship of such physical measurable implication is normally interpreted in what is known as a linear regression equation given as:

$$y' = a + bx$$

A hypothetical example in which the regression equation can be used is in predicting final achievement of students in the Senior

School Certificate Examinati0n (SSCE) using their test anxiety. In this example achievement in SSCE is the dependent variable (Y) while test anxiety is the independent variable (X). This is shown in figure 10.1.

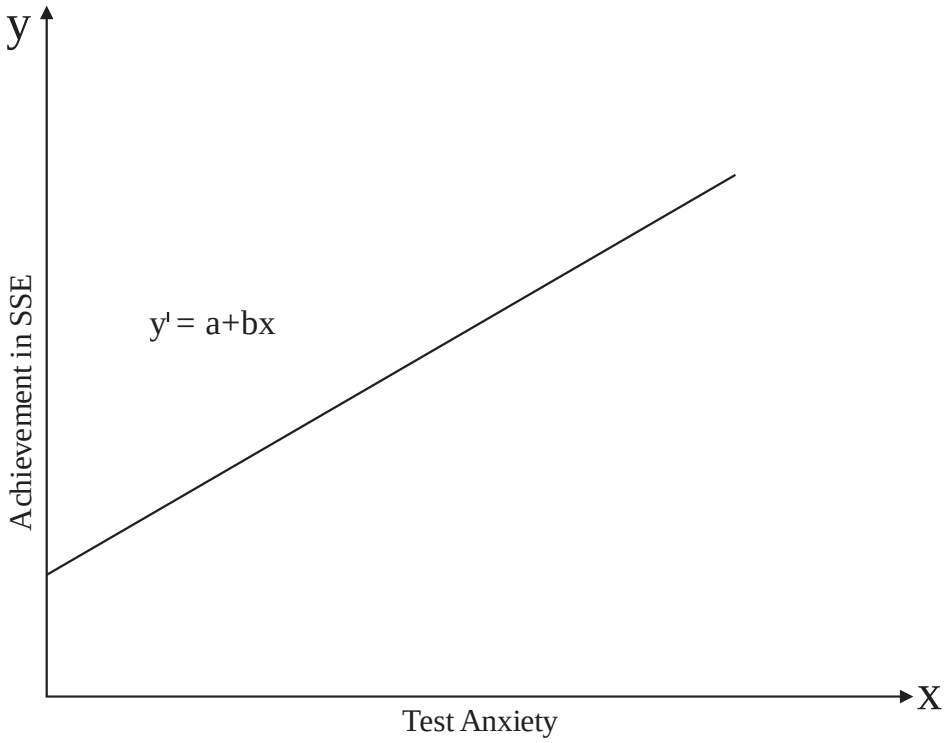


Figure 10.1: A hypothetical linear regression line

The graphical representation in figure 10.1 shows that the independent variable (X) is represented along the x-axis while the dependent variable (Y) is represented along the y-axis. From the equation 'a' is the intercept of the line and 'b' the slop of the line explaining the increase in y-axis resulting from per unit increase in

x-axis. When only one independent variable is involved, the equation is called simple linear regression equation. If the independent variables are two or more the equation is called multiple linear regression equation.

10.2 SOME KEY CONCEPTS USED IN REGRESSION ANALYSIS

Correlation coefficient (R):

This is the correlation coefficient between the independent variable (s) and the dependent variable. It measures the strength of association between the two (normally considered as x & y).

Coefficient of Determination (R^2)

This is a value that explains the power of the independent variable in predicting the dependent variable. It shows the extent to which the variation of the dependent variable about its mean is explained by the independent variable (s). Statistically, the higher the R^2 value, the better the prediction of the dependent variable by the independent variable (s) of the study. Consequently, $1-R^2$ represents the extent to which the variation of the dependent variable about its means is as a result of variables extraneous to the study. The lower the value of $1-R^2$, the higher will be the value of R^2.

Intercept (a)

The intercept in a regression equation is that value on the y-axis of a regression graph where the regression line crosses the axis. It is designated by the constant term 'a' in any given regression equation.

Regression coefficient (b)

This is the numerical value of any parameter estimate that is directly associated with the independent variables. In a regression equation $y = a + b_1x_1 + b_2x_2 + b_3x_3$, for instance, b_1, b_2 and b_3 are regression coefficients for the variables x_1, x_2 and x_3 respectively.

