DEMYSTIFIED STATISTICS

FOR

EDUCATION

AND

BEHAVIORAL SCIENCES

Omachi, Daniel. ***Ph.D. California***

First Edition, Published in 2021 by
Angusco Nigeria Enterprise
No. 37 Edinburgh Road,
Ogui New Layout, Enugu State
Nigeria

Dedication

This book is dedicated to my wife Mrs. Grace Omachi Daniel

and

children: Elyon Omachi Daniel

Stainless Omachi, Daniel

Eliora Omachi Daniel

ACKNOWLEDGEMENTS

I wish to acknowledge first, all my lovely students in all universities and colleges whose encouragement informs this edition of the book. I also acknowledge my dear Vice Chancellor Godfrey Okoye University, Very Rev. Fr. Professor Christian Anieke for his love for education. Thank you sir for all your encouragement, supports and advice. A very big thank you to other management and staff members of Peaceland College of Education, Enugu. Many thanks to the management and staff of Elyland College of Education, Ankpa for their encouragement and support. A special thanks to Mr. Arome Alfa, Mr. Raluchukwu Okwudili, Mr. Nwadike Bornaventure, Mr. James Monday and others.

May I thank in a very special way my academic giants and worriors whose shoulder I stand. Prof. Romey Okoye of Nnamdi Azikiwe University, Prof. Ngozi Agu, Prof. Kan Nwankwo, Prof. Esomuonu, and Prof. Sunday Abonyi for their wealth of knowledge and what they have made out of the author. God bless you all.

I appreciate greatly the understanding of my dear wife Mrs. Grace Omachi Daniel, my children Elyon, Stainless and Eliora. My Dad Dr. Adams Baba Daniel and all my siblings. I appreciate all your efforts and understanding that led to the success of this work.

To all the people worthy of this acknowledgement but not mentioned, I tell you, you are the best and I appreciate your prayers, encouragement and supports. God bless you all.

PREFACE

Demystified Statistics for Education and Behavioral Sciences is concerned with application of statistical methods in analysis of data. It is no longer news that students in education and other behavioral sciences approach statistics with anxiety and looks at it as a mountain of science and figures that cannot be climbed. To alleviate some of the negative feelings of students in these areas towards statistics, this book have demystified and presented statistical analysis in a very simple way with clear language devoid of technicalities.

The goal of this book is to develop in students the understanding of assumptions of different statistical procedures. The students will learn how to apply statistics in analysis of data collected for any research work of interest. The students will learn how to read and understand statistical researches. The students will also be made to drop the feeling of trepidation towards statistics and even develop an interest in it. The book is divided into eight major chapters:

I. **The Nature and Concept of Statistics:** Meaning of statistics and significance; statistical variables; and scales of statistical measurements are simply presented in this chapter. These make the students to have a clear understanding of statistical concepts .and types of data for statistical applications.

2. **Presentation and Organization of Data:** Different ways of organizing data for statistical analysis are conveniently presented here. This covers the arrangement of data in both frequency distributions and in graphs.

3. **Descriptive Statistics:** Statistical procedures used in describing and summarizing data are presented in a step-by-step format.

4. **Measures of Relationship (Association) or Correlation Analysis:** A description of correlation analysis and its uses are presented in this chapter. Different correlation coefficients for different types of variables are' described with simple illustrations.

5. **Regression Analysis:** This book attempts to describe the concept, and procedures of regression analysis in a simplified manner. Experience has shown that very few teacher education students attempt to use regression analysis in their research work. Going through the simplified presentation of regression analysis, it is hoped that more students will attempt to venture into predictions, particularly for variables whose relationships are linear.

6. **Hypothesis Testing.** This book presents what hypothesis testing is and how to formulate null and alternative hypotheses. Students are made to know what is meant by level of significance, one-tailed and two-tailed tests, degrees of freedom, and their uses in hypothesis testing.

7. **Inferential Statistics and Tests:** Assumptions of inferential statistics; assumptions and computational procedures of different parametric statistics are presented in a simplified way. This is very vital for students who venture into hypotheses testing in their research works. Conscientious study of this chapter will alleviate most of the confusion students have in applying parametric statistics in their research.
8. **Non-Parametric Statistics and Tests:** Alternative procedures for testing hypotheses which do not necessitate restrictive assumptions about the distributions of the population are presented in this chapter. Students are either unaware of their existence or just ignorant of their procedures as they are rarely used in students' data analysis. Students, however, attempt to use only one out of the non-parametric procedures, which is the chi-square (x^2)

Each chapter in the book contains exercises which are meant to help the students test the level of their understanding of each concept and procedure.

Probability theory was not covered in this book. This does not mean that probability theory is not important to students in Education. This book concentrated on those statistical procedures which the students frequently use in their research work.

CONTENTS

CHAPTER ONE

THE NATURE AND CONCEPT OF STATISTICS

The Purpose of this Book

a. To enable you read and understand professional literature with symbols, concepts and ideas of statistics. For instance, when it is expressed that the class mode in examination is '15' or that the standard deviation of a set of scores is '3.6' or the meaning and significance of N, t test and T score, you will easily understand what it means.
b. To prepare you for advanced course and study, especially if one has to collect, analyze and present data from projects, researches or other personal works.
c. As a part of your professional training as a teacher or otherwise, basic statistics helps you to appreciate, and evaluate the performance of the children, or to standardize raw marks. For instance, when you rank students, after a test, you will be able to know how the 20^{th} position in one class compares with the 10^{th} position in another class,

Objectives of the Book

You are expected at the end of this book to:

i) Be familiar with and recognize the methodology and typology used in educational statistics.
ii) Compare mean, mode and standard deviation etc.
iii) Convert raw data into standard form.
iv) Organize data for meaningful presentation.
v) Compute correlation coefficient by product moment or rank order method.
vi) Test hypothesis using inferential statistical procedures.

Even though mathematics is involved in educational statistics, the emphasis in this book is not on mathematical calculations but on the understanding of the basic concepts and their applications. The emphasis is on the total comprehension of the procedures and mechanics of operations of statistics to ameliorate phobia.

What is Statistics?

Statistics may be defined as a science that studies numerical values assigned to objects, attributes or variables in order to characterize and make the variables more meaningful. It is, therefore, the science of classifying, organizing, and analyzing data (King & Minium, 2003). Usually, in statistics, information is presented not in words but in numerical form- numbers.

Statistics deals with populations and samples. The average or mean or other descriptions of the population are referred to as parameters but the descriptions of the sample (such as the mean or standard deviation) are referred to as statistics.

The subject, statistics may be divided into two broad areas:

a). Descriptive Statistics:

This refers to the statistical procedures that are used to describe and summarize data for easy comprehension. Examples of descriptive statistics are the mean, standard deviation, range etc which describe and give you information about a sample or population.

b). Inferential Statistics

This area of statistics is often based on the information derived from descriptive statistics. It is often referred to as analytical statistics and enables us draw conclusions or make informed statistical inferences from a mass of data. The purpose of inferential statistics is to draw a conclusion about conditions that exist in a population from study of a sample. Therefore, a researcher uses it when he bases his conclusions of a research on limited information (information from 1 sample of a population). Examples of inferential statistics are: t-test, ANOVA, Chi square etc. Inferential statistics starts with formulation of hypothesis - null hypothesis or alternative hypothesis, which in facts are guesses. At the end of the analysis, we confirm or reject the hypothesis.

Significance of Statistics

Statistics to Mendenhall (1993, p.3) is about describing and predicting. It is a response to the limitations of our ability to know. Any complex aspect of human knowledge is either difficult for us to understand, or lends itself to misunderstanding - unless we possess the right tool with which to examine and analyze its complexity.

Statistics is that tool.

Statistics is basic to the world of manufacturing, social sciences, business, politics, pure sciences, behavioral sciences, and research in all spheres of life because:

i. It provides the necessary exact descriptions.
ii. It forces researchers by its methods to follow defined procedures in their work and thinking.
iii. Its method provides a means of summarizing results of research in meaningful and convenient form.
iv. It enables researchers to make predictions based on the data available at hand.
v. It provides easy means of analyzing cause(s) of complex events.

Variables in Statistics

As stated earlier, in statistics, we use numbers to describe, measure or give other information e.g. about performance of students, intelligence level of students, height of class pupils etc. The characteristics being measured is referred to as **"variables"**. If we are studying the scores of students in mathematics, the scores become the variable. Sometimes we study the relationship between two variables. One of the two or more variables is called independent variable, while the other variable is referred to as dependent variable. Suppose we are interested in school children's age and their scores in mathematics, 'age' is the independent variable while 'score' is the dependent variable. The dependent variable is one which is affected or which changes (varies) as changes occur in the other variable. In this our case, scores change as age changes. The independent variable (age) is that which is more or less not affected by changes in the other variable. In statistics, variables are also regarded as either continuous or discontinuous (discrete) variables.

Continuous – Variables

These are variables that can take on infinite number of values within a range and may not necessarily be determined by mere counting. For example, given the range of scores 60 to 70, there is every possibility that some students may have scores between two scores like 60.5, 61.2, 62.4, 62.8, 64.7 etc. The fact that students' score are rounded up to whole numbers does not remove the possibility of occurrence of fractional scores. Actually, variables that are not seen with naked eyes and cannot be touched but are innate like psychological variables tend to be continuous and educational variables belong to this group.

Discontinuous or Discrete Variables

These are variables which take on finite number of values within a range. These refer to data that can be fully counted and are not fractional. Discontinuous variables refer to variables that are tangible and can be measured directly.

Variate in Educational Statistics

A variate is the single value of a variable. This means that if a variable of interest is test scores of students in statistics, a variate refers to one single score of a student in the distribution. While analyzing any data, we deal with a number of variates which we may refer to as raw scores or data.

Scales of Statistical Measurement

Once more, it is emphasized that variables are expressed in numerical form. The process used in determining or assigning these numbers informs the interpretation that can be made from them and the statistical procedures that can be used meaningfully. This means that during the process of data collection, different instruments are used and each instrument enables the researcher to collect a particular type of data. Each set of data collected lends itself to a particular statistical analysis. The most important aspect of application of statistics to research data is choice of appropriate statistical procedure. Inappropriate statistical procedure will lead to inaccurate analysis and interpretation. To ensure appropriate application of these statistical procedures, data are classified into different scales of measurement wherein we put different types of data into different categories. There are basically the following scales of measurement, starting with the least rated:

Nominal Scale

This is the lowest scale of measurement and involves placing objects or individuals into categories which are qualitative rather than quantitative. For example, you may categorize individuals into male or female; people you meet in the university can be divided into students or lecturers; or success in an examination as either pass or fail. There is no order of magnitude involved in this categorization. In filling forms sometimes you tick "1" for male and "2" for female. These numbers do not indicate any magnitude but are used for convenience; in fact they are arbitrarily assigned and can be changed at will. The numbers cannot be added, subtracted, divided or subjected to any mathematical treatment.

Ordinal Scale

The ordinal scale shows you that things are different but does not signify the direction or degree of the difference. It indicates the order of magnitude or rank e.g. 1^{st}, 2^{nd}, 3^{rd}, 4^{th}. In other words, ordinal scale categorizes variables in order of magnitude. It does not the exact interval or difference between the categories. When we rank students (without their marks) or when we use grades like grade 1, grade 2, etc, in WAEC examinations we are using ordinal scale. We shall use ordinal scale in correlation studies.

Interval Scale

In this scale, the exact difference or interval between categories is specified. Suppose we have four students with scores A, 90%; B, 80%, C, 70%; D. 60%, by ordinal scale, A is first, B is second, C is third, D is fourth. When we state their scores in percentages, we are using interval scale. Interval scales can be subjected to multiplication, addition or subtraction. For instance, we can subtract 10 marks from each student. Interval scale provides order as well as difference between the intervals. To get the difference, we subtract one category from another.

However, interval scale does not have absolute or genuine zero. 0% does not mean absolute nothingness or complete non existence of the attribute. Interval scales feature when we determine means, compute standard deviation or F – ratio.

Ratio Scale

Ratio scale is the highest scale of measurement. This scale has absolute zero. Ratio scale is used more in the physical sciences than in behavioural sciences. For instance, when we measure temperature in degrees Kelvin, we are certain of the real meaning of zero degree on the Kelvin scale i.e. the temperature at which water freezes. Other examples of ratio scale measurements are length, weight, and measures of elapsed time.

Exercises

1. What is statistics? Demarcate between descriptive and inferential statistics.
2. Using specific illustrations, explain the meaning of variables and variates.

3. Describe different scales of measurement in statistical analyses highlighting the differences between them.

CHAPTER TWO

PRESENTATION AND ORGANIZATION OF DATA

Data such as scores which are collected and recorded in the way they occur without any order or arrangements or processing are called raw data. The table below represents unorganized raw scores of 50 students.

Table 1: ***Raw Scores of 50 Students in Geography***

25	25	29	26	27	22	24	38	39	32	34	44	33	51	41
25	21	28	14	33	33	15	27	36	20	55	16	33	47	16
15	27	42	37	10	11	29	21	18	28	46	19	21	36	17
46	40	34	27	29										

Before we can meaningfully analyze any data like the raw scores above, the data has to be carefully organized. One common way of organizing data such as in the table above is to arrange them in frequency distribution form and present them in graphic form as the case may be.

Frequency Distributions

Frequency distribution is a systematic arrangement of data (scores) from the lowest to the highest taking into consideration the number of occurrence of each datum. The process involved is simply writing out the scores and as we come across that score in our distribution we put a stroke. When four strokes have been made, a fifth stroke is used to cross the four strokes (~~////~~). A crossed bundle is thus five strokes. The frequencies are the sum of the strokes for the range of scores.

Table 2: Frequency Distribution of Scores

Record (x)	**Tallies**	**Frequency**
10 – 14	///	3
15 – 19	~~////~~ //	7
20 – 24	~~////~~ /	6
25 – 29	~~////~~ ~~////~~ //	12
30 – 34	~~////~~ //	7
35 – 39	~~////~~ /	6

40 – 44	////	4
45 – 49	///	3
50 – 54	/	1
55 – 59	/	1

We started with 50 raw scores of 50 students. When at the end we sum up the frequencies they make up 50. In statistics, N usually represents the number of scores or categories while f represents frequency. Thus we may write $\sum f = N$.

Grouped and Ungrouped Data

When scores are less than thirty, we can just write down the number and carry out the tallying. For numbers more than thirty e.g. 100, tallying without grouping the scores would be difficult. This leads to what is called grouped and ungrouped data or frequency distribution.

Ungrouped Frequency Distribution

This refers to systematic arrangement of scores from highest to lowest in such a way that each individual score is listed and the frequency recorded.

Suppose we have the following distribution:

20, 22, 25, 20, 15, 14, 24, 21

26, 27, 20, 26, 16, 15, 15, 21

We can put it in a frequency distribution table as presented in Table 3. Scores or data are presented in ungrouped form when they are less than thirty and when the range of the data is small.

Table 3: ***Ungrouped Frequency Distribution of Scores***

Scores (x)	**Tally**	**Frequency**
27	/	1
26	///	3
25	///	3
24	/	1
22	//	2
21	//	2
20	~~////~~	5
16	//	2
15	///	3
14	//	2

Grouped Frequency Distribution

When the raw scores obtained are more than thirty, there is a need to group them for convenience. Given the 50 raw scores of 50 students in Table 2 above, there is need to group them as they are more than 30. To group such data, the following steps will be taken.

1. **Determine the Class Interval**

   This means the number of groups or classes one wants to have in a distribution. Statistically, class interval is mostly denoted with letter 'K'. The decision on this class interval depends on the researcher which he takes considering the number of scores he has. For our data in Table 1, we will take class interval to be 10. K

2. **Compute the class or interval width or size:**

   Class (interval) width (size) refers to the number of scores which will be in each class. This is determined by dividing the range of scores with the class interval. Class width is statistically denoted with small letter 'c' or 'i'. The formula is:

   $$c = \frac{R}{K}$$

   Where R = Range, K = Class interval

   For the data in Table 1 above, class width is: $\frac{55 - 10}{10} = 4.5 = 5$

NOTE: Always approximate class width to the next higher whole number. Also prefer an odd numbered class width because it will reduce the tediousness of working with fractions later in the statistical analysis of grouped data. Where you have an even numbered class width, adjust the class interval until it yields an odd numbered class width.