10.3 ASSUMPTIONS OF REGRESSION ANALYSIS

There are basically four assumptions of regression analysis that are not robust to violation and can be dealt with by a researcher if violated. These assumptions are normality, linearity, reliability of measurement and homoscedasticity. Let's briefly explain these assumptions. Detailed presentation of these assumptions will be beyond the scope for this text. See Osborne and Waters (2002) for an indept explanation of these assumptions.

i. Assumption of normality

Variables of study are assumed in regression to have normal distribution. This means that variables should not be highly skewed, Kurtotic or have substantial outliers. Simple histograms or frequency distributions of the data can reveal whether or not the variables have normal distribution.

ii. Assumption of linearity

This assumption means that the independent variable should have a linear relationship with the dependent variable. A non-linear relationship can increase the chance of a Type II error for the independent variable (under-estimation) and increase the chance of Type I error (over-estimation) for other independent variables that share variance with that independent variable in a multiple regression.

iii. Assumption of reliability

This assumption emphasizes that variables are measured without error.

iv. Assumption of homoscedacity

This is the assumption that the variance of errors is the same across all levels of the independent variable. Heteroscedascity refers to a situation when the variances of error differ at different values of the independent variable. High heteroscedacity can lead to serious distortion of findings that lead to Type I error.

10.4 SIMPLE REGRESSION

In a simple regression, we have only two variables, one of them being the predictor (independent) variable and the other being the criterion (dependent variable). A simple regression equation is given as:

$y' = a + bx$

Where:

$a = \bar{y} - b\bar{x}$, and

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

y' = the predicted y

a = the intercept (constant)

b = the regression coefficient

x = the independent variable

EXAMPLE 10.1
COMPUTATION OF SIMPLE REGRESSION

The following are scores obtained by six pupils in two separate examinations. Score X is Entrance Examination and score Y is First Term Examination.

Person	X	Y
1	6	10
2	7	8
3	8	11
4	9	9
5	10	14
6	11	15

i. Estimate the intercept 'a' and slop 'b' for the regression equation.
ii. Fit a regression equation to predict performance in First term examination.
iii. Estimate Y' (y predicted) for all the persons.
iv. Calculate the Correlation coefficient (R), coefficient of determination (R^2), and 1-R^2.

Solution

X	Y	$X-\bar{X}$ x	$Y-\bar{Y}$ y	x^2	y^2	xy	Y' $(a+bx)$	$(Y-Y')^2$
6	10	-2.5	-1.1667	6.25	1.3612	2.9168	8.23857	3.1026
7	8	-1.5	-3.1667	2.25	10.0280	4.7500	9.41114	1.9913
8	11	-0.5	-0.1667	0.25	0.0278	0.0834	10.58371	0.1733
9	9	0.5	-2.1667	0.25	4.6946	-1.0834	11.75628	7.5971
10	14	1.5	2.8333	2.25	8.0276	4.2500	12.92885	1.1474
11	15	2.5	3.8333	6.25	14.6942	9.5833	14.10142	0.8074
51	**67**	**0**	**-0.0002**	**17.50**	**38.833**	**20.5001**		**14.8191**

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

$$= \frac{6+7+8+9+10+11}{6} = 8.5$$

$$\bar{Y} = \frac{\Sigma y}{n}$$

$$= \frac{10+8+11+9+14+15}{6} = 11.1667$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{20.5001}{17.50}$$

$$= 1.171$$

$$a = \bar{Y} - b\bar{X}$$

$$= 11.1667 - 1.171(8.5)$$

$=11.1667-9.9535$

$=1.2132$

10.5: GRAPH OF X AND Y, X AND Y'

The graph of X and Y will be horizontal (parallel to the x-axis) to indicate deviations of Y from the mean of $Y(Y-\bar{Y})$. When the deviations are added the result is approximately zero (0). This is shown in Figure 10.2.