3. **Group the Data Starting from Lowest Score**

   Ensure that each class has a class width of 5 each. This should be constant in all classes. Then determine the frequencies by using the tally column as in table 4 below.

Table 4: Grouped Frequency Distribution of Scores

Record (x)	Tallies	Frequency
10 – 14	///	3
15 – 19	~~////~~ //	7
20 – 24	~~////~~ /	6
25 – 29	~~////~~ ~~////~~ //	12
30 – 34	~~////~~ //	7
35 – 39	~~////~~ /	6
40 – 44	////	4
45 – 49	///	3
50 – 54	/	1
55 – 59	/	1

Cumulative Frequency (cf) Distribution

After data has been put in frequency distribution, there is need to determine the cumulative frequency distribution of such data. Cumulative frequency is the progressive summation of the frequencies from that of the lowest score to the highest. The frequency of the lowest score is carried over to form a base. This is then added to the frequency of the next higher score or class continuously until that of the last score. This is represented in Table 5 below.

Table 5: Cumulative Frequency Distribution of a Grouped Data

Classes	F	Cf
55-59	1	50
50-54	1	49
45-49	3	48
40-44	4	45
35-39	6	41
30-34	7	35
25 - 29	12	28
20-24	6	16
15-19	7	10
10-14	3	3

Graphic Representation of Frequency Distributions

In discussing this section, the entry behaviour assumed is that the reader has known how to draw graphs. If you do not know this or have forgotten, please revise arithmetic sections that deal with graphs. We plot graphs on graph paper. We may represent the frequency distribution in a graphic form because:

a) the pictorial effect easily catches the eye, that is

b) the graph acts as a seductive slogan that holds attention. There are three methods of representing a frequency distribution graphically which we shall consider.

   i. Frequency polygon

   ii. Histogram

   iii. Cumulative frequency graph or ogive

Frequency Polygon of an Ungrouped Data

For an ungrouped data, the frequency polygon is plotted by listing the scores (x) on the x axis of the graph and the frequencies of the scores (f) on the y axis. A dot or mark is made at the intersect between each score and its frequency, after which the marks are connected with lines.

Using the frequency distribution in Table 3 above, the frequency polygon is thus represented in fig. 1:

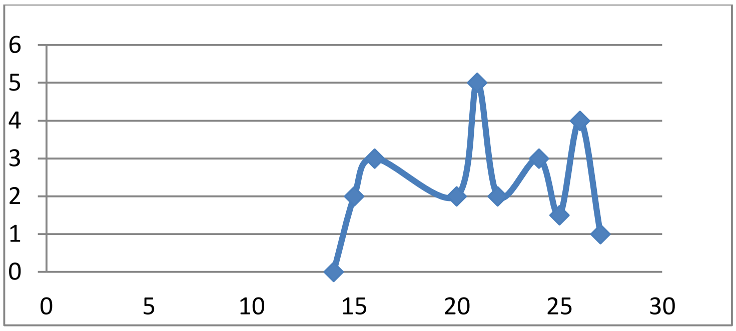


Fig. 1: Frequency Polygon of ungrouped data

Frequency Polygon of a Grouped Data

To construct the frequency polygon of a grouped data, we need to first of all determine the class marks or midpoints of each class interval. This is written on the y axis.

The class mark of any class is the midpoint of the class which is by listing determined by dividing the sum of the extreme scores by two. Using of the data in Table 5 above, the class marks are presented in the table 1 below as:

Table 6: Marks (Midpoints) of a Grouped Data

Classes	**F**	**Cf**	**Class Marks (x)**
10-14	4	4	12
15-19	6	10	17
20-24	2	12	22
25 - 29	8	20	27
30-34	4	24	32
35-39	6	30	37
40-44	4	34	42
45-49	4	38	47
50-54	8	46	52
55-59	9	55	57

Using the data in Table 6 above, we construct the frequency polygon using the x axis for the scores (mid-point of class intervals) and the y axis for the frequencies. We then join the lines neatly. We choose the scales for the graph making sure for a good graph, that the y unit is about 75% of the x units.

In the polygon graph, the total area of the polygon represents the total frequency N.

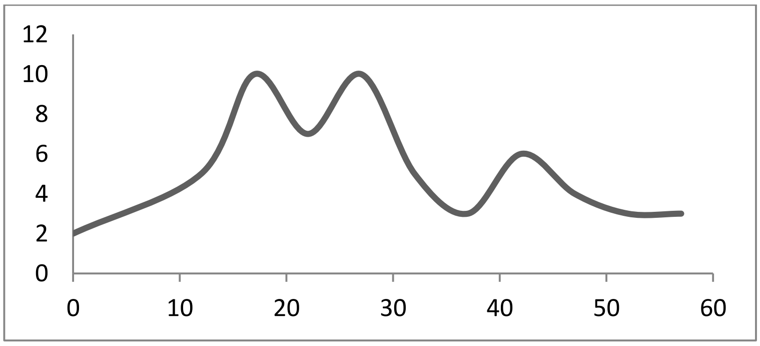


Fig. 2: Frequency Polygon of a Grouped Data

Smoothing the Frequency Polygon

The frequency polygon is usually very irregular and jagged in outline. To iron out the irregularities, the frequency polygon is often smoothened.

In smoothening, we use the adjusted values of "f”. These values are found by adding together the frequency above the one we want to adjust, the particular frequency being adjusted plus the frequency below it. We divide the sum by 3. See table 7 below.

Table 7: Smoothened Frequencies of a Grouped Data

Class Interval	F	Mid Point (x)	Cum f.	Smoothened F
195 – 199	2	197	50	1.33
190 – 194	2	192	49	2.67
185 – 189	4	187	47	3.67
180 – 184	5	182	43	5.67
175 – 179	8	177	38	7.67
170 – 174	10	172	30	8.00
165 – 169	6	167	20	6.67
160 – 164	4	162	14	4.67
155 – 159	4	157	10	3.33
150 – 154	2	152	6	3.00
145 – 149	3	147	4	2.00
140 – 144	1	142	1	1.33

If we want to adjust ft. frequency 2 in the class interval 190 - 194, we take the given frequency which is 2, the frequency above this, which is also 2, the frequency below which is 4, we add up these and divide by 3. This gives us 2.67.

For the next frequency (for the class interval 195 - 199), we add the frequency of the next higher class, the frequency of the class and the frequency of the next lower class and divide by 3 i.e.

$$\frac{0 + 2 + 2}{3} = \frac{4}{3} = 1.33$$

Now a graph is plotted using the adjusted or smoothened “f” instead of the original ***f***.

Exercise: Now use the data in Table 7 to plot the frequency polygon. On the same graph paper plot the smoothened frequency polygon. What do you notice?

The Shape of the Frequency Polygon

In statistical measurements, all things being equal, when the frequency polygon of a set of data is plotted, we expect to get graph of the normal curve which is symmetrical and dumb bell shaped. In practice, the graph we obtain may be skewed.

In the case where the frequency polygon is normally curved, the mean, the median, and the mode all coincide and there is perfect balance between the right and left halves of the polygon.

Where the mean and the median fall at different points in the distribution, and the balance is shifted to one side or the other-to left or right, the distribution is said to be skewed.

Distributions are said to be skewed negatively or to the left when scores are massed at the high end of the scale (the right end) and are spread out more gradually toward the low end (or left) as in fig 4 below. In such cases, the mode is greater than the median and the median is greater than the mean. Distributions are skewed positively or to the right when scores are massed at the low (or left) end of the scale and are spread out gradually toward the high or right end as shown in fig 5 below (Garrett, 1966). In such cases, the mean is greater than the median and the median is greater than the mode.

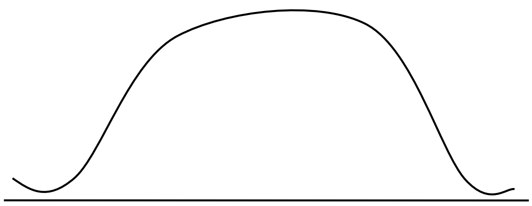
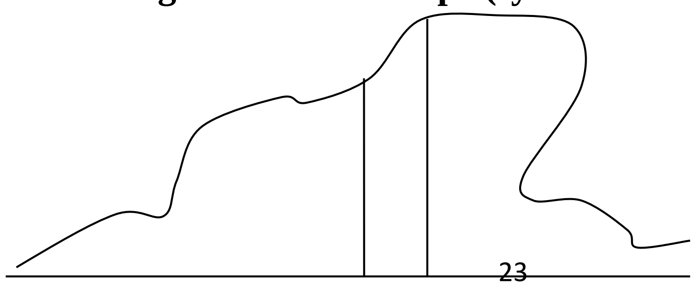
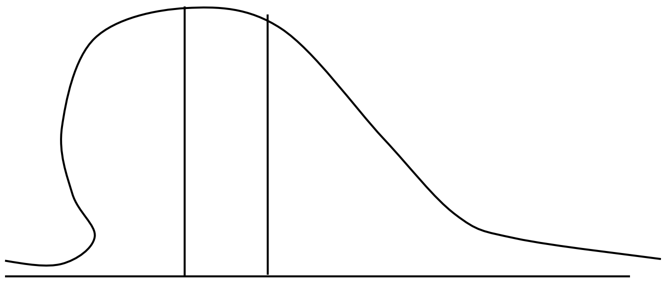


Fig. 3: Showinga dumb bell shape (symmetrical) distribution



Mean Median

Fig. 4: Showing negative skewness to the left



Median Mean

Fig. 5: Showing positive skewness to the right

We will notice the figures above (figs 4 and 5) are skewed to the left and to the right respectively. To determine the direction of the skew, we also use the tail of the polygon. If the tail is on the left side, we say that it is negatively skewed. If the tail is on the right side, we say it positively skewed. To calculate the skewness of a distribution, we apply the formula:

$$SK = \frac{3(\text{mean} - \text{median})}{SD}$$

Kurtosis of Data

The term "Kurtosis," to Garrett (1966, p. 101), "refers to the 'peakedness' or flatness of a frequency distribution as compared with the normal."

Kurtosis is also described as the 'curvedness' of the graph of a distribution. Kurtosis is frequency used in a relative sense. There are different forms of curves or peaks which the frequency polygon of data distributions may take. These forms depend on the data collected, and they are:

a. Mesokurtic

This refers to a symmetrical shaped distribution or a normally curved distribution as represented in the figure below:

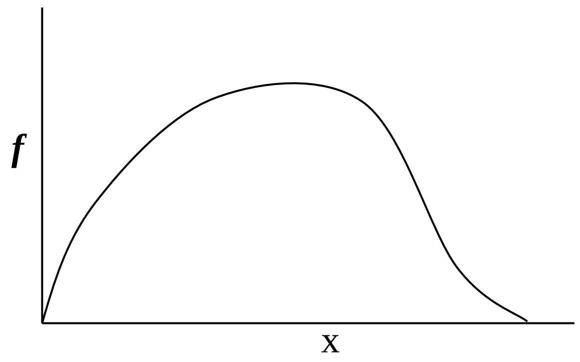
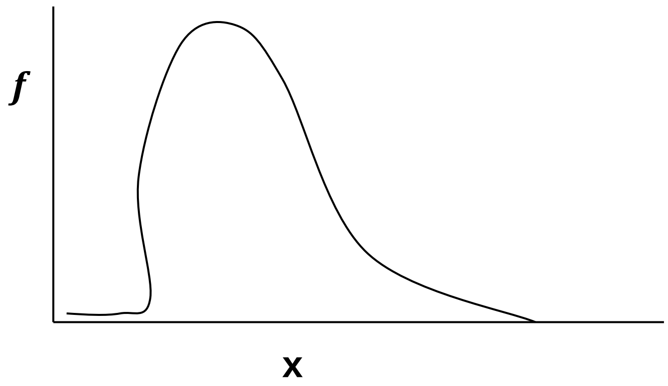


Fig. 6: A Mesokurtic Distribution (Normal Curve)

b. Leptokurtic

The Greek word lepto means thin, so leptokurtic implies a thin distribution. Another way of describing leptokurtic distribution is a distribution with a high peak as in figure 7 below.



c. Platykurtic:

Platy means flat, therefore, a platykurtic distribution is a distribution with a flatter curve than that of normal distribution. A platykurtic distribution is represented in fig. 8 below.

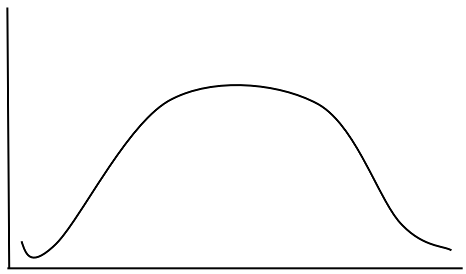


Fig. 8: Platykurtic Distribution

It is necessary to determine the kurtosis of a data when one wants to ascertain whether the data is normally distributed or not. Assumptions of a normally distributed population helps one to decide whether the test statistics will be parametric or non parametric. The formula for measuring kurtosis is:

$$Ku = \frac{Q}{(P_{90} - P_{10})}$$

where Q = Quartile deviation of the distribution.

For a normally distributed data, the Ku = .263. If Ku is greater than .263 the distribution is platykurtic, if less than .263 the distribution is leptokurtic.

Histogram or Column Diagram

This has the same shape as the frequency polygon. Instead of using the midpoint of the class interval to draw a line graph as in the frequency polygon, the frequencies in the histogram are represented by a column or a rectangle respectively.

For example, in Table 7, for the class interval 140 - 144, the frequency is 1 and should be plotted against the midpoint of the class interval which is 142.

In the histogram, the point, 142, can be represented by a line of 1cm.

The lower and upper limits are by the ends of the line while the middle is really 142. The frequency is indicated along the Y axis.

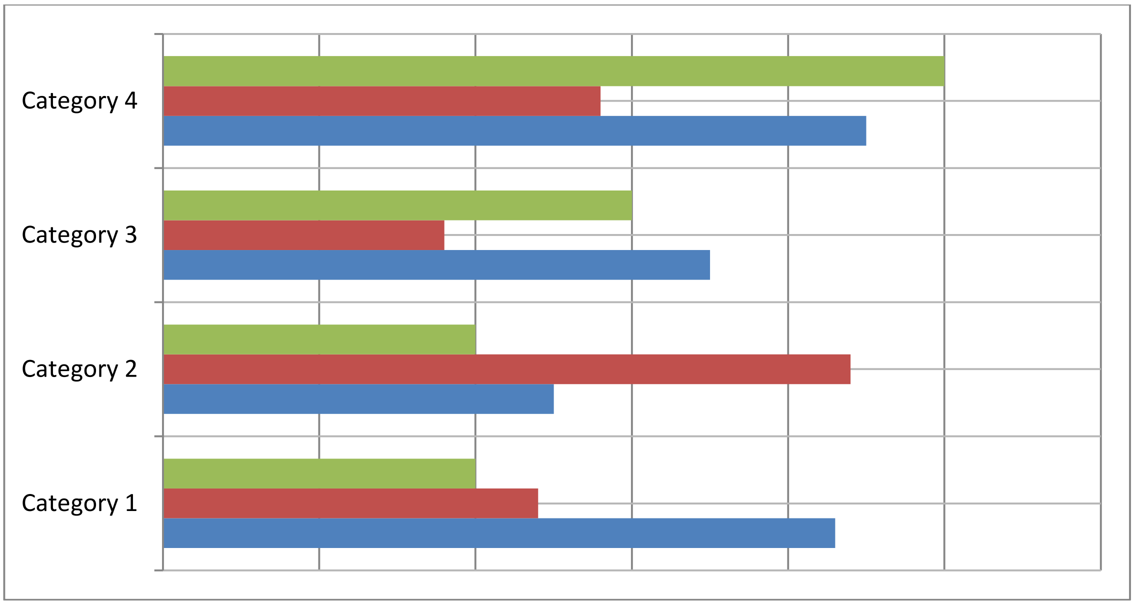


Fig. 9: Histogram or column diagram of a grouped data

The graph looks like series of rectangles of the same base (arbitrarily chosen) but different heights which depend on the frequency.