The graph of X and Y' is a linear relationship showing that observed scores will always predict a slop (b). This is shown in Figure 10.3

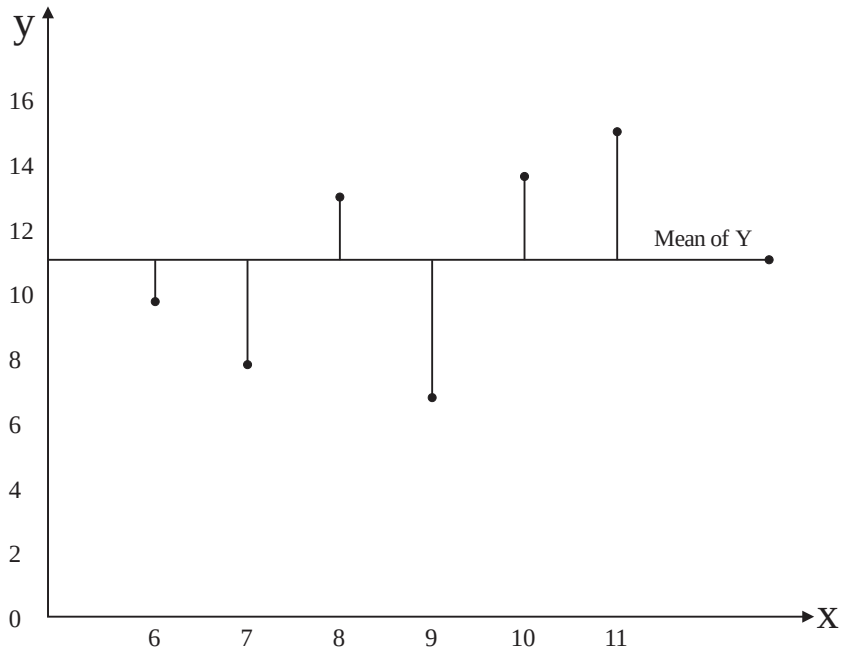


Figure 10.2: Graph of X and Y

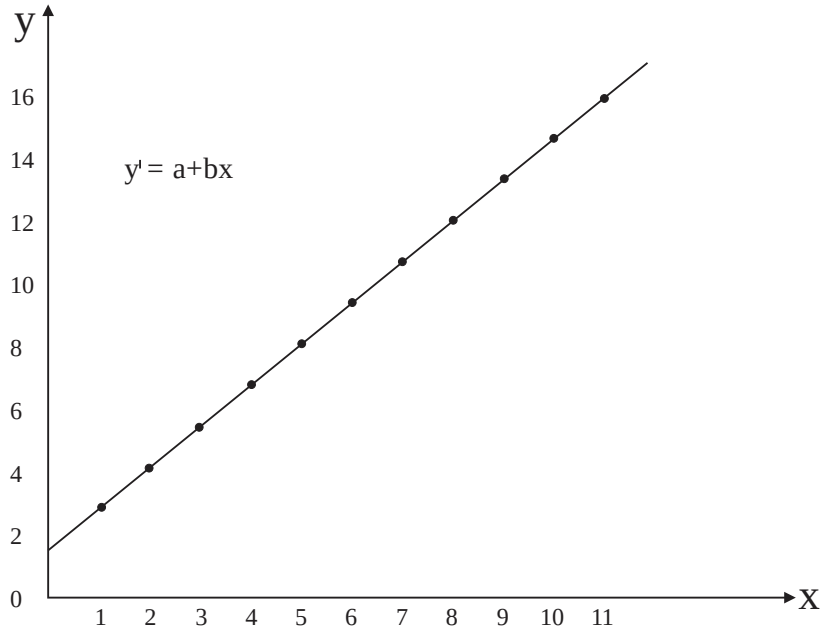


Figure 10.3: Graph of X and Y' showing the intercept (a)

10.4.2 Test of significance

$$SS_{total} = \Sigma y^2 = \Sigma(y - \bar{y})^2$$

$$SS_{regression} = b\Sigma xy$$

$$SS_{residual} = \Sigma(y - y')^2$$

or

$$SS_{residual} = SS_{total} - SS_{regression}$$

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

so $R = \sqrt{\frac{SS_{regression}}{SS_{total}}}$

$$F = \frac{MS_{regression}}{MS_{residual}}$$

but, $MS_{regression} = \frac{SS_{regression}}{k-1}$

$$MS_{residual} = \frac{SS_{residual}}{n-k}$$

$$t = \frac{R}{\sqrt{1-R^2}} \times \sqrt{n+2}$$

From results in example 10.1

$SS_{total} = 38.833$

$SS_{regression} = 1.1714(20.5001)$

$= 24.014$

$SS_{residual} = 14.819$

$R^2 = \frac{24.014}{38.833}$

$= 0.6184$

$R = \sqrt{0.6184}$

$= 0.7864$

$$MS_{regression} = \frac{24.014}{2-1} = 24.014$$

$$MS_{residual} = \frac{14.819}{6-2} = \frac{14.819}{4}$$

$= 3.705$

$$F = \frac{24.014}{3.705}$$

$= 6.482$

$$t = \frac{0.7864}{\sqrt{1-0.6184}} \times \sqrt{6-2}$$

$$= \frac{0.7864}{\sqrt{0.3816}} \times 2$$

$$= \frac{1.5728}{0.6177}$$

$= 2.546$

Note that, test of significance for regression analysis is normally reported with the F-ratio and t-value(s).