Note that the area of each rectangle in a histogram is directly proportional to the number of scores or measures within the interval. The histogram presents accurate relative proportions of the total number.

Exercise:

With the knowledge of histogram, construct a histogram using the data on Table 2.

Class Limits:

Each class in a grouped data has two limits viz: lower-class limit and upper class limit. These class limits are the extreme scores in a distribution. Using the data in Table 8, for example, the class limits of class interval 1 are 8 and 10 respectively, 8 becomes the lower one while. 10 is the upper limit. These types of class limits are appropriate for discrete variables. However, for continuous variables to which education data belong, true or actual class limits are used in statistical analysis. These class limits are boundaries between one class and another. .To determine the class boundaries, we simply subtract .5 from the lowest score in each class for lower class limits, and add .5 to the highest score for the upper class limit. The class limits are presented in the Table 9 below.

Table 8: Upper and Lower Class Limits of a Grouped Data

Classes	F	Lower Class Limit (LCL)	Upper Class Limit (UCL)
1-15	4	0.5	15.5
16-30	6	15.5	30.5
31- 45	8	30.5	45.5
46- 60	2	46.5	60.5
61- 75	6	60.5	75.5
76 - 90	8	75.5	90.5
91 - 104	9	90.5	104.5
105 - 119	4	104.5	119.5
120 - 134	6	119.5	134.5

Percentage Frequencies and Cumulative Proportion Frequencies

Percentage frequencies are determined by dividing each ***f*** by N, and then multiply the result by 100. The percentage frequency of class 2 in Table 9 is 4/86 x 100 = 4.7. Proportion frequency is obtained by simply dividing each *f* by N. For our data in Table 9, the proportion frequency for the class interval 11 - 13 with a frequency of 4 is 4/86 =0.047.

To get the proportion frequency for all other classes, we use the same procedure.

Cumulative percentage frequency and cumulative proportion frequency are computed by adding up progressively the percentage or proportion frequencies respectively already determined starting from the lowest class. This is because in adding progressively from the bottom up, each cumulative frequency carries through to the exact upper limit of the interval. The cumulative frequency graph of the data in Table 7 is presented below.

The Cumulative Frequency Graph (Ogive)

Table 9: Cumulative frequencies, percentages and proportions for memory test scores

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Scores	**Upper Limit**	**Lower Limit**	**F**	**CF**	**PF**	**CPF (Cumulative Percentage Frequency)**	**PF (Proportion Frequencies)**	**CPRF (Cumulative Proportion Frequencies)**
41 – 43	42.5	40.5	1	86	1.2	100.1	0.012	1.00
38 – 40	40.5	37.5	4	85	4.7	98.9	0.047	0.99
35 – 37	37.5	34.5	5	81	5.8	94.2	0.058	0.943
32 – 34	34.5	31.5	8	76	9.3	88.4	0.093	0.885
29 – 31	31.5	28.5	14	68	16.3	79.1	0.163	0.792
26 – 28	28.5	25.5	17	54	19.8	62.8	0.198	0.629
23 – 25	25.5	22.5	9	37	10.4	43.0	0.105	0.431
20 – 22	22.5	19.5	13	28	15.1	32.6	0.151	0.326
17 – 19	19.5	16.5	8	15	9.3	17.5	0.093	0.175
14 – 16	16.5	13.5	3	7	3.5	8.2	0.035	0.082
11 – 13	13.5	10.5	4	4	4.7	4.7	0.047	0.047
8 – 10	10.5	7.5	0	0	0	0.0	0.000	0.000

N = 86

The construction of the ogive or cumulative frequency graph using the above data is simple. It is the graph which uses the upper limit of the class interval and the cumulative frequency, (cf) cumulative percentage frequencies (cpf) or cumulative proportion frequencies (cpfr).

However, the cumulative frequency graph can be constructed without the cpf and cpfr. Before we can plot a cumulative frequency polygon (graph), the scores of the distribution must be added serially or cumulated as in Table 9. We then write out the exact upper limits of the class intervals along the x axis and their cumulative frequencies along the y axis. This is because in adding progressively from the bottom, each cumulative frequency carries through to be exact upper limit of the interval. The cumulative frequency graph of the data in Table 9 is presented below:

Note that if we are plotting the frequency polygon, we use the midpoints of the class intervals with their respective frequencies instead.

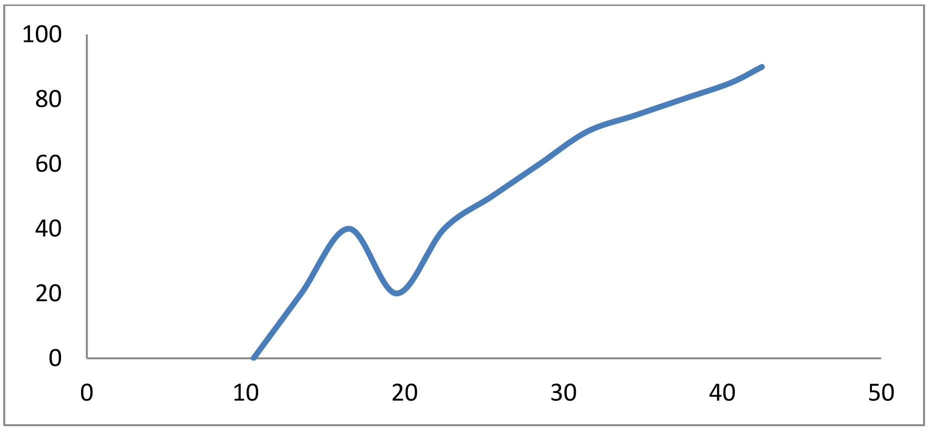


Fig. 10: Cumulative Frequency Polygon

Exercise

1. The following are marks obtaiend by a group of first year technical education students

33	88	37	98	93	73	82
56	80	52	66	54	73	62
80	62	53	69	56	81	75
52	65	49	67	59	88	80
60	74	72	91	82	89	67

Prepare a frequency distribution and a cumulatvie frequency distribution for this data using a class interval of 5.

2. Prepare a frequency distribution for the following test scores.

6	2	15	11	6	4	18	1	9	5
3	8	7	16	12	11	17	13	3	7
5	11	7	9	5	16	16	16	4	9
11	4	18	10	4	4	15	15	5	5
10	9	1	19	8	8	7	7	2	12

Obtain a cumulative percentage frequency distribution for the data.

3. Prepare histogram for the data in number 2 above.

4. Suppose a set of scores numbered (a) to (e), have the ranges given below, indicate how large an interval and how many intervals you would suggest for use in drawing up a frequency distribution of each set.

Range of Mark Size of Intervals Number of Interval

a). 16 to 87

b) 0 to 46

c) 110 to 212

d) 63 to 151

e) 4 to 12.

5a. The following represent the scores of students in Educational Measurement and evaluation.

Prepare a frequency distribution table of five classes for the data.

50	71	90	51	70	30	87	79	81
58	76	72	51	76	90	71	72	62
89	90	71	88	66	40	91	71	63
65	76	79	80	76	54	80	72	63
91	90	45	69	66	79	71	75	68
90	67	67	52	64	88	70	80	90

b. Plot the cumulative frequency graph with the data in 5a.

6. Explain with illustration the meaning of skewness and kurtosis of distributions.

CHAPTER THREE

DESCRIPTIVE STATISTICS

The statistical procedures that are used for just describing and summarizing data are called descriptive statistics. There are many procedures under this type of statistics. These are:

a. Measures of central value or tendency (Averages)

b. Measures of variability or dispersion.

c. Measures of association or relationship.

d. Measures of relative standing.

The Measures of Central Value or Tendency (Averages)

The central value is of two folds. Firstly, it is an average of a group. 'It serves as a description of a mass of quantitative data from a sample. As an average, it stands for a group and saves us the problems of the details of each member of the group. The average mark in examination immediately gives an idea of the general group performance. Secondly, measures of central value help us to compare two or more groups in terms of typical performance. In statistics, there are three kinds of averages:

i) The mean (arithmetic average)

ii) The median

iii) The mode.

Note: The term 'average' is popularly used for the arithmetic mean.

In statistics analysis, however, 'average' is the general term for any measure of central tendency.

The Mean

The mean is arithmetic average of a distribution. This equals dividing the summation of total scores in a distribution by number of the scores.

Analysis of Mean for Ungrouped Data

Example 1:

Suppose we have a set of marks:

45, 65, 94, 30, 75, 20, 10, 56, 45, 20

We may represent each of these numbers by x. The mean, X of all the numbers is given by:

$$\frac{x}{N}$$

where X

$$= \frac{45 + 65 + 94 + 30 + 75 + 20 + 10 + 56 + 45 + 20}{10}$$

$$= \frac{460}{10} = 46$$

Where X = arithmetic mean.

The Mean of a Frequency Distribution

In the example above, each value of x was listed individually even when any occurs more than once. There may be a situation where each x may occur more than once but it is listed once with the frequency of occurrence recorded as below.

X	***f***	***Fx***
18	1	18
17	2	34
16	2	32
15	3	45
14	2	28
13	5	65
12	3	36
11	2	22

∑f or N = 20 ***∑fx = 280***

The score 18 has occurred once - frequency is one.

The score 17 has occurred twice - frequency is two

The score 12 has occurred three times, etc.

In the above illustration, x is the score, *f* is the frequency, fx is the product of the score and its frequency of occurrence.

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

In this case x will be replaced by fx

N remains as number of scores.

For such frequency distributions,

$$x = \frac{\sum fx}{N} \quad \text{i.e. } \frac{18 + 34 + 32 + 45 + 28 + 65 + 36 + 22}{280}$$

In general, where X_1, x_2, x_3, ... x_4 occur with frequencies f,, $f_{2,}$ f_3 . . f_4

The arithmetic mean = $\underline{fx_1 + fx_2 + fx_3 + fx_4}$.

In the example above = $\frac{280}{20}$ = 14

Analysis of Mean for Grouped Data

Using the distribution below, the mean computation is:

Table 10: Frequency Distribution of Grouped Data for Mean Computation

Class Interval	**Mid Point x**	**Frequency (*f*)**	**Frequency Mid Point FX**
45 – 49 40 – 44	47 42	1 2	47 84
35 – 39 30 – 34	37 32	3 6	111 192
25 – 29 20 – 24	27 22	8 17	216 374
15 – 19 10 – 14	17 12	26 11	442 132
5 – 9 0 – 4	7 2	2 0	14 0
		76	**1612**

For a problem of this type involving the class interval, we use the mid point of the class interval as x

$$X = \frac{\sum fx}{\sum f} = \frac{1612}{76} = 21.2$$

Example 2

Consider the data below, been scores generated from a test administered by a Physics teacher in City girls secondary school

~~22~~	~~10~~	~~20~~	~~32~~	~~30~~	~~70~~	~~88~~	~~40~~	~~33~~	~~16~~
~~33~~	~~8~~	~~16~~	~~23~~	~~35~~	~~76~~	~~87~~	~~27~~	~~15~~	~~27~~
~~44~~	~~15~~	~~26~~	~~37~~	~~47~~	~~78~~	~~90~~	~~66~~	~~37~~	~~40~~
~~51~~	~~26~~	~~37~~	~~64~~	~~58~~	~~69~~	~~10~~	~~25~~	~~86~~	~~51~~
~~62~~	~~74~~	~~85~~	~~90~~	~~69~~	~~80~~	~~15~~	~~72~~	~~71~~	~~20~~

Categorize the distribution into seven classes, and ascertain the mean.

Solution

Step 1

$$C = \frac{R}{K}$$

$$C = \frac{90-8}{7}$$

$$C = 11.7$$

$$C \approx 12$$

Step 2. Construction of the Frequency Distribution Table

C. I	Midpoint (X)	Tallies	Frequency	FX
8 – 20	14	~~IIII~~ ~~IIII~~	10	140
21 – 33	27	~~IIII~~ ~~IIII~~ I	11	297
34 - 46	40	~~IIII~~ II	7	280
47 – 59	53	IIII	4	212
60 – 72	66	~~IIII~~ III	8	528
73 – 85	79	~~IIII~~	5	395
86 - 98	92	~~IIII~~	5	460
			ΣF = 50	ΣFX = 2312

Mean $(\overline{X}) = \frac{\Sigma FX}{\Sigma F}$

Mean $(\overline{X}) = \frac{2312}{50}$

$= 46.24$

Computation of Mean Using Assumed Mean Formula

The formula used in the computation of the mean above is the ordinary mean formula. The mean can also be computed using the Assumed or Guessed mean formula. This formula is appropriate when the data is large and use of raw scores with ordinary mean formula becomes tedious.

In assumed mean formula one assumes that one of the scores or mid points is the mean. The differences between the assumed mean and each of the other scores or mid points are gotten. These differences are summed and divided by number. The difference can also be determined arbitrarily starting from any origin. Examples of determining the mean with assumed mean formula are given below.

Computation of Mean with Assumed Mean Formula Using Deviation Scores

Using the distribution in Table 11 below, and taking 22 as the assumed mean, the mean becomes:

$$Xa \quad = \quad A + \left(\frac{\sum fd}{N}\right)$$

Where:

A = the assumed mean

D = difference between the assumed mean and other scores (x).

Table 11: Computation of Mean with Assumed Mean Formula Using Deviation Scores

Class Interval	**Mid Point x**	**Frequency (*f*)**	**d**	**Fd**
45 – 49	47	1	25	25
40 – 44	42	2	20	40
35 – 39	37	3	15	45
30 – 34	32	6	10	60
25 – 29	27	8	5	40
20 – 24	22	17	0	0
15 – 19	17	26	-5	-130
10 – 14	12	11	-10	-110
		$\sum f$ = 74		**$\sum Fd$ = -30**

Illustration

Consider the data below obtained from the sale day book of Mr. Okafor who operate in shop rite Enugu.

3, 6, 9, 12, 16, 19, 24, 22, 27, 1.

4, 8, 10, 14, 16, 19, 21, 25, 2, 1

29, 3, 7, 11, 13, 16, 18, 23, 26, 29

5, 7, 11, 13, 15, 18, 21, 1, 4, 1

8, 6, 9, 10, 9, 14, 17, 15, 20, 5

2, 4, 7, 14, 20, 12, 12, 14, 6, 3.

Given that the assume mean is 8, calculate the mean of the distribution and compare your answer with the mean calculated in the above example.

Solution:

Class Interva	**Class Boundary**	**Class Mark (X)**	**Tally (T)**	**Freq (F)**	**X-X (D)**	**FD**
0-2	-0.5-2.5	1	*~~IIII~~ I*	6	-7	-42
3-5	2.5-5.5	4	*~~IIII~~ III*	8	-4	-32
6-8	5.5-8.5	7	*~~IIII~~ III*	8	-1	-8
9-11	8.5-11.5	10	*~~IIII~~ II*	7	2	14
12-14	11.5-14.5	13	*~~IIII~~ III*	8	5	40
15-17	14.5-17.5	16	*~~IIII~~ I*	6	8	48
18-20	17.5-20.5	19	*~~IIII~~ II*	7	11	77
21-23	20.5-23.5	22	*IIII*	4	14	56
24-26	23.5-26.5	25	*III*	3	17	51
27-29	26.5-27.5	28	*III*	3	20	60
Total				∑F =60		∑FD= 264

Average deviation = $\frac{\sum FD}{\sum F}$

A.D = $\frac{264}{60}$

A.D = 4.4

Mean (X) = 8+4.4

Mean (X) =12.4

Computation of Mean with Assumed Mean Formula Starting From Arbitrary Origin

The assumed mean formula used above was done by obtaining deviation scores through subtraction of the assumed mean from each of the X. We can, however, compute the mean with assumed mean formula where we obtain the deviation scores by starting from arbitrary origin. We start from the centre of the distribution to obtain our deviations and in this case, we take the deviation score of that class interval to be 'o', all the subsequent higher class intervals, we affix 1, 2, 3, etc as the deviation scores and -1, -2, -3, etc-to the lower class intervals. However, the class mark of the class interval which is the starting point will automatically be taken as the assumed mean. The formula of this procedure is:

$$Xa = A + \left(\frac{\sum fd}{N}\right)$$

Where;

A = the assumed mean

d = deviation scores from arbitrary origin

c = class size of the assumed mean.

We have to note that whenever we compute the mean using assumed mean formula where the deviation scores are obtained arbitrarily, the $\sum fd/N$ is always multiplied by c(class size) before we add the result to the assumed mean. Using the data in Table 11, the computation goes in this way.

Class Interval	**Mid Point x**	**Frequency (*f*)**	**d**	**Fd**
45 – 49	47	1	4	4
40 – 44	42	2	3	6
35 – 39	37	3	2	6
30 – 34	32	6	1	6
25 – 29	27	8	0	0
20 – 24	22	17	-1	-17
15 – 19	17	26	-2	-52
10 – 14	12	11	-3	-33
		$\sum f$ **= 74**		$\sum Fd$ **= -80**

Mean of Combined Groups

If we are given the mean of two or more groups and their numbers, N, we can calculate the mean of the combined groups. Let there be two groups A & B. Let the number in group A = 30; Mean = 16. Let the number in group B = 40, Mean =. 14., To find the mean of the groups combined.

Total number for A = 30 x 16 = 480

Total number for B = 40 x 14 = 560

Mean of the combined group

$$= \frac{480}{30} + \frac{560}{40} = \frac{1040}{70}$$

= 14.86

In general, let N for the groups A, B, C, etc be NI, N_{2i} N_3, etc respectively and their means X1, X2, X3, etc. The mean of the combined groups =

$$\frac{N_1X_1 + N_2X_2 + N_3X_3}{N_1 + N_2 + N_3}$$

Sometimes, the reporting of the mean cannot give a needed picture of a distribution. Suppose in an examination involving 40 pupils, one scored 100%, two scored 80%, ten 0%, seven scored 20%, 20 scored 30%.

No. of Students (f)	**Scores (x)**	**Fx**
1	100	100
2 10	80 0	160 0
7 20	20 30	140 600
40	230	1000

N =40; X = 25

This mean, does not tell much about the group and is not descriptive of the group scores. This is so because if we plotted the graph, we would find out that it is much skewed. In every skewed condition, the mean may not be a good measure of the group characteristic. The mean is supposed to be the center of symmetry of distribution in the group. If the mean is not the center of symmetry, it gives a distorted picture.

The Median

The median is a point in the scale of distribution wherein half of the scores falls below it. It is necessary that in discussion about the - median, the distribution must be arranged in an order, either ascending or descending order of magnitude. The median, therefore, is an ordinal statistics.

Suppose there is a record of marks.

2, 27, 20, 7, 19, 25, 16.

In order to determine the median, we arrange the marks in ascending or descending order.

2, 7, 16, 19, 20, 25, 27.

We look for the median, i.e. the score that is mid way. That is 19.

Three scores fall above 19 and three scores fall below it. **Note:** The median can be easily determined by taking the score at the middle after ranking the scores. If two scores fall in the middle in the case of even numbered distributions, add the two scores and divide by 2.

You may notice that ungrouped scores are arranged in order of size. Two things can happen.

i. N can be odd.

ii. N can be even.

When N is odd, there is no problem finding the median. When N is even as in the series below:

7, 8, 9, 10, 11, 11

The median lies between 9 and 10 i.e. the 3.5th score. Add scores 9 and 19 and divide by 2, which gives us 9.5. The median of the distribution is 9.5

Calculation of the Median When Data are Grouped Table 13: Computation of the Median for Grouped Data

Class Interval	**f**	
40 – 44	1	32 cases above the interval containing the median
35 – 39	0	
30 – 34	3	
25 – 29	15	
20 – 24	13	
15 – 19	10	Median lies here
10 – 14 5 – 9	11 11	
0 – 4	4	26 cases below
	68	

The median lies between 34^{th} case above or 34^{th} case below in the table. Counting from the top, the number of cases above the interval containing the median is 32. The median must, therefore, be contained in the frequency 10.

10 cases lie within 15-19 intervals more exactly between 14.5-19.5.

The score must lie within this.

The class interval width is 5

We divide this by 10 to get 0.5

We have 32 cases above the median class. To get 34 which is half the number of scores, we need 2 more marks.

To get 2 out of this we multiply 0.5 by 2. This gives 1.00.

When we subtract this from the upper limits of 19.5 we get the, median, which is 18.5.

We may approach the problem counting from below. We have 26 cases below ten. We need 8 more to obtain exactly half of our distribution and these lie within the frequency 10 in the class intervals 14.5 to 19.5.

If 10 contains 5 scores, 8 contains $\frac{5}{10} \times \frac{8}{1} = 4$

This time we add to the lower limit of 14.5.

i.e. 14.5 + 4 = 18.5. You must decide which way to consistently use.

Calculation of the Median of Grouped Data Using Median Formula

The median calculated above can be determined with a formula as:

$$Md = L_1 + \left(\frac{N/2 - cfb}{f_1}\right) c$$

where L_1 = lower limit of the median class

N = number of scores

cfb = cumulative frequency of the next lower class to the median class.

Procedure

Steps in calculation of the median

1. Determine the cumulative frequencies of the scores in the distribution.
2. Divide the number of scores by 2 and look for the result in the cumulative frequency column. If not seen, take the one nearest to it but not below it and call the class with that cumulative frequency the median class.
3. Calculate the median, using the formula as in the example below.

Table 14: Data for Computation of the Median

Class Interval	***f***	***Cf***
40-44 35-39	1 0	68 67
30 - 34 25 - 29	3 15	67 64
20 - 24 15-19	13 10	49 36 median class
10-14 5-9 0-4	11 11 4	26 15 4
	68	

$$Md = 14.5 + \left(\frac{68/2 - 26}{\quad}\right)5 = 18.5$$

10

Example 2

The below data, has been fetched from the scattered information kept by Nwatu Ltd. Who is into the sales of sandal at the modern market Enugu.

Required: Calculate the median of the distribution.

22, 23, 25, 31, 32, 33, 38, 41, 44, 47
51, 54, 55, 21, 23, 26, 30, 32, 36, 39
42, 44, 47, 52, 53, 58, 20, 25, 28, 30
32, 36, 39, 42, 44, 48, 51, 55, 24, 27
28, 33, 35, 38, 43, 46, 47, 50, 25, 29
34, 37, 39, 42, 46, 49, 47, 29, 33, 35
36, 40, 43, 46, 32, 36, 38, 43, 45, 45
33, 37, 37, 40, 41, 40, 41, 38, 42, 40

Solution

Table Eight

Class interval	Class Limit	Frequency	Cumulative frequency
20-22	19.5-22.5	2	2
23-25	22.5-25.5	4	6
26-28	25.5-28.5	5	11
29-31	29.5-31.5	6	17
32-34	31.5-34.5	8	25
35-37	35.5-37.5	10	35
38-40	37.7-40.5	12	47
41-43	40.5-43.5	10	57
44-46	43.5-46.5	8	65
47-49	46.5-49.5	6	71
50-52	49.5-52.5	4	75
53-55	52.5-55.5	3	78
56-58	55.5-58.5	2	80
		$\sum$ F=80	

Median = $L_i + \frac{(\frac{N}{2} - \frac{CFb}{1})I}{f}$

Where Li =37.5
N = 80
F = 35
f=12

Median = 37.5 + $\left(\frac{80}{2} - \frac{35}{12}\right) \times 3$

$37.5 + \left(40 - \frac{35}{12}\right) \times 3$

$37.5 + (5/12) \times 3$

$37.5 + (15/12)$

$37.5 + 1.25$

Median = 38.75.

Example 3

Consider the table below

C.I	Frequency	C.L	Cumulative frequency
1-5	3	0.5-5.5	3
6-10	2	5.5-10.5	5
11-15	8	10.5-15.5	13
16-20	4	15.5-20.5	17
21-25	7	20.5-25.5	24

To identify the median class

Divide Σf by 2 = $\frac{24}{2}$= 12

The median class lies on class 3

L_1 = 10.5

N = 24

Cbf = 5

C = 5

F = 8

$= 10.5 + \frac{(\frac{24}{2} - \frac{5}{1})}{8}$

$= 10.5 + \frac{(12-5)5}{8}$

$=10.5 + \frac{35}{8}$

Median = 14.9

The Mode

The mode is that single measure or score which occurs most frequently. In other words, the mode is said to be the most occurring score or variate. This measure of central tendency is the simplest and most unreliable. It is the statistical procedure applied when the data is on the nominal scale, and it determines the height of the peak of a distribution.

The mode for an ungrouped data can simply be determined by tallying the frequencies of the scores and taking the score that occurs most frequently. For example, in a distribution with the scores: 3, 4, 5, 5, 6, 7, 7, 7, 7, 8, 8, 9, 10, the most occurring score which is 7 becomes the mode.

There can be more than one mode in a distribution. Where the mode is only one, we call the distribution a uni-modal distribution. If the mode is two, it becomes a bi-modal distribution and where it is more than two, it becomes a multimodal distribution.

Computation of the Mode for a Grouped Data

We can compute the mode for a grouped data in two different ways. First, we can simply determine the mode of a grouped data by taking the class midpoint, 'x' of the modal class as the mode. Secondly, the mode of a grouped data can be determined using the formula.

Where:

M_0 = the mode

LI = lower limit of the modal class

D1 = the difference between the modal frequency and the frequency of the next
lower class.

D2 = the difference between the modal frequency and the frequency of the next higher class.

c = class size (width) of the modal class.

Note: To compute the mode, you will first determine the modal class which is the class with the highest frequency.

Table 15: Computation of the Mode for a Grouped Data.

Example 1

Class interval	F
20-22	2
23-25	4
26-28	5
29-31	6
32-34	8
35-37	10
38-40	12
41-43	10
44-46	8
47-49	6
50-52	4
53-55	3
56-58	2
Total	$\sum$ F=30

$M_o = L_1 + [\frac{D1}{D1+D2}]I$

Where: Li=37.5

Di=12-10

D_2=12-10

C=3

$Mo = 37.5 + \frac{(2)}{2+2} \times 3$

Mo=37.5 + 6/4

Mo=37.5 +1.5

Mo=39

Example 2

Consider the table below, determine the mode

C. I	8 - 20	21 - 33	34 - 46	47 - 59	60 - 72	73 - 85	86 - 98
Frequency	10	11	17	4	8	5	5

Solution

C. I	Class limit	Frequency
8 - 20	7.5 – 20.5	10
21 - 33	20.5 – 33.5	11
34 - 46	33.5 – 46.5	17
47 - 59	46.5 – 59.5	4
60 - 72	59.5 – 72.5	8
73 - 85	72.5 – 85.5	5
86 - 98	86.5 – 98.5	5

Mode = $L_1 + \left(\frac{D_1}{D_1 + D_2}\right)C$

$L_1 = 33.5$

$D_1 = 17 - 11 = 6$

$D_2 = 17 - 4 = 13$

$C = 13$

Mode = $33.5 + \left(\frac{6}{6+13}\right)13$

$33.5 + \left(\frac{6}{19}\right)13$

$33.5 + (0.32)13$

$33.5 + 4.16 \Rightarrow 37.66$

Example 3

Consider the table below

Class Interval	**X**	***f***
40-44 35-39	42 37	1 0
30-34 25-29	32 27	3 15 modal class
20-24 15-19	22 17	13 10
10-14 5-9	12 7	11 11
0-4	2	4
		68

$$\mathbf{Mo = 24.5 + \left(\frac{15 - 13}{(15 - 13) + (15 - 3)}\right) 5}$$

$$\mathbf{= 25.2}$$

Uses of the Mean, Median and Mode

1. Where the distribution appears balanced from either side, the mean should be used, and is most appropriate. Where there is obvious skewness the median is preferable. The mode is used to get an idea of where there is the greatest bunching of marks.

When there is even distribution, we get a normal curve in which the mean and the median coincide. The curve is abnormal to the extent of the differences between the mean and the median.

Exercises

1a. What are measures of central value (tendency) used for?

b. Given the distribution of scores below, compute the median.

c. What is mode of the distribution?

X	22	24	26	28	30	32	34
F	2	4	4	12	6	2	3

2a. Determine the mean of the following students' test scores in Physics using the assumed mean formula.

Classes	**F**
20-26	12
27-33	22
34-40	25
41 -47	8
48-54	10
55-61	6

2b. What are the median and mode of the distribution?

3a. A researcher collected different sets of data from three groups as follows:

Group A: X = 35; N = 20

Group B: X = 44; N = 15

Group C: X = 50; N = 32

Determine the grand mean for the three groups.

Measures of Variability

Suppose two groups of students have taken, an examination, it is possible for both groups to have the same mean score and yet, for the individual scores of members of each group to differ very widely.

Example:	
Group A	**Group B Scores**
0	40
25	42
36	50
58	60
78	65
80	70
81	72
90	74
92	77
100	90
Sum = 640	**Sum = 640**
Mean = 64	**Mean = 64**

To compare the two groups on the basis of central value like the mean, would not tell the whole story and may mislead. We could get more meaning from the scores if perhaps we found the range of marks for each group (the marks for group A ranges from 0 – 100 and for group B, from 40 to 90). The measures that will help us determine how the students actually performed in the two groups are the measures of variability. These measures are measures that help us to know the spread of the scores of data in a distribution. If a group is homogeneous, that is, made up of individuals of nearly the same ability, most of its scores will fall around the same point in a scale, the range will be relatively short and the variability small. But if the group contains individuals of widely differing capacities, the range will be relatively wide and the variability large (Garrett, 1966).

The measures of variability are:

The range

The mean deviation

The quartile deviation or semi-interquartile range

The variance

The standard deviation.

The Range

The range is the simplest measure of variability. It is the difference-between the largest and smallest scores in a distribution. For the two) groups A and B mentioned earlier, the ranges are:

100 - 0 = 100

90 - 40 = 50

The first group shows greater variability of scores.

Disadvantages of the Range

i. For large samples, it is an unstable measure

ii. The range depends on the size of the sample.

iii. It caters only for the two extreme scores.

The Quartile Deviation or Semi-Interquartile Range

The median has been defined as the value of the variable which divides the group into two equal parts. The quartile is related to the median. The quartile divides each of the two groups of the median into 2. Quartile Deviation uses three quartiles in its analysis. These quartiles are the following:

Q_1 - 1st quartile = $^1/_4$ from the one end.

Q_2 - 2nd quartile i.e. the median.

Q_3 - 3rd quartile = % from the same one end.

The semi inter-quartile range or quartile deviation is half of the difference in value of Q_3 and Q_1 i.e. Q_3 minus $Q^1/_2$. The semi inter-quartile range is a better estimate than a full range because in the calculation of the semi inter-quartile range, we neglect or do not bother about the two extreme values of the scores. The semi inter-quartile range concentrates on the scores around the middle.

If the ogive is divided into 100 equal parts, each of these is called a percentile. In that case Q_3 is the 75th percentile while Q_1 is the 25th percentile, Q_2 is the 50th percentile.

Calculation of the Quartile Deviation for Ungrouped Data

As has been mentioned above, quartile deviation or semi interquartile range uses three quartiles in its analysis.

These are Q_1 = 1st quartile ($^1/_4$), Q_2 = 2nd quartile ($^1/_2$) and Q_3 = 3rd quartile ($^3/_4$).

Given the data below, the quartile deviation with the formula

$$QD = \frac{Q_3 - Q_1}{2}$$ is calculated thus:

22, 20, 23, 24, 18, 17, 16, 14

1. Arrange the scores in order of magnitude

   24, 23, / 22, 20, 18, 17, / 16, 14

   (Q_3 at the first /, Q_1 at the second /)

2. Determine position of Q_1 by first calculating ¼ of the score = ¼ x 8 = 2
3. Starting from the lowest score, count off two scores between the 2nd and 3rd scores lies Q_1. To get it exactly, add the 2nd and 3rd scores and divide by 2.

   $$Q1 = \frac{17 + 16}{2} = 16.5$$

4. Calculate Q_2 by first determining ¾ of the scores which is = 3 x 6 = 6. Starting from the lowest score, count off 6 scores. Q is the middle of the 6th and 7th scores.
5. Apply the QD formula:

   $$\frac{22.5 - 16.5}{2} = 3$$

Calculation of Quartile Deviation for a Grouped Data

Given the distribution below, take the following steps in calculating Quartile Deviation.