10.6: MULTIPLE REGRESSIONS

Multiple regression analysis is appropriate when the research problem involves analyzing the relationship between a single

dependent variable and two or more independent variables. To apply multiple regressions, the dependent and independent variables must be continuously measured. The objective of multiple regression analysis is to predict a dependent variable using multiple independent variables. The decision on which variable is dependent and which is independent is that of the researcher.

A multiple regression equation is given as:

$Y' = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$, and

$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_n\bar{x}_n$

Where: Y' = The predicted Y

a = the intercept

$b_1, b_2, \cdots, b_n$ = The regression coefficients for variables 1, 2 up to nth term.

$x_1, x_2, \cdots, x_n$ = independent variables 1, 2, up to the nth term.

10.7: DERIVATIVE OF THE REGRESSION COEFFICIENTS ($b_1, b_2, \cdots, b_n$)

Assume that the number of variables range from variable 1 up to variable n. Then, the multiple regression equation will be:

$$Y' = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad (1)$$

Similarly, $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_n x_n$ (2)

substitute equation (2) in equation (1)

i.e $Y' = \bar{Y} + b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) + \cdots + b_n(X_n - \bar{X}_n)$

If $SS_{total} = SS_{regression} + SS_{residual}$

i.e $SS_{total} = \Sigma(Y' - \bar{Y})^2 + \Sigma(Y - Y')^2$

then, to minimize $\Sigma(Y - Y')^2$, the value of 'b' must satisfy the following equations:

$$b_1X_1^2 + b_1\Sigma X_1X_2 + \cdots + b_n\Sigma X_1X_n = \Sigma X_1Y$$

$$b_1X_1X_2 + b_2\Sigma X_2^2 + \cdots + b_n\Sigma X_1X_n = \Sigma X_2Y$$

$$b_1X_1X_2 + b_2\Sigma X_nX_2 + \cdots + b_n\Sigma X_n^2 = \Sigma X_nY \quad (3)$$

If the independent variables are two, we will likely have two equations with two unknowns.

i.e., $b_1\Sigma X_1^2 + b_1\Sigma X_1X_2 = \Sigma X_1Y$, and (4)

$$b_1\Sigma X_1X_2 + b_2\Sigma X_2^2 = \Sigma X_2Y \quad (5)$$

Using algebraic elimination method multiply (4) by ΣX_2^2 and (6) by ΣX_1X_2

Thus; $b_1(\Sigma X_1^2)(\Sigma X_2^2) + b_2(\Sigma X_1 X_2)(\Sigma X_2^2) = (\Sigma X_1 Y)(\Sigma X_2^2)$, and (6)

$$b_1(\Sigma X_2 X_1)^2 + b_2(\Sigma X_1 X_2)(\Sigma X_2^2) = (\Sigma X_2 Y)(\Sigma X_1 X_2) \quad (7)$$

Subtracting (7) from (6), we have

$$b_1(\Sigma X_1^2)(\Sigma X_2^2) - b_1(\Sigma X_2 X_1)^2 = (\Sigma X_1 Y)(\Sigma X_2^2) - (\Sigma X_2 Y)(\Sigma X_1 X_2)$$

i.e., $b_1 = \dfrac{(\Sigma X_1 Y)(\Sigma X_2^2) - (\Sigma X_2 Y)(\Sigma X_1 X_2)}{(\Sigma X_1^2)(\Sigma X_2^2) - (\Sigma X_2 X_1)^2}$

Similarly, $b_2 = \dfrac{(\Sigma X_2 Y)(\Sigma X_1^2) - (\Sigma X_1 Y)(\Sigma X_1 X_2)}{(\Sigma X_1^2)(\Sigma X_2^2) - (\Sigma X_1 X_2)^2}$

Users of this text should note that, detailed explanations to these derivatives are not adequate for this text. Those wanting detailed explanations may consult Edwards (1979).