Table 16: Cumulative Frequency Distribution of Students' Scores

Class Intervals	***F***	***Cf***
52 – 58	10	77
45 – 51	8	67
38 – 44	22	59
31 – 37	16	37
24 – 30	15	21
17 – 23	6	6
	77	

1. Build a cumulative frequency
2. Determine Q_1 class by calculating ¼ of 77 = 19.25. Look up the result in the cumulative frequency column and take the cf nearest to it but not below it. The class that has it becomes Q_1 class.
3. Determine Q_3 class by calculating ¾ of 77 and apply the same procedure as in number 2 above.
4. Calculate Q_1 and Q_3 using the formulae:

$$Q_1 = L_1 + \left(\frac{{}^{N}/_{4} - Cfb}{f_1}\right) c; \qquad Q_3 = L_1 + \left(\frac{{}^{3N}/_{4} - Cfb}{f_1}\right) c;$$

where; L_1 is class limit lower of Q_1 or Q_3 class respectively.

Cfb = cumulative frequency of the next lower class to Q_1 or Q_3.

f_1 = frequency of Q_1 or Q_3

5. Apply Quartile deviation formula

Using the data in Table 16

Q1 class = $^{77}/_{4}$ = 19.25 approx. 19

This of nearest to 19 but not below it is 21. The class that has that cumulative frequency becomes Q_1 class.

$$Q_1 = 23.5 + \left(\frac{{}^{77}/_{4} - 6}{}\right) 7$$

15

= 29.68

Q_3class = ¾ x 77 = 57.75 approx. 58

The cf nearest to 58 but not below it is 59. The class that has the cumulative frequency 59 becomes Q3 class.

$$Q3 = 37.5 + \left(\frac{{}^{3 \times 77}/_{4} - 37}{22}\right)7$$

= 44.10

$$QD = \frac{Q_3 - Q_1}{2} = \frac{44.10 - 29.68}{2}$$

= 7.21

The Mean Deviation or Average Deviation

The mean deviation is interval statistics which makes use of all the scores in the distribution rather than only quarters or extreme score like the quartile deviation and range respectively.

The mean deviation is defined as the mean of deviation scores. Put in figurative form, it is presented thus:

AD or MD = $\frac{\sum(x - x)}{N}$ *For non frequency distribution;*

$\frac{\sum f(x - x)}{N}$ *For frequency distribution;*

Where;

X = Any score

AD or MD = Average Deviation or Mean Deviation

∑ = Sum of

N = Total number of scores

In finding the mean of average deviations, we use absolute deviations denoted by //. This means that we ignore the signs, add up the deviations and divide by the sample size. If we do not ignore the signs and add up the deviations, we may get zero as the mean deviation. This use of absolute deviations is the problem with mean deviation. A measure that handles negative values better rather than ignoring them when they occur in determining deviations is more preferable.

Calculation of the Mean Deviation

Using the distribution of the scores in group A in page 48 whose mean is calculated as 64, we can calculate the mean deviation as:

Table 17: Data for Computation of Mean Deviation

Group A X	**Difference from Mean x – x**	**Absolute Deviation /x – x/**
0	0 – 64	64
25 36	25 – 64 36 – 64	39 28
58 78	58 – 64 78 – 64	6 14
80 81	80 – 64 81 – 64	16 17
90 92	90 – 64 92 – 64	26 28
100	100 – 64	36

The mean deviation for group A is the arithmetic mean of the absolute deviation (disregarding the signs) from the mean i.e.

$$\frac{64 + 39 + 28 + 6 + 14 + 16 + 10 + 26 + 36}{10}$$

i.e. $\frac{274}{11} = 27.4$

If the data above is in frequency distribution, you obtain the deviations, multiply with the various frequencies, sum and divide by number to obtain the mean or average deviation.

Exercises

A teacher, teaching English Language to two classes of Primary Six with 10 pupils in each class, gave an essay examination to the pupils and collected the results below:

CLASS X 62, 63, 62.5, 64, 60.5, 99, 64, 60, 98, 60

CLASS B 98, 96, 63, 61, 99.5, 62, 62, 98.5, 62, 60.5

Calculate the deviations from the mean of the scores of each class and compare the performance of the two classes. What do you observe about the marking from the scores?

The Variance and Standard Deviation

These are the measures of variability commonly used. Like the mean deviation, the standard deviation is a kind of average of all deviations from the mean. The variance is the mean of squared deviation scores. This measure tries to overcome the deficiency of negative signs in mean deviation by squaring all the deviation scores before summing or multiplying with frequencies as the case may be. The standard deviation, on the other hand, is the square root of variance. In calculating the standard deviation, therefore, one follows the same procedure as the variance plus one further step which is getting the square root of the variance. There are various formulae which can be used to calculate the variance and standard deviation. These are the mean deviation formula, assumed mean formula and raw score formula. We will discuss the mean deviation formula for an ungrouped data; the mean deviation formula, assumed mean formula and the raw score formula for a grouped data.

The Mean Deviation Formula for Calculating the Variance and Standard Deviation of an Ungrouped Data

The Variance

The formula for determining the variance of non-frequency distribution is given below:

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

Where:		
∑	=	Sum of
X	=	Mean
X	=	any score
N	=	Sample size

Sometimes, instead of using N, N - 1 is used. The symbol used for variance is S^2 or ʅ^2. For this book we shall not bother about N - 1.

Calculate the variance for each of the scores of the class.

Group A	**Group B**
0	40
25	42
36	50
58	60
78	65
80	70
81	72
90	74
92	77
100	90

Put your work in a tabular form as follows:

Table 18: Computation of Variance of a Non-Frequency Ungrouped Data

Scores (x)	**Deviation scores (x – X)**		**Square Deviation Scores (x – X)^2**
0	0 – 64	(-64)	4096
25	25 – 64	(-39)	1521
36	36 – 64	(-28)	784
58	58 – 64	(-6)	36
78	78 – 64	(4)	196
80	80 – 64	(16)	256
81	81 – 64	(17)	289
90	90 – 64	(26)	676
92	92 – 64	(28)	784
100	100 – 64	(36)	1296
∑X = 640	**X = 64**		**9934**

$$= \frac{9934}{10} = 993.4$$

Procedure for the computation above

Step 1: Determine the mean of the distribution.

Step 2: Calculate the deviation scores for each score. This is done by subtracting the mean from each score (X-X).

Step 3: Square the deviation scores (X -X)

Step 4: Add the squared deviation scores and divide by number of scores to obtain the variance.

Exercise

Calculate the variance for Group B above.

Standard Deviation

This is the square root of variance. We will determine the standard deviation using the formula:

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

NOTE: *The smaller the standard deviation the less the variability of the scores or data. If the scores do not differ at all we have zero standard deviation.*

Example; In a class of 9 students of SSII, the geography teacher gave a test of which the scores are shown below:

Table 19: Calculation of the Standard Deviation of a Non Frequency Ungrouped Distribution

Scores (x)	**Deviation scores (x – X)**	**Square Deviation Scores (x – X)^2**
17	3.6	12.96
17 15	3.6 1.6	12.96 2.56
13 13	-0.4 -0.4	0.16 0.16
13 11	-0.4 -2.4	0.16 5.76

11 11	-2.4 -2.4	5.76 5.76
∑X = 121	**X = 13.4**	**46.24**

Procedure followed in the calculation in table 19

Calculate the mean, then the deviation of each score from the mean; square each deviation score, sum up and apply the formula.

NOTE: The variance and standard deviation calculated above are for non-frequency distributions.

For a frequency distribution, after squaring the deviations, the squared deviation scores are multiplied with their frequencies before summing and dividing by number.

The Variance (S^2) and Standard Deviation (SD or S) of Grouped Data

The variance and standard deviation are calculated using either the same mean deviation formula already discussed above for ungrouped data, the raw score formula or assumed mean formula. In this section, we are going to see how the variance and standard deviation can be computed using the mean deviation formula, the assumed mean formula and the raw score formula.

The Mean Deviation Formula for Calculating S^2 and SD of Grouped Data

Procedure

1. Determine the midpoints or class marks of all classes
2. Calculate the mean of the distribution.
3. Subtract the mean from each of the mid points to get deviation scores (x -x.)
4. Square all the deviation scores $(x - x)^2$
5. Multiply the squared deviation scores with their respective frequencies $= F(x - x)^2$
6. Sum step 5 and divide by number to get variance.
7. Calculate the square root of the variance in step 6 for standard deviation.

Example: Given the distribution of scores below, the S^2 and S are calculated as:

Table 20: Computation of Variance and Standard Deviation for a Grouped Data Using the Mean Deviation Formula

Consider the data below
Table 16

X	**F**	**FX**	**X - $\bar{X}$**	**(X - $\bar{X}$)^2**	**F(X - $\bar{X}$)^2**
10	2	10	4.2	17.64	35.28
8	5	40	2.2	4.84	24.2
6	12	72	0.2	0.04	0.48
4	4	16	-1.8	3.24	12.96
2	4	8	-3.8	14.44	37.76
	27	156			130.68

Variance $\frac{f(x-\bar{x})^2}{N}$

$= \frac{130.68}{27}$

$= 4.84$

$\delta^2 = \sqrt{4.84}$

$\delta^2 = 2.2$

Example 2

Consider the table below and determine the standard deviation.

C. I	5 - 15	16 - 26	27 - 37	38 - 48	49 - 59	60 - 70
Frequency	2	3	4	7	5	8

Solution

C. I	Frequency	X	FX	X - X	$(X-X)^2$	$F(X-X)^2$
5-15	2	10	20	-349	1218.01	2436.02
16-26	3	21	63	-23.9	571.21	1713.63
27-37	4	32	128	-12.9	166.41	665.64
38-48	7	43	301	-1.9	3.61	26.27
49-59	5	54	270	9.1	82.81	414.05
60-70	8	65	520	20.1	404.01	3232.08
	Total = 29		Total = 1302			Total = 8487.69

Step 2. Determine the mean (x)

Mean (x) = $\frac{\Sigma FX}{\Sigma F}$

$= \frac{1302}{29}$

= 44.90

$S^2 = \frac{\Sigma f(x-x)^2}{\Sigma f}$

$S^2 = \frac{8487.69}{29}$

S^2= 292.68

S. D = $\sqrt{292.68}$

S. D = 17.12

Example 3

Class Interval	***F***	**Mid-Point (x)**	***fx***	***x – x***	**(x – x)^2**	***f*(x – x)^2**
81 – 91	5	86	430	32	1024	5120
70 – 80 59 – 69	10 12	75 64	750 768	21 10	441 100	4410 1200
48 – 58 37 – 47	18 11	53 42	954 462	-1 -12	1 144	18 1584
26 – 36 15 – 25	6 6	31 20	186 120	-23 -34	529 1156	3174 6936
	∑f = 68	***∑fx = 3670***				***∑f(x – x)^2 = 22442***

$X = 54$

$S^2 = \frac{\sum f(x - x)^2}{N}$ $\qquad$ $S = \sqrt{\frac{\sum f(x - x)^2}{N}}$

$S^2 = \frac{22442}{68} = 330.03$ $\qquad$ $SD = \sqrt{\frac{22442}{68}}$

$X = 18.17$

Assumed Mean Formula for Calculating the Standard Deviation of a Grouped Data where deviation is taken from arbitrary origin.

Procedure

1. Determine the frequencies and class mid points.
2. Start from an arbitrary origin and determine the deviations of the scores = d
3. Multiply each of the deviations with respective frequencies =fd.
4. Multiply each fd with respective d = fd^2.
5. Sum the column titled fd and fd^2
6. Apply the formula as in example given below.

Class Interval	***f***	**x**	**d**	***Fd***	**Fd^2**
81 – 91	5	86	3	15	45
70 – 80	10	75	2	20	40
59 – 69	12	64	1	12	12
48 – 58	18	53	0	-0	0
37 – 47	11	42	-1	-11	11
26 – 36	6	31	-2	-12	24
15 – 25	6	20	-3	-18	54
	68			**6**	**186**

$$\delta^2 = c\sqrt{\frac{\sum Fd^2}{N} - \left(\frac{\sum Fd}{N}\right)^2}$$

Where:

i = interval width or class width

d = deviation of scores from arbitrary origin

N = number of scores

$$\delta^2 = 11\sqrt{\frac{186}{68} - \left(\frac{6}{68}\right)^2}$$

$$11\sqrt{2.728}$$

S^2 = 18.17^2 = 330.01

Table 21: Computation of S^2 and SD Using Assumed Mean Formula

C.I	F	X	d	Fd	Fd^2
82-90	2	86	3	9	18
73-81	6	77	2	12	24
64-72	10	68	1	10	10
55-63	12	59	0	0	0
46-54	18	50	-1	-18	18
37-45	6	41	-2	-12	24
28-36	8	32	-3	-24	72
19-27	4	23	-4	-16	64
10-18	2	14	-5	-10	50
				-52	280

$$\delta^2 = c\sqrt{\frac{\sum Fd^2}{N} - \left(\frac{\sum Fd}{N}\right)^2}$$

$$\delta^2 = 9\sqrt{\frac{280}{68} - \left(\frac{-52}{68}\right)^2}$$

$$9\sqrt{4.12 - 0.58}$$

$$9\sqrt{3.54}$$

9 x 1.88

= 16.93

Computation of S^2 and SD Using the Raw Score Formula

Procedure

1. Determine the class midpoint (x) of the different class intervals.
2. Get the product of f and x = fx.
3 Determine fx" by multiplying the x of each class interval with their respective fx's.
4. Sum fx^2 and fx.
5. Apply the formula.

$$S^2 = \frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2$$

$$SD = \sqrt{\frac{\Sigma Fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$

Table 22: Computation of S2 and SD for Grouped Data with Raw Score Formula

Consider the table below

Table 15.

Classes	F	X	FX	FX^2
82-90	2	86	172	14792
73-81	6	77	462	35574
64-72	10	68	680	46240
55-63	12	59	608	35872
46-54	18	50	900	45000
37-45	6	41	246	10086
28-36	8	32	252	8192
19-27	4	23	92	2116
10-18	2	14	28	392
	68		3544	204164

$$\delta^2 = \frac{\Sigma FX^2}{N} - \left(\frac{\Sigma FX}{N}\right)^2$$

$$\delta^2 = \frac{204164}{68} - \left(\frac{3544}{68}\right)^2$$

$3002.41176 - (52.1176471)^2$

3002.41176 – 2716.24914

= 286.16262

s.d = $\sqrt{286.166}$

s.d = 16.92

Example 2

C. I	5 - 15	16 - 26	27 - 37	38 - 48	49 - 59	60 - 70
Frequency	2	3	4	7	5	8

Solution

C. I	Frequency	X	FX	FX^2
5 – 15	2	10	20	200
16 – 26	3	21	63	1323
27 – 37	4	32	128	4096
38 – 48	7	43	301	12943
49 – 59	5	54	270	14580
60 - 70	8	65	520	33800
	Total = 29		Total = 1302	= 66942

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

$S^2 = \frac{66942}{29} - \left(\frac{1302}{29}\right)^2$

S^2 = 2308.34 – $(44.90)^2$

S^2 = 2308.34 – 2016.01

S^2 = 292.33

S. D = $\sqrt{292.33}$

S. D = 17.10

EXAMPLE 3

Class Interval	***F***	**X**	**fx**	***fx²***
81 – 91	5	86	430	36980
70 – 80	10	75	750	56250
59 – 69	12	64	768	49152
48 – 58	18	53	954	50562
37 – 47	11	42	462	19404
26 – 36	6	31	186	5766
15 – 25	6	20	120	2400
	68		**3670**	**220514**

$$S^2 = \frac{220514}{68} - \left(\frac{3670}{68}\right)^2 = 330.03$$

$$SD = c\sqrt{\frac{220514^2}{68} - \left(\frac{3670}{68}\right)^2} = 18.17$$

From the computation above, we can see that the raw score formula involves use of large scores which can be very tedious for a large distribution. In such cases, we can use the other formulae already described above which are more convenient.