EXAMPLE 10.2

A teacher was interested in finding out whether two students' variables, Attitude (X_1) and Intelligence (X_2) significantly predict their academic achievement in a test. The following were the raw scores:

Student	1	2	3	4	5	6
X_1	4	5	6	8	5	3
X_2	3	2	2	4	3	4
Y	8	6	3	4	6	3

Compute all the required statistics and test for significance

Solution

X_1	X_2	Y	x_1	x_2	Y	x_1^2	x_2^2	y^2	x_1x_2	x_1y	x_2y	y'	$(y-y')^2$
4	3	8	-1.17	0	3	1.37	0	9	0	-3.51	0	5.32	7.1824
5	2	6	-0.17	-1	1	0.03	1	1	0.17	-0.17	-1	5.55	0.2062
6	2	3	0.83	-1	-2	0.69	1	4	-0.83	-1.66	2	5.28	5.1797
8	4	4	2.83	1	-1	8.01	1	1	2.83	-2.83	-1	3.74	0.0697
5	3	6	-0.17	0	1	0.03	0	1	0	-0.17	0	5.05	0.9103
3	4	3	-2.17	1	-2	4.71	1	4	-2.17	4.34	-2	5.09	4.3510
31	**18**	**30**				**14.84**	**4**	**20**		**-4**	**-2**		**17.9214**

$\bar{X}_1 = 5.17$,

$\bar{X}_2 = 3$,

$\bar{Y} = 5$

$$b_1 = \frac{(-4)(4)-(2)(0)}{(14.84)(4)-(0)^2}$$

$$\frac{-16}{59.36}$$

$$\cong -0.27$$

$$b_2 = \frac{(-2)(14.84)-(4)(0)}{(14.84)(4)-(0)^2}$$

$$\frac{-29.68}{59.36}$$

$$= -0.50$$

$$a = 5 - (-0.27)(5.17) - (-0.50)(3)$$

$$= 5 + 1.3959 + 1.5$$

$$= 7.8959$$

$$\cong 7.90$$

$$SS_{regression} = (-0.27)(-4) + (-0.5)(-2)$$

$$= 1.08 + 1$$

$$= 2.08$$

$$SS_{residual} = 17.921$$

$$SS_{total} = 20$$

$$R^2 = \frac{2.08}{20} = 0.104$$

$$R = \sqrt{0.104} = 0.322$$

$$1 - R^2 = 1 - 0.104$$

$$= 0.896$$

$$MS_{regression} = \frac{2.08}{2} = 1.04$$

$$MS_{residual} = \frac{17.921}{4} = 4.480$$

$$F = \frac{1.04}{4.480} = 0.232$$

Test of significance

$$t_1 = \frac{b_1}{\sqrt{\frac{MS_{residual}}{\Sigma x_1^2 (1 - R^2)}}}$$

$$= \frac{-0.27}{\sqrt{\frac{5.9737}{14.84 \times 0.896}}}$$

$$= \frac{-0.27}{0.67025}$$

$$= -0.4028$$

Result: Calculated t = -0.4028

Critical t-value at 0.05 with 4 degrees of freedom = 2.776. Based on this result, we conclude that attitude did not significantly predict students' performance.

$$t_2 = \frac{b_2}{\sqrt{\frac{MS_{residual}}{\Sigma x_2^2 (1 - R^2)}}}$$

$$= \frac{-0.5}{\sqrt{\frac{5.9737}{4 \times 0.896}}}$$

$$= \frac{-0.5}{1.291}$$

$$= -0.387$$

Result: Calculated t-value = -0.387

Critical t-value at 0.05 with 4 degrees of freedom = 2.776. Based on this result, we conclude that intelligence did not, as well, significantly predict students' performance.

10.8 EXERCISE TEN

1. The following are scores obtained by ten persons in two examinations. Assume that scores on X were for JAMB-UTME while scores on Y were for Post-UTME from a particular university.

person	1	2	3	4	5	6	7	8	9	10
X	49	53	64	51	72	43	37	50	57	51
Y	55	56	50	62	48	46	52	48	55	50

i) Estimate the intercept 'a' and slop 'b' for the regression equation.

ii) Fit a regression equation to predict performance in Post-UTME using JAMB-UTME.

iii) Estimate Y' for all the ten persons

iv) Calculate the correlation coefficient (R), coefficient of determination (R^2), and $1-R^2$

v) Test for significance at 0.05 level to say whether or not JAMB-UTME predicted performance in Post-UTME of that University among those candidates.