Note: Variance is the square of standard deviation. So we can get the \ variance in the example above by squaring the standard deviation f calculated.

The Uses of different measures of variability

a. The Range

This is used:

i. When the quickest index of dispersion or variability is wanted.

ii. When information about the two extremes of the scores is taken into account

iii. When the data are too scanty or too scattered to justify the computation of a more precise measure of variability

iv. When knowledge of extreme scores or of total spread is all that is wanted (Garrett, 1966).

b. Semi-Interquartile Range or Quartile Deviation (QD)

This is used:

i. When the extreme scores are likely to influence the dispersion very much.
ii. Whenever the median score is reported
iii. Whenever the extreme scores are truncated or incomplete
iv. When the actual score limits of the middle 50% is needed.

c. Average Deviation (AD)

The average deviation is used:

i. When calculating the standard deviation, squaring the extreme scores would bias the mean
ii. When it is desired to weight all deviations from the mean according to their size (Garret, 1966).

d. Standard Deviation (SD)

This is the most dependable value and helps in further computation like the Z score. It is used:

i. When the statistic having the greatest stability is sought
ii. When extreme deviations should exercise a proportionally greater effect upon the variability
iii. When coefficients of correlation and other statistics are subsequently to be computed (Garrett, 1966, P.60).

The Coefficient of Variation

When dealing with positive observations, two groups which have completely different mean values or are measured in completely different units can be compared by calculating the coefficient of variation in each case. This is defined by:

$$C.V = \frac{SD}{X}$$

Or $$C.V = \frac{100 \times SD}{X} \%$$

The above two definitions mean that the standard deviation is expressed either as a fraction or as a percentage of the mean. The coefficient of variation is independent of the units in which the mean and standard deviation are measured. The standard deviation is an absolute measure, the coefficient of variation a relative measure of dispersion (Loveday, 1970).

Using the computation of mean and standard deviation above, the coefficient of variation (C.V.) is:

$$C.V. = \frac{18.17}{54} = 0.36$$

Or

$$C.V. = \frac{100 \times 18.17}{54} = 33.6\%$$

Exercise

1. Calculate the mean deviation and quartile deviation for the distribution of scores below:

Scores	**F**
25	1
24	2
23	6
22	8
21	5
20	2
19	1
	N = 25

2. Given the following grouped data, calculate the variance and standard deviation using both mean deviation and assumed mean formulae.

X =	10-16	17-23	24-30	31-37	38-44.
F =	5	12	12	14	4

Measures of Relative Standing

When we want to accurately interpret the score or data collected concerning any phenomenon of interest, when comparing the standing of a score among a group of scores, raw scores (data) can be quite misleading in such cases. For precise and accurate interpretation to be done there is a need to transform the raw data or scores obtained. The measures which are used to carry out this transformation are called measures of relative or positional standing.

When the raw scores or data are transformed with these measures we obtain what is called 'derived' scores or data. This name, 'derived' refers to the fact that the scores now obtained are not the raw scores but derived from the raw scores.

Derived scores give the true picture of the standing of a score or variate in a distribution and lead to accurate and better interpretation. For example, suppose a student, Emeka, obtains a raw score of 80 in an English test and 45 in a mathematics test, the impression given by the raw scores is that of better performance in English than in Mathematics. This may, however, not be correct. To be able to interpret Emeka's performance in the two tests appropriately, the scores will be transformed with measures of relative standing. This transformation gives us the true position of Emeka in the two tests and enables us to derive the true scores from the raw scores given above.

There are many measures of relative standing, which can be used to transform scores. These are:

- Simple Ranks.
- Raw Score Letter Grades.
- Grade Point Average.
- Percentile Ranks.
- Standard Scores: Z and T scores.
- Stanines.

Simple Ranks

Simple ranks refer to awarding positions to raw scores of students or to any data collected. This is a situation where one gives 1^{st}, 2^{nd}, 3^{rd} ... positions to scores or data according to their order of magnitude.

Example: Given a distribution of scores below, positions can be assigned to them in this manner.

X	**Ranks**
80	1
71	2
65	3
65	3
44	6
50	5

We will notice that in the example above two students obtained the same score. In such cases of ties, the same position is assigned to the scores either in whole number as in our example, or by adding the positions occupied by the scores and dividing by number. Whenever the ties are assigned a position, like in our example where the 2 scores of 65 occupied 3^{rd} and 4^{th} positions but are awarded 3^{ld} position each, the 4^{th} position will no more be awarded to another score but jumped.. An alternative way of ranking identical scores is by adding the positions they will occupy and dividing by then number. That is, assigning them the mean of the positions they will occupy. In the distribution above, the two students that scored 65 may be ranked 3.5 respectively while the next score takes the 5^{th} position.

Raw Score Letter Grade

In some cases, rather than using ranks to transform scores, letter grades are used instead. This transformation method is common in institutions of higher learning and in secondary schools especially in external examinations. This is a situation where students' raw scores are assigned some letter grades like 'A', 'B', 'C' ... according to particular criteria. In the Nigerian universities, the grades are assigned in this manner:

Raw Scores		**Grade**	**Grade Point**
70-100	=	A	5.00
60- 69	=	B	4.00
50-59	=	C	3.00
45-49	=	D	2.00
40 - 45	=	E	1.00
0-39	=	F	0.00

Grade Point Average (GPA)

Grade point average is a measure of relative standing commonly used in tertiary institutions in computing semester, sessional and final results of students. This measure is calculated with credit load units and grade points. It first transforms the raw scores .to grades which are weighted with some points as in Raw score letter grade above. The grade points are multiplied with the respective credit units, summed and divided with the total credit load of the courses taken by the student.

For example, suppose a student obtained the following results in a semester, the GPA is computed thus:

Table 23: Computation of GPA of a Student in a Semester

Course Code	**Course Title**	**Credit Unit**	**Grade**	**Grade Point**	**Quality Grade Point**
EDU 212 EDU 221	Intro. To Educational Psychology Fun. Of Curriculum Development	3 3	B A	4.00 5.00	12.00 15.00
GSS 103 GSE 101	Logic and Philosophy Use of English	2 2	F D	0.00 2.00	0.00 4.00
EDA	Management of Conflict	3	B	4.00	12.00

246 EDA 241	Leadership in Administration	3	C	3.00	9.00

GPA = $\frac{52.00}{16} = 3.25$

Refer back to 'Raw Score Letter Grade' and see how the grades are weighted. The weight of each grade is multiplied with the respect credit unit.

Percentile Ranks

Students' raw scores can be transformed to derived scores by determining the percentage number of scores surpassed by the student in a test. This gives the true position of the student in the test. The percentage number of scores that are below a particular score is called the percentile of that score. This means that Nneka scored '64' in a test; we will not know the true standing of Nneka in that test until we transform the score '64' to a derived score. This we may do by calculating the percentage of the scores below Nneka's score of 64. This may give us 48. Score 48 is now Nneka's percentile score which tells us that the actual standing of Nneka in that test is 48 and not 64. 64 is the raw score while 48 is the percentile score.

Percentile Rank for Ungrouped Data

Given the distribution of scores below, calculation becomes:

$$PR_X = \frac{.5f+Cfb}{N'} \times \frac{100}{1}$$

Where PR_X = Percentile rank of any score of interest

.5f or $f^1/_2$ = $^1/_2$ of frequency of the score of interest => Number of scores in a distribution

Cfb = Cumulative frequency of the score below that of the score of interest

Table 24: Computation of Percentile Rank for Ungrouped Data

X	**F**	**Cf**
88	2	28
72	4	26
60	3	22
55	6	19
46	5	13
40	4	8
32	4	4
	28	

Taking score 72 for example, the percentile rank becomes:

$$PR_{72} = \frac{4/2 + 22}{28} \quad x \quad \frac{100}{1} = 85.7 = 86$$

Note: For practice, you can calculate the percentile ranks of other scores in the distribution above.

Percentile Rank computation for Grouped Data

To determine the percentile rank of a grouped data, we apply the formula below:

$$PR_X = \left(\frac{Cfb + \left(X - L \right)/c \; f}{N} \right) 100$$

Where;

PR_x = Percentile rank of the score of interest.

Cfb = Cumulative frequency below that of the class interval containing the score (X) of interest.

X = the score of interest

c = width of the class interval.

N = number of scores, in the distribution.

Using the distribution in Table 22, the percentile rank of score 58 is computed thus:

Table 25: Computation of Percentile Rank of a Grouped Data

Classes	**F**	**Cf**
81 - 91	5	68
70 – 80	10	63
59 – 69	12	53
48 – 58	18	41
37 – 47	11	23
26 – 36	6	12
15 – 25	6	6
	68	

Step 1: Determine the cf below that of the class containing the score of interest (X). To do this, build a cumulative frequency column in your table as is done above. Then get the cfb of X. For our example, score 58 is located in the class 48 - 58. The cumulative frequency below that of class 48-58 is 23.

Step 2: Subtract the lower class limit of the class containing the score of interest from the score itself.

58 - 47.5 = 10.5

Step 3: Determine the class width of the class containing X (i.e. 58). For our analysis above, the class width is 11.

Step 4: Divide the result in step 2 with the class width. $^{10.5}/_{11} = 0.955$.

Step 5: Multiply the result in step 4 with the f of the score of interest.

0.955 x 18 = 17.19.

Step 6: Add the result in step 5 to the cfb obtained in step 1.

17.19 + 23 = 40.19

Step 7: Divide the result in step 6 with the number of scores (N).

$^{40.19}/_{68} = 0.591$.

Step 8: Multiply the result in step 7 with 100 to obtain your PR_{58}.

$$PR_{58} = \left(\left(23 + \left(\underline{58 - 47.5} \right) 18 \right) \right) 100$$

$$\frac{11}{68}$$

Standard Scores

Standard scores are the measures used to standardize raw scores or data obtained concerning any variable. These scores can be applied where one wants to know the relative standing of a student's score in a distribution. The scores are Z and T scores.

Z score is defined as the distance of a score from the mean as measured by standard deviation units.

$$Z = \frac{X - Y}{SD}$$

Where; X = Score of interest

X = Mean of the distribution

SD = Standard deviation of the distribution

Suppose a student obtains the following scores in two exams with the respective means and standard deviations as:

Geography	**Statistics**
X = 72	X = 58
X = 66	X = 35
SD = 10	SD = 12

The Z scores become:

$$Z = \frac{72 - 66}{10} = 0.6; \qquad \frac{58 - 25}{12} = 1.92$$

The Z scores obtained for the two subjects indicate a better performance in statistics than in geography which are opposites of what the raw scores portray.

Z score has the disadvantages of negative signs and decimals which makes it difficult to be properly interpreted. To avoid negative values, Z scores can be transformed to T scores.

T score is defined as a score with a mean of 50 and a standard deviation of 10.

T = 10Z + 50

We can use T score to transform the scores above. From the formula of T score given above, we can see that before we transform a score to T standard measure, we will first obtain the Z value and then go further to T score. For the example above, T score for Geography becomes:

10 x 0.6 + 50= 56.

T score for statistics = 10 x 1.92 + 50 = 69.2 approximately 69. With the T scores above, we can easily and better interpret the relative performance of the students in geography and statistics.

Stanine

Stanine is a short form of standardized nine or standard nine. This measure standardizes raw scores by appending numbers 9 - 1 to a particular percentage of the scores in a descending order. The percentages are constant and by affixing numbers 9 - 1 to them, Stanine at the end normalizes the distribution as we can see below:

Stanine	9	8	7	6	5	4	3	2	1
Percentages	4%	7%		12%	17%		20%	17%	
12%	7%	4%.							

The information above implies that Stanine 9 is assigned to the top 4% of scores in a distribution, Stanine 8 to the next 1% of scores in the distribution and so on. We must note that the percentages of scores to which each stanine number is assigned is constant. To calculate the %'s of the scores to which Stanine is assigned, the scores are first arranged in order of magnitude. The percentage of scores relative to each Stanine is now determined. For example, suppose a teacher gave 20,000 students a test on English Language, marks the test and arranges the scores in order of magnitude, to assign Stanine to these scores, the teacher will calculate the 4%, 7%, 12%, etc of the scores and assign Stanine numbers accordingly.

For Stanine 9, we first calculate 4% of 20000 scores which is

$\frac{4}{100} \times \frac{20000}{1} = 800$

Stanine 9 is awarded to the top 800 scores in the distribution

For Stanine 80 which is 7% of the scores, the number of the scores involved

$= \frac{7}{100} \times \frac{20000}{1} = 1400$

Others are computed thus:

Stanine 7 $= \frac{12}{100} \times \frac{20000}{1} = 2400$

Stanine 6 $= \frac{17}{100} \times \frac{20000}{1} = 3400$

Stanine 5 $= \frac{20}{100} \times \frac{20000}{1} = 4000$

Stanine 4 $= \frac{17}{100} \times \frac{20000}{1} = 3400$

Stanine 3 $= \frac{12}{100} \times \frac{20000}{1} = 2400$

Stanine 2 $= \frac{7}{100} \times \frac{20000}{1} = 1400$

Stanine 1 $= \frac{4}{100} \times \frac{20000}{1} = 800$

Table 25: Computation of Stanine

Stanine	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
Percentages %	**4%**	**7%**	**12%**	**17%**	**20%**	**17%**	**12%**	**7%**	**4%**
No. of scores	**800**	**1400**	**2400**	**3400**	**4000**	**3400**	**2400**	**1400**	**800**

The table above indicates that the top 800 scores in the test are assigned Stanine 9; next 1400 scores, Stanine 8; next 2400 scores, Stanine 7, etc. Stanine is a very convenient way of transforming scores when the data is very large.

Many external examining bodies use stanine to present the scores of candidates. A notable example is West African Examination Council (WAEC) which uses an inverted stanine to present examinee results. It is inverted because the normal stanine measure assigns 9 to the top 4% of the scores and 1 to the lowest 4% of the scores but WAEC inverts this procedure by assigning 1 to the top 4% of the scores and 9 to the lowest 4% of the scores.

This is why WAEC presents results in this manner.

Al = Stanine 1

B2 = Stanine 2

C3 = Stanine 3

C4 = Stanine 4

C5 = Stanine 5

C6 .= Stanine 6

D7 = Stanine 7

D8 = Stanine 8

F9 = Stanine 9

Exercises

1. Calculate the grade point average of your first and second semester examinations in the past session.
2. Give a better interpretation of the performance of Ogechi in the three tests whose distributions are presented by calculating her percentile ranks in the test

Test A		**Test B**		**Test C**	
X	**F**	**X**	**F**	**X**	**F**
76 68	6 7	88 84	8 6	92 78	2 8
52	4	82	2	66	9

33	8	66	10	58	4
20 18	5 3	52 45	2 4	44 30	2 2

Ogechi's scores in the tests are 52 for Test A, 45 for Test B and 78 for Test C.

3. With the data below, calculate the Z and T scores of student X. Give the appropriate interpretation of the student's relative standing in the distributions.

		A	**B**
Score	=	62	48
Mean	=	78	34
Standard deviation	=	15	10

4. Suppose you, as a researcher or a teacher collected data of 3250 students' test scores, show how you can use stanine to transform the raw scores to standard scores.

CHAPTER FOUR

MEASURES OF RELATIONSHIP (ASSOCIATION) OR CORRELATION ANALYSIS

The measures of central tendency (e.g. mean, median, mode) and variability (range, standard deviation etc) discussed in the preceding chapters dealt with the description of a single variable. In most cases, interest may center on the strength of relationship which exists between variables that is on how well the variables are correlated. This means that one may wish to discover if any relationship exists between two or more variables, how strong the relationships appear to be and whether one variable of primary interest can be effectively predicted from information on the values of the other variables (Johnson & Bhattacharyya, 1996). Correlation analysis provides such information. In correlation analysis we consider two variables or two sets of scores or measurements (bivariate data). For example, suppose there are marks obtained by a class in English Language and those obtained in Mathematics and the scores in the bivariate data are paired for each student, the study of the two sets of results helps in the determination of the degree of relationship between them. It may also help us predict, knowing the results, in say English Language and the results in Mathematics. Correlation analysis provides a simple summary description of the degree of relationship between the two sets of variables.