2. (a) In the regression line $Y' = a+b_1x_1+b_2x_2+\ldots+b_nx_n$ prove that:

i. $b_1 = \frac{(\sum X_1Y)(\sum X_2^2)-(\sum X_2Y)\sum X_1X_2)}{(\sum X_1^2)(\sum X_2^2)-(\sum X_2X_1)^2}$

ii. $b_2 = \frac{(\sum X_2Y)(\sum X_1^2)-(\sum X_1Y)\sum X_1X_2)}{(\sum X_1^2)(\sum X_2^2)-(\sum X_1X_2)^2}$

(b) Show that $\sum x^2 = b\sum xy + \sum(Y-y')^2$ in a linear regression model.

3. List the four assumptions of multiple regressions and write briefly but exhaustively on each of them.

4. The following regression results were obtained for a study:

Variable	Coefficient	Standard error
Intercept	0.5292	0.2712
Attitude	0.1810	0.1412
Anxiety	0.8827	0.0708
$R^2 = 0.9942$		

i. Write out the regression function if the independent variable is student's academic performance in an end of semester examination.
ii. Estimate the t-statistic for the intercept, attitude and anxiety
iii. Using the rule of thumb, are they significant?

REFERENCES

Adegoke, B. A. (2014*). Statistical methods for behavioural and social science research*. Ibadan: Everlasting Printing Ventures.

Afonja, B. (2001). *Introductory statistics: A learner's activated approach.* Ibadan: Evans Brothers (Nigerian Publisher) Limited.

Anikweze, C. M. (2013). Statistical analysis of research data. Workshop proceedings by the Nationaational Mathematical Centre, Abuja 11th-15th November, 2013.

Ary, D., Jacobs, L. C. & Razavieh, A. (1985). *Introduction to research in education.* New York: CBS College Publishing.

Edwards, A. L. (1979*). Multiple regression and the analysis of variance and covariance*. San Francisco: W. H. Freeman and Company.

Gupta, S. C. (2013*). Fundamentals of statistics*. Delhi: Himalaya Publishing House.

Joshua, M. T. (2008). *Fundamentals of texts and measurement in education.* Calabar: University of Calabar Press.

Kothari, C. R. (2011). *Research methodology; methods and techniques.* New Delhi: New Age International Limited Publishers.

Koul, L. (2012). *Methodology of educational research.* Jangpura: Vikas Publishing House.

Lucy, T. (2002). Quantitative techniques. China: C& C Offset Printing

Osborne, J. W. & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation Journal*, 8(2), 1-5.

Spiegel, M. R. & Stephens, L. J. (2011). Statistics. New York: MCGraw-Hill Publishing Company.

Ubi, I. O. & Bassey, B. A. (2012). Multivariate statistics. In Abang J. Isangedighi, *Essentials of Research and Statistics in Education and Social Sciences*.

APPENDIX

Critical values of t

Degrees of Freedom	Level of significance for one-tailed test					
	.01	.05	.025	.01	.005	.0005
	Level of significance for two-tailed test					
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	1.110
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.449	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.973	1.833	2.262	2.821	3.258	4.781
10	1.382	1.812	2.228	2.764	3.169	1.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.937	4.073
16	1.337	1.746	2.120	2.583	2.921	4.051
17	1.333	1.740	2.110	2.567	2.898	5.965
18	1.328	1.734	2.093	2.552	2.878	5.922
19	1.328	1.729	2.093	2.539	2.861	5.883
20	1.325	1.725	2.086	2.528	2.831	2.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.767
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.319	1.711	2.064	2.492	2.797	3.745
25	1.317	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.282	1.645	1.960	2.426	2.576	3.291

Source: Denga, D. I & Ali, A. (1998). An introduction to research methods and statistics in education and social sciences. Calabar: Rapid Education Publishers Limited.

Critical Values of the Pearson Correlation Coefficient

Degrees of Freedom	Level of significance for one-tailed test			
	.05	.025	.01	.005
	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.576	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.582	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.496
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.486
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.303
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254`

Source: Ary, D., Jacobs, L. C. & Razavieh, A. (1985). Introduction to research in education. New York: CBS College Publishing.