We may also wish to:

I. know the degree of association or relationship between the performance of students in any school subject and their gender or

II. determine the degree of relationship between achievement test scores of students and their aptitude tests. If the two variables in the above two cases are highly related, interpretations will be done accordingly. .

This means that for the first case, it will be taken that performance of students in the subject of interest has something to do with their gender and in the second case, aptitude tests can be used to determine the achievement performance of students,

In correlation analysis, interest centers on the direction and strength of the linear relationship between the variables, that is, on how the variables are correlated, whether positively or negatively, whether the correlation is high or low. The assumptions of two-variable correlation model are as follows:

I. Both X and Y are random variables.

II. Both X and Y are normally distributed. The two distributions need not be independent.

III. The standard deviation of the Y's are assumed to be equal for all values of X, and the standard deviations of X's are assumed to be equal for all values of Y (Hamburg, 1983).

The Correlation Coefficient (C.C)

The statistical procedure for establishing the degree of relationship between two or more variables is what, we refer to as correlation coefficient. The correlation coefficient, therefore, according to Nworgu (1992) "is a statistic which provides a measure of the extent or degree of association between two or more variables."

Correlation coefficient takes values of +1 through 0 to -1. A value of -I describes perfect negative correlation, all points lie on straight line. A value of +1 describes a perfect positive correlation, here all points also lie on straight line i.e. one variable is independent of another. We do not always get correlations that are perfect, so values of correlation coefficient may be 0.75, 0.21, 0.62, etc, Correlation coefficients do not take any value above +1 or below -1. This means that one cannot obtain a C.C. of 1.16 or -1.32 or example. The decisions as to the acceptable value which will enable us describe bivariate data as showing correlation depends on some other statistical calculations.

However, that two variables are correlated does not necessarily imply that one variable is the cause of the other. Their correlation may as well have been due to the fact that the two variables are related to a third variable. For example, the performance of students in Mathematics and their performance in English may correlate not because performance in one subject causes performance in the other, but because performances in both subjects are related to intelligence (Nworgu, 1992). The relationship between test results can be depicted diagrammatically or expressed statistically in quantitative terms. The diagram presentation is known as the scatter diagram (or scatter graph or scatter plot).

The Scatter Diagram or Scatter Graph

The scatter diagram, in this context, gives a visual picture of the closeness of the relationship between two sets of test scores which have been plotted on it, by showing how scattered, spread out or dispersed the scores are. A scatter diagram has two axis to which are affixed the two sets of scores that are being con-elated. It is a kind of graph. In plotting the scatter diagram, the position of each point (dot) is established by locating on the horizontal axis each pupil's test, score.in the first test and then moving-upward, parallel .to the vertical axis, until the score of that same pupil in the second test is reached and a dot is made at this point of intersection. Let us consider the two sets of results in case 1, case 2, and case 3. The marks for each student are paired.

Case 1			**Case 2**			**Case 3**		
Pupil	**English**	**Maths**	**Pupil**	**English**	**Maths**	**Pupil**	**English**	**Maths**
Glad	9	20	Paul	15	6.4	Eddy	15	2
Joy	8	19	James	14	6.5	Jim	14	0
Nnenna	7	18	Usman	13	6.6	Sam	13	4
Martha	6	17	Sampson	12	6.7	Emma	12	3
Florenc	5	16	Aaron	11	6.8	Uju	11	1

e								

The relationship of the marks given above may be examined by plotting, using suitable scales for X and Y axis, the paired scores on a graph paper, each pair of scores being represented by a point.

Inspection of the scatter graph gives a layman an idea of the degree of relation between two variables. If you plot the graph, you will get for:

Case 1	**Case 2**	**Case 3**
Linear perfect	Linear perfect	Random Correlation
Positive	Negative	
Correlation	Correlation	
Correlation Coefficient = 1	Correlation Coefficient = -1	Correlation coefficient = 0

CASE 1 **CASE 2** **CASE 3**

9 **15** **15**
8 **14** **14**
7 **13** **13**
6 **12** **12**
5 **11** **11**
16 17 18 19 20 **6.0 6.4 6.5 6.7 6.8** **6.0 6.4**
6.5 6.7 6.8

Between the two cases of perfect positive correlation and the perfect negative correlation is a large number of other possible arrangements in which the correlation is not perfect.

A positive correlation indicates that the scores in one variable increase or decrease as the scores in the other variable increase or decrease. A zero correlation indicates no consistent relationship.

Exercises

I. How would you interpret the following values of correlation ' +0.82, +0.25, -0.6, 0.9.

II. Can we ever have a correlation coefficient of +1.25? Explain.

III. Draw a scatter diagram using the following two tests taken by 10 pupils.

Test A:	4	6	8	10	9	8	5	6	9	10
Test B:	**5**	**7**	**8**	**9**	**10**	**7**	**6**	**6**	**10**	10

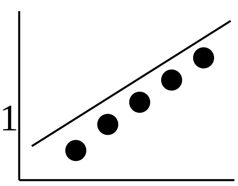
Methods of Calculating Correlation Coefficient (c.c)

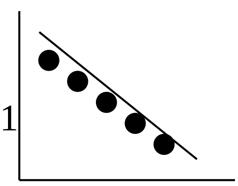
Various methods have been derived to calculate correlation coefficients for various types of data. A description of the most popular, common method which is the Person Product-moment correlation coefficient 'r' will be given in this chapter coupled with a detailed discussion of the special con-elation methods which are applicable in special situations. These special correlation methods are:

i. Spearman Rank-Ordered correlation coefficient
ii. Correlation Ratio.
iii. Biserial r correlation coefficient.
iv. Point Biserial r correlation coefficient.
v. Partial con-elation coefficient.

The Pearson Product Moment Correlation Coefficient

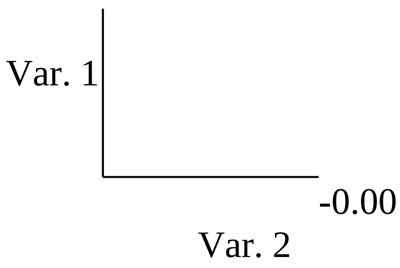
This is the most frequently used correlation coefficient introduced by Karl Pearson in 1896 and denoted by 'r'. It is a measure of the degree of linear relationship between two variables. The value of Pearson r always lies between -1.0 and +1.0. A Pearson r of - 1 indicates a perfect negative relationship, + 1 indicates perfect positive relationship while 0.00 indicates no relationship at all. This can be represented diagrammatically thus:

Var. 1 

Var.1 

r =   r = - 100

Var. 2   Var. 2



Source of Diagram: Nworgu, B. (1992)

Before Pearson r could be computed, the basic condition of linearity of relationship between the variables has to be satisfied. This is because, to the extent that there is a departure from linearity, the two variables. Construction of a scatter gram like the one done above helps to reveal the relationship between two variables being considered.

Pearson r can be computed using different methods which are:

a, Standard Score Method

b. . Mean Deviation Method

c.. Raw Score Method

d. Difference Method

We are going to use two of these methods in calculating Pearson r in this chapter. These are the Mean Deviation Method and the Raw Score Method.

Mean Deviation Method for Calculating Pearson r

The following steps are used in calculation of Pearson r using the Deviation Method.

1. List the two scores X and Y for each student.
2. Determine the deviation scores (x) by subtracting the mean (X) from each score.
3. Do same as above for Y to get deviation scores (y).
4. Square each x i.e. $(X-X)^2$ or x^2

5. Do same for each y deviation.
6. Find x, y products for each pair and find their sum = (X-X~) (Y-X) or xy.
7. Find r by applying the formula.

$$r = \sqrt{\frac{\sum(X - X) \quad (Y - Y)}{\sum(X - X)^2 \quad (y - y)^2}} \quad \text{or} \quad \sqrt{\frac{\sum xy}{\sum x^2 \sum y^2}}$$

We may set out the work in a form, using the scores in English Language and History.

Data For Computation of C. C.

Pupils	A	B	C	D	E	F	G	H	I	J
Scores in English	4	5	6	7	11	15	19	23	24	25
Scores in History	6	7	8	9	13	17	21	25	26	26

Table 26: Calculation of Pearson r C.C. Using Mean Deviation Method

X	**X – X or (x)**	**$(X – X)^2$ or $(x)^2$**	**Y**	**Y – Y or (y)**	**$(Y – Y)^2$ or $(y)^2$**	**(X – x) (Y – y) or (xy)**
4 5	-10.7 -9.7	114.49 94.09	6 6	-10.6 -10.6	112.36 112.36	113.42 102.82
6 11	-8.7 -3.7	75.69 13.69	9 13	-7.6 -3.6	57.76 12.96	66.12 13.32
15 19	0.3 4.3	0.09 18.49	17 21	0.4 4.4	0.16 19.36	0.12 18.92
23 24	8.3 9.3	69.89 86.49	25 26	8.4 9.4	70.56 88.36	69.72 87.42
25	10.3	106.09	26	9.4	88.36	96.82
132		**578.01**	**149**		**562.24**	**568.68**

(Mean of X) X = 14.7; Y = 16.6 (Mean of Y)

$$r = \sqrt{\frac{568.68}{578.01 \times 562.24}} = 100$$

Note that in this formula, we have not used any standard deviation values. If we had the mean of each score X and Y and the standard deviations of X and Y, we would have used another formula to find the 'coefficient of correlation and this would serve as a check.

The Pearson product moment correlation coefficient presented above can be calculated with other formulae. We will now use one of the most popular formula to calculate it again. This is the raw score formula.

Raw Score Formula for Calculating Pearson Correlation Coefficient

Procedure

1. Build five columns with headings X, Y, XY, X^2,'and Y^2 respectively.
2. List the scores of one variable under column X, and the scores, of the other variable under column Y.
3. Get a product of X and Y = XY.
4. Square all the X = X^2.
5. Square all the Y = Y^2.
6. Sum all the five columns and apply the formula =

$$r = \frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$$

Example: Using the data in Table 26 above, the Pearson r is calculated thus with raw score formula:

Table 27: Calculation of Pearson r Using Raw Score Method

Example 1:

Consider the table below being the data obtained from the scores of female and male students in mathematics' test:

X	6	5	8	6	7	5
Y	5	7	9	4	3	1

Solution:

S/N	**X**	**Y**	**XY**	**X^2**	**Y^2**
1	6	5	30	36	25
2	5	7	35	25	49

3	8	9	72	64	81
4	6	4	24	36	16
5	7	3	21	49	9
6	5	1	5	25	1
	35	29	187	235	181

Pearson r $= \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$

$r = \frac{6(187) - (35 \times 39)}{\sqrt{[6(235) - (35)^2][6(181) - (29)^2]}}$

$r = \frac{1122 - 1073}{\sqrt{(1410 - 1369)(1086 - 84)}}$

$r = \frac{49}{\sqrt{(41)(245)}}$

$r = \frac{49}{\sqrt{10045}}$

$r = \frac{49}{100.224747}$

r = 0.49 The relationship is a positively weak one.

Example 2:

Consider the table below, determine and interprete the correlation index.

X	5	3	2	4	3
y	9	4	7	8	6

Solution

	x	Y	x2	y2	xy
	5	9	25	81	45
	3	4	9	16	12
	2	7	4	49	14
	4	8	16	64	32
	3	6	9	36	18
Total	**17**	**34**	**63**	**246**	**121**

$\Sigma x^2 = 63$

$(\Sigma x)^2 = 289$

$\Sigma y^2 = 246$

$(\Sigma y)^2 = 1156$

N = 5

$r = \frac{5(121) - 17 \times 34}{\sqrt{[5(63) - 17^2)][5(246) - (34)^2]}}$

$r = \frac{605 - 578}{\sqrt{(315 - 289)(1230 - 1156)}}$

$r = \frac{27}{\sqrt{(26)(74)}}$

$r = \frac{27}{\sqrt{1924}} = \frac{27}{43.86}$

r = 0.616

Example 3

The table below was extracted from the scores of student taught in the urban and rural area after the same test has been administered.

Urban (X)	20	60	85	72	45	50	32
Rural (Y)	33	52	62	40	81	30	71

Using Pearson's correlation, is there any relationship between the scores.

Solution

X	Y	X^2	Y^2	XY
20	33	400	1089	660
60	52	3600	2704	3120
85	62	7225	3844	5270
72	40	5184	1600	2880
45	81	2025	8561	3645
50	30	2500	900	1500
32	71	1024	5041	2272
364	**369**	**21958**	**23739**	**19347**

$\Sigma X = 364$, $\Sigma Y = 369$, $\Sigma X^2 = 21958$, $\Sigma Y^2 = 23730$, $\Sigma XY = 19347$, $N = 7$

$r = \frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[N(\Sigma X^2) - (\Sigma X)^2][N(\Sigma Y^2) - (\Sigma Y)^2]}}$

$r = \frac{7(19347) - (364 \times 369)}{\sqrt{[7(21958) - (364)^2][7(23739) - (369)^2]}}$

$r = \frac{135429 - 134316}{\sqrt{(153706 - 132496)(166173 - 136161)}}$

$r = \frac{1113}{\sqrt{(21210)(30012)}}$

$r = \frac{1113}{\sqrt{636554520}}$

$r = \frac{1113}{25230.03}$

r = 0.0441

r = ≈0.04

The interpretation is that the relationship is positive but very low.

	X	Y	Xy	X^2	Y^2
A B	4 5	6 6	24 30	16 25	36 36

C	6	9	54	36	81
D	11	13	143	121	169
E	15	17	255	225	289
F	19	21	399	361	441
G	23	25	575	529	625
H	24	26	624	576	676
I	25	26	650	625	676
	135	**149**	**2754**	**2514**	**3029**

$$r = \frac{N\sum XY - \sum X \sum Y}{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}$$

$$r = \frac{9 \times 2754 - 132 \times 149}{(9 \times 2514 - (132)^2 \quad (9 \text{ X } 3029 - (149)^2)}$$

$$r = \frac{24786 - 19668}{(22626 - 17424)\ (27261 - 22201)} = 100$$

Interpretation: The relationship between X and Y above is a perfect positive one.

Special Correlation Methods

Pearson Product - moment coefficient is the standard index of the amount of correlation between two variables, and we prefer it whenever its use is possible and convenient. Pearson r may not be applicable to some data and in some cases where it is applicable, other procedures may be more expedient for practical purposes (Guilford, 1965). Some of these situations where Pearson r may not be applicable according to Guilford are:

1. Where two variables cannot be measured on a continuous metric scale.
2. In situations where the regression of the variables are not linear.
3. Heterogenous samples.
4. Samples restricted in variability or forced into a smaller number of categories than is needed for good estimation of correlation.

In all the cases, other methods of determining correlation must be employed. These methods are special correlation coefficient methods, and they are presented below:

The Spearman's Rank - Difference Correlation Coefficient or Spearman's Rho Correlation Coefficient

There are cases where statistical series comprise items for which the exact magnitude cannot be ascertained but which are ranked according to size. To determine the correlation between the two variables of interest, Charles Spearman developed a measure used for determining the degree of rank association between two variables X and Y (Nwabuokei, 1986). This measure is based on the ranks (or order) of the observations and does not depend on the specific observations of x and y. That is, instead of using the actual marks in calculation of c.c, the marks are ranked and the ranking order used.

The formula for computation of Spearman's rank correlation coefficient is:

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

Where;

d = the difference between the correlative ranks of each set of paired X, Y values,

N = number of paired observations.

Procedure for Calculation

1. Rank data for each group giving rank "1" to the highest score.
2. Obtain the differences (d) between each pair of ranks.
3. Square each of the differences (d^2)
4. Add all the squares together (£d^2).
5. Calculate n, where n is the number of pairs of scores,
6. Find Rho or f_s.