Critical values of F

.05 (roman) and .01 (bold face) levels of significance

	Degrees of freedom for greater mean square									
	1	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	
1	161.45 **4052.10**	199.50 **4999.03**	215.72 **5403.49**	224.57 **5625.14**	230.17 **5764.08**	233.97 **5859.39**	238.89 **5981.34**	243.91 **6105.83**	249.04 **6234.16**	254.32 **6366.48**
2	18.51 **98.49**	19.00 **99.01**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.37 **99.36**	19.41 **99.42**	19.45 **99.46**	19.50 **99.50**
3	10.13 **34.12**	9.55 **30.81**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.84 **27.49**	8.74 **27.05**	8.64 **26.60**	8.53 **26.12**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.04 **14.80**	5.91 **14.37**	5.77 **13.93**	5.63 **13.46**
5	6.61 **12.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.82 **10.27**	4.68 **9.89**	4.53 **9.47**	4.36 **9.02**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.15 **8.10**	4.00 **7.72**	3.84 **7.31**	3.67 **6.88**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.73 **6.84**	3.57 **6.47**	3.41 **6.07**	3.23 **5.65**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.41 **6.03**	3.28 **5.67**	3.12 **5.28**	2.93 **4.86**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.23 **5.47**	3.07 **5.11**	2.90 **4.73**	2.71 **4.31**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.07 **5.06**	2.91 **4.71**	2.74 **4.33**	2.54 **3.91**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	2.95 **4.74**	2.79 **4.40**	2.61 **4.02**	2.40 **3.60**
12	4.75 **9.33**	3.38 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.85 **4.50**	2.69 **4.16**	2.50 **3.78**	2.30 **3.36**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.77 **4.30**	2.60 **3.96**	2.42 **3.59**	2.21 **3.16**
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.70 **4.14**	2.53 **3.80**	2.35 **3.43**	2.13 **3.00**
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.64 **4.00**	2.48 **3.67**	2.29 **3.29**	2.07 **2.87**
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.59 **3.89**	2.42 **3.55**	2.24 **3.18**	2.01 **2.75**
17	4.45 **8.40**	3.59 **6.11**	3.20 **5.18**	2.96 **4.67**	2.81 **4.34**	2.70 **4.10**	2.55 **3.79**	2.38 **3.45**	2.19 **3.08**	1.96 **2.65**
18	4.41 **8.28**	3.55 **6.01**	3.16 **5.09**	2.93 **4.58**	2.77 **4.25**	2.66 **4.01**	2.51 **3.71**	2.34 **3.37**	2.15 **3.01**	1.92 **2.57**
19	4.38 **8.18**	3.52 **5.93**	3.13 **5.01**	2.90 **4.50**	2.74 **4.17**	2.63 **3.94**	2.48 **3.63**	2.31 **3.30**	2.11 **2.92**	1.88 **2.49**
20	4.35 **8.10**	3.49 **5.85**	3.10 **4.94**	2.87 **4.43**	2.71 **4.10**	2.60 **3.87**	2.45 **3.56**	2.28 **3.23**	2.08 **2.86**	1.84 **2.42**
21	4.32 **8.02**	3.47 **5.78**	3.07 **4.87**	2.84 **4.37**	2.68 **4.04**	2.57 **3.81**	2.42 **3.51**	2.25 **3.17**	2.05 **2.80**	1.81 **2.42**
22	4.30 **7.94**	3.44 **5.72**	3.05 **4.82**	2.82 **4.31**	2.66 **3.99**	2.55 **3.75**	2.40 **3.45**	2.23 **3.12**	2.03 **2.75**	1.78 **2.30**
23	4.28 **7.88**	3.42 **5.66**	3.03 **4.76**	2.80 **4.46**	2.64 **3.94**	2.53 **3.71**	2.38 **3.41**	2.20 **3.07**	2.00 **2.70**	1.76 **2.26**

Critical values of F

.05 (roman) and .01 (bold face) levels of significance

	Degrees of freedom for greater mean square									
	1	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.36**	**3.93**	**2.66**	**4.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.32**	**2.99**	**2.62**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.29**	**2.96**	**2.58**	**2.13**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.78**	**3.56**	**3.26**	**2.93**	**2.55**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1. 91	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.75**	**3.53**	**3.23**	**2.90**	**2.52**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.20**	**2.87**	**2.49**	**2.08**
30	4.17	3.32	2,92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.17**	**2.84**	**2.47**	**2.01**
35	4.12	3.26	2.87	2.64	2.48	2.37	2.22	2.04	1.83	1.57
	7.42	**5.27**	**4.40**	**3.91**	**3.59**	**3.37**	**3.07**	**2.74**	**2.27**	**1.90**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.52
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**2.99**	**2.66**	**2.29**	**1.82**
45	4.06	3.21	2.81	2.58	2.42	2.31	2.15	1.97	1.76	1.48
	7.23	**5.11**	**4.25**	**3.77**	**3.45**	**3.23**	**2.94**	**2.61**	**2.23**	**1.75**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	1.95	1.74	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.19**	**2.89**	**2.56**	**2.18**	**1.68**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.82**	**2.50**	**2.12**	**1.60**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.07	1.89	1. 67	1.35
	7.01	**4.92**	**4.07**	**3.60**	**3.29**	**3.07**	**2.78**	**2.45**	**2.07**	**1.53**
80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.88	1.65	1.31
	6.96	**4.88**	**4.04**	**3.56**	**3.26**	**3.04**	**2.74**	**2.42**	**2.03**	**1.47**
90	3.95	3.10	2.71	2.47	2.32	2.20	2.04	1.86	1.64	1.28
	6.92	**4.85**	**4.01**	**3.53**	**3.23**	**3.01**	**2.72**	**2.39**	**2.00**	**1.43**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.03	1.85	1.63	1.26
	6.90	**4.82**	**3.98**	**3.51**	**3.21**	**2.99**	**2.69**	**2.37**	**1.98**	**1.39**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.01	1.83	1.60	1.21
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.66**	**2.33**	**1.94**	**1.32**
150	3.90	3.06	2.66	2.43	2.27	2.16	2.00	1.82	1.59	1.18
	6.81	**4.75**	**3.91**	**3.45**	**3.14**	**2.92**	**2.63**	**2.31**	**1.92**	**1.27**
200	3.89	3.04	2.65	2.42	2.26	2.14	1.98	1.80	1.57	1.14
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.89**	**2.60**	**2.28**	**1.88**	**1.21**
300	3.87	3.03	2.64	2.41	2.25	2.13	1.97	1.79	1.55	1.10
	6.72	**4.68**	**3.85**	**3.38**	**3.08**	**2.86**	**2.57**	**2.24**	**1.85**	**1.14**
400	3.86	3.02	2.63	2.40	2.24	2.12	1.96	1.78	1.54	1.07
	6.70	**4.66**	**3.83**	**3.37**	**3.06**	**2.85**	**2.56**	**2.23**	**1.84**	**1.11**
500	3.86	3.01	2.62	2.39	2.23	2.11	1.96	1.77	1.54	1.06
	6.69	**4.65**	**3.82**	**3.36**	**3.05**	**2.84**	**2.55**	**2.22**	**1.83**	**1.08**
1000	3.85	3.00	2.61	2.38	2.22	2.10	1.95	2.76	1.53	1.03
	6.66	**4.63**	**3.80**	**3.34**	**3.04**	**2.82**	**2.53**	**2.20**	**1.81**	**1.04**
for smaller	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	
mean square	**6.64**	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.51**	**2.18**	**1.79**	

Source: Koul, L. (2012). Methodology of educational research. New Delhi: Vikas Publishing House.

Critical values of X^2 distribution

TABLE E: χ^2 Table, P gives the probability of exceeding the tabulated value of χ^2 for the specified number of degrees of freedom (df). The values of χ^2 are printed in the body of the table

df	*0.95*	*0.90*	*0.80*	*0.70*	*0.50*	*0.30*	*0.20*	*0.10*	*0.05*	*0.02*	*0.01*
1	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.083	16.812
7	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.663	32.000
17	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.760	27.587	30.995	33.409
18	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.869	28.869	32.346	34.805
19	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	10.851	12.443	15.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	28.885	42.856	45.642
27	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	19.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45. 9	48.278
29	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.69	49.588
30	18.493	20.599	23.364	25.508	29.386	33.530	36.250	40.256	43.773	47.962	50.892

Source:Koul, L. (2012). Methodology of educational research. New Delhi: Vikas Publishing House.

SUBJECT INDEX

NAME INDEX

View publication stats