Example 1

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

X	40	50	60	40	58	55	59	20
Y	65	80	80	85	70	82	65	50

The data above are scores obtained from Jeba International College Wukari, showing the scores of male and female students in an examination. Using Spearman's ranking order correlation, ascertain the correlation index.

Solution:

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

S/N	X	y	Rx	Rx	D	d^2
1	40	65	6.5	6.5	0	0
2	50	80	5	3.5	1.5	2.25
3	60	80	1	3.5	-2.5	6.25
4	40	85	6.5	1	5.5	30.25
5	58	70	3	5	-2	4
6	55	82	4	2	2	4
7	59	65	2	6.5	4.5	20.25
8	20	50	8	8	0	0
						67

rho = 1 – $\frac{6(67)}{8(8^2-1)}$

= 1 - $\frac{402}{8\times63}$

= 1 - $\frac{402}{504}$

= 1 – 0.798

= 0.20

Example 2:

Consider the scores below generated from the scores of test administered to students taught in the rural and urban area.

Urban (X)	20	60	85	72	45	50	32
Rural (Y)	33	52	62	40	81	30	71

X	Y	R_X	R_y	d	d^2
20	33	7	7	0	0
60	52	3	4	-1	1
85	62	1	3	-2	4
72	40	2	5	-3	9
45	81	5	1	4	16
50	30	4	6	-2	4
32	71	6	2	4	16
					Total = 50

$rh_o = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

$rh_o = 1 - \frac{6 \times 50}{7(7^2-1)}$

$rh_o = 1 - \frac{300}{7(49-1)}$

$rh_o = 1 - \frac{300}{7(48)}$

$rh_o = 1 - \frac{300}{336}$

$rh_o = 1 - 0.8929$

$rh_o = 0.12$

The relationship is positive but very low

Example 3: Find the correlation coefficient of the following scores of students of a class

Individuals	**X**	**Y**
A	5	3
B	8	7
C	9	8
D	7	8
E	6	5
F	1	9

Table 28: Computation of Spearman Rho c.c

X	Score	Rank	Y	Score	Rank	d	d^2
A	5	5	A	3	6	-1	1
B	8	2	B	7	4	-2	4
C	9	1	C	8	2.5	-1.5	2.25
D	7	3	D	8	2.5	0.5	0.25
E	6	4	E	5	5	-1	1
F	1	6	F	9	1	5	25
						$\sum d^2$ =	**33.5**

Note that for two scores that are the same we may assign to them the average of the ranks e.g. C and D in Y variable has 8 marks each and should share the second and third ranks, average of 2 and 3 is 2. We may also assign the same ranks to the two scores with tie recognize that they have taken two position while assigning subsequent ranks. For e.g. we may assign ranks 2.5 each for subject C and D in y variable then the next subject will be assigned rank 4.

Applying the formula

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

Substitution:

$$1 - \frac{6 \text{ x } 33.5}{6(36 - 1)}$$

$$= 0.04$$

The correlation coefficient obtained indicates low positive correlation between the two sets of scores.

Example II: Given two sets of distribution of test scores of 10 students in English and Physics as in the table below, the Spearman rho c.c. becomes:

Table 29: Computation of Spearman Rho c.c.

Scores in English (x)	Scores in Physics (y)	Rank (x)	Rank (y)	d	d^2
80	56	2	7	-5	25
76	54	4	9	-5	25
70	60	5	6	-1	1

52	77	7	3	4	16
50	66	8	4	4	16
63	62	6	5	1	1
77	78	3	2	1	1
82	80	1	1	0	0
44	52	10	10	0	0
50	55	8	8	0	0
				$\sum d^2$ =	**85**

$$r = 1 - \frac{6 \times 85}{10(10^2 - 1)}$$

$$= 0.48$$

Like the product-moment coefficient of correlation, the Spearman's rank correlation coefficient takes values between -1 and +1. The Spearman rho rank correlation coefficient is part of the same statistical family as the median. It is an ordinal statistic designed for use with ordinal data. The product-moment method deals with the size of the measures (or scores) as well as with their positions in the series. Rank differences, on the other hand, take account only of the positions of the items in the series, and make no allowance for gaps between adjacent scores. Individuals, for example, who score 90, 89 and 70 on a test would be ranked 1, 2, 3 although the difference between 90 and 89 is much less than the difference between 89 and 70. Accuracy may be lost in translating scores over into ranks especially when there are a number of ties.

In spite of its mathematical disadvantages, r_s provides a quick and convenient way of estimating the correlation when N is small, or when we only have ranks.

Correlation Ratio

Correlation ratio is a measure of relationship which is used when there is a curved regression. This type of regression is not very common as linear relationships occur more frequently especially in psychological and educational measurements. However, a curved relationship may exist in non-psychological matters or when one correlates educational variable with one that is not educational (Guilford, 1965). A typical example of this is correlation of achievement scores with chronological age. To determine the correlation of the above variables in our example, two regressions may result. The regression of test score on age and the regression of age on test score, each of which is curved in nature due to two correlation ratios or what is known as eta coefficients for each of the regressions and according to Guilford, these will not necessarily be the same in value.

Illustration of Correlation Ratio

Supposing one is interested in knowing whether the test score of students are dependent on age, the correlation ratio becomes:

Table 30: Computation of Correlation Ratio

Students	**Test Score (x)**	**Age (Y)**	**X – X**	**Y – Y**	**xy**	**X/dx**	**Y/dx**	**Xdx Y/dy**
A	70	20	1.6	-1.6	-2.56	0.86	-0.66	-0.57
B	69	22	0.6	0.4	0.24	0.32	0.17	0.05
C	65	25	-3.4	3.4	-11.56	-1.84	1.40	-2.58
D	68	18	-0.4	-3.6	1.44	-0.22	-1.49	0.33
E	70	23	1.6	1.4	2.24	0.86	0.58	0.50
	∑ = 342 mx = 68.4 dx = 1.85	**108 my = 21.6 dy = 2.42**						**-2.27**

$$r = \frac{\sum(x/dx \,.\, y/dy)}{N}$$

$$r = \frac{-2.27}{5} = -0.45$$

The reason for computing correlation ratio is that we cannot take $^{y}/_{s}$ as r because this is not a stable measure. We need a ratio which is- a pure number as correlation coefficient, otherwise when the units of measurement change, correlation coefficient will change. Correlation ratio implies that we need a pure number independent of the unit of measurement. The product moment provides us with a procedure that converts the X and Y as standard scores and this throws away the effect of the units on our computation. Correlation ratio represents at best the average poduct of standard scores pertaining to variables X and Y (Guilford, 1965).

Biserial Correlation

In many problems, it is important to be able to compute the correlation between traits and other attributes when the members of the group can be measured in the one variable, but can be classified into two categories in the second or dichotomous variable. For example, we may wish to know the correlation between achievement scores and aptitude scores of students who are classified under higher achievers and lower achievers. Other examples are dichotomous variables like pass and fail, right and wrong, etc. The important thing about this dichotomous variable is that we have to assume that the trait in which we have made a two way split would be found to be continuous and normally distributed. In this case, a biserial r between the set of scores and the dichotomous variable may be computed (Garrett, 1966).

The formula of Biserial correlation coefficient =

$r_b \quad = \quad m_p$ - mt x $^p/_y$ ~S,

where;

m_p = mean of X values for the higher group in the dichotomized variable, the one having more of the ability on which the sample is divided into two sub-groups.

M_t = mean of the total sample.

Standard deviation of the total sample in the continuously measured variable

X. proportion of the cases in the higher group ordinate of the unit normal distribution curve with surface equal to 1.00 at the point of division between segments containing p and q proportions of the cases.

To illustrate the computation of biserial r, a distribution of grouped scores obtained by 88 individuals on a statistics test were sub divided into pass and fail respectively.

Table 31: Computation of Biserial Correlation

1	2	3	4	5	6	7	8
Classes	**F_1**	**f_q**	**f_r**	**x^1**	**fx^1**	**fx^2**	**f_px^1**
46 – 52	8	6	14	+2	28	56	16
39 – 45	14	6	20	+1	20	20	14
32 – 38	20	4	24	0	0	0	0
25 – 31	14	6	20	-1	-20	20	-14
18 – 24	10	0	10	-2	-20	40	-20
	66	**22**	**88**		**8**	**136**	**-4**

To compute the biserial r, we take the following steps:

1. Calculate the mean of those who passed using the formula:

X, +

$$+\left(\frac{\sum f_p x^1}{\sum f_p}\right) i$$

Where

X = class midpoint of the interval chosen as arbitrary origin.

i = interval size

$$m_p = 35 + \left(\frac{-40}{66}\right)7$$

$$= 34.56\ 2.$$

2. Calculate the mean of the total score using the formula:

$$m_t = X_i + \left(\frac{\sum ft\ X^1}{\sum f_t}\right) i$$

$$= 35 + \left(\frac{8}{88}\right) 7$$

$$= 35,64$$

3. Calculate the standard deviation of the total score:

$$S1 = i \quad \frac{\sum f_1 x^1\ 2}{\sum f_t} \quad \left(-\frac{\sum f_1 x^1}{\sum f_t}\right)^2$$

$$= 7 \quad \frac{136}{88} \quad \left(-\frac{8}{88}\right)\ 2$$

$$= \quad 8.68$$

Uses of Biserial r

The biserial r gives an estimate of the product moment r to be expected for the given data when certain assumptions have been met. These assumptions are:

1. Continuity in the dichotomized trait;
2. Normality of distribution, underlying the dichotomy;
3. a large sample;
4. a split that is not extreme but near to the median (Garrett, 1966).

Limitations of Biserial r

1. Biserial r cannot be used in a regression equation. This is because it has no standard error of estimate, and the score predicted for all members of the group is simply the mean of that category (Garrett, 1966).
2. Biserial r is less reliable than the Pearson r. It is favoured against the latter only in cases where the sample is very large and computation time an important consideration.

The Point Biserial r

When one of the two variables in a correlation problem is a genuine dichotomy, the appropriate type of coefficient to use is point biserial r (Guilford, 1965). This means that point biserial r lends itself to situations where items are scored simply as I if correct and 0 if incorrect, that is, as right or wrong. Although this case is not a case of fundamentally dichotomous variables but they are treated in practice as if they are genuine dichotomies. Example of true or genuine dichotomous variables where biserial r is applicable are male - female, living - dead, loyal - disloyal, delinquent - non delinquent, psychotic - normal etc. The formula for the point biserial r is

$$r_{pbi} = \frac{m_p - m_q}{S_t} \quad pq$$

Where m_p and m_q are the means of the two categories;

p is the proportion of the sample in the first group;

q is the proportion of the sample in the second group;

s is the standard deviation of the sample.

Illustration 1: Calculation of point biserial r in the analysis of items of a test where 1 = item passed; 0 = item failed.

Table 32: Computation of Point Biserial c.c. for Ungrouped Data

1	**2**	**3**	**4**	**5**	**6**
Students	**Test Criterion X**	***Item No. 15 Y***	**x^2**	**y^2**	***XY***
1	20	1	400	1	20
2	16	0	256	0	0
3	14	1	196	1	14
4	15	0	225	0	0
5	15	1	225	1	15
6	18	1	324	1	18
7	19	0	361	0	0
8	20	1	400	1	20
	137	**5**	**2387**	**5**	**87**

N_p (number who passed) = 9

N_q (number who failed) = 3

$$M_p = \frac{87}{5} = 22.33 \qquad p = .63$$

$$M_q = \frac{50}{3} = 16.67 \qquad q = .38$$

$$M_t = \frac{137}{8} = 17.13$$

S_t = 2.42

r_{pbi} = $\frac{17.4 \; - 16.67}{2.42}$ .63 x 38

The low positive correlation obtained indicates that its No. 6 is a bad item and should be discarded.

Example 11: Calculation of point biserial r for grouped obtained by 50 students in a test.

Tale 33: Computation of Point Biserial c.c. for Grouped Data

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Classes	f_p	f_q	f_r	x^1	fx^1	fx^2	f_px^1
40 – 44	5	4	9	+3	27	81	15
35 – 39	6	4	10	+2	20	40	12
30 – 34	10	2	12	+1	12	12	10
25 – 29	4	3	7	0	0	0	0
20 – 24	3	1	4	-1	-4	4	-3
15 – 19	3	2	5	-2	-10	20	-6
10 – 14	2	1	3	-3	-9	27	-6
$\sum$ =	**33**	**17**	**50**		**26**	**184**	**22**

Significance of Point Biserial r

The point biserial r may be tested against a null hypothesis. Since r_{pbi} depends directly upon the difference between the means m_p and m_q, a significant departure from a mean difference of zero also indicates a significant correlation. A direct t test of the correlation coefficient can be made using r_{pbi} but only for the hypothesis of a correlation of zero. The t- ratio can be computed for r_{pbi} in the same manner as for a Pearson Product - moment r (Guilford, 1965).

Partial and Multiple Correlation

Portial and multiple correlation represent an important extension of the theory and techniques of simple or 2 - variable linear correlation to problems which involve three or more variables (Garrett, 1966). Partial correlation is often useful in analyses in which the effects of some variable or variables are to be eliminated. Its chief value, however, lies in the fact that it enables us to set up a multiple regression equation of two or more variables by means of which we can predict another variable or criterion. A partial correlation between two variables, therefore, is one that nullifies the effects of a third variable (or a number of other variables) upon both the variables being correlated (Guilford, 1965). For example, if we want to determine the correlation between vocabulary score and intelligence level which may be affected by age, and we want to control the age factor, this may be done:

i. experimentally by selecting children all of whom are of the same age or
ii. statistically, by holding age variability constant through partial correlation. The first method will drastically reduce the size of the sample while partial correlation will utilize all of the data which makes it preferable.

If we let 1 = vocabulary score, 2 = intelligence level, and 3 = age, Then $r_{12.3}$ represents the partial correlation between 1 and 2 (vocabulary score and intelligence level) when 3 (age) has been held constant or "partialed out." The subscripts 12.3 mean that variable 3 is rendered constant, leaving the net correlation between 1 and 2. The subscripts in the partial correlation coefficient, (112.345) mean that 3 variables, namely, 3,4,5, are partialed out from the correlation between 1 and 2. The numbers to the right of the decimal point represent variables whose influence is ruled out; those to the left represent the two correlated variables (Garrett, 1966).

The correlation between a set of obtained scores and the same scores predicted from the multiple regression equation is called a coefficient of multiple correlations. It is designated by the letter R (called multiple R). If $R_{1\ (234)}$ = .72, this means that scores in variable (1) -predicted from a multiple regression -equation containing variables' (2), (3), and (4) correlate .72 with' scores obtained in variable (1), Expressed in another way, RI (234) gives the con-elation between a criterion (1) and a team of tests (2,3, 4). The variables in parentheses ( ) are the independent variables in the regression equation, whereas the variable outside of the parentheses, namely, (1), is the criterion to be predicted or estimated.

The multiple regression equation is set up by way of partial correlation and its accuracy as a predicting instrument is given by the coefficient of multiple correlation, R. The meaning of partial and multiple correlation will be better understood when computed (Garrett, 1966).

Illustration:

Calculation of partial and multiple correlation of vocabulary scores, intelligence level and age of 450 students adapted from Garrett (1966, pp404 – 406).

1. Vocabulary Score	2. General Intelligence	3 Ageof Student
M_1 = 15.6	M_2 = 99.8	M_3 = 22
S_1 = 9.5	S_2 = 10.5	S_3 = 4
r_{12} = .65	r_{13} = .48	r_{23} = .38

Step 1: Equations for multiple regression are:

$$X_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{(deviation form)}$$

Note: The equation above is in deviation form.

Where:

Xi = Vocabulary score

X_2 = General intelligence

Xs = Age of students

We can rewrite the equation is score form which becomes

$(X1 - m1) - b_{12.3} (X_2 - m_2) + b_{13.2} (X_3 - m_3)$.

Step 2: Computation of partial r'

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r^2_{13} \quad 1 - r^2_{23}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r^2_{12} \quad 1 - r^2_{23}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{1 - r^2_{12} \quad 1 - r^2_{13}}$$

Substituting the formula with our data:

$$r_{12.3} = \frac{.65 - .48\ (.38)}{.877 \times .925} = .58$$

$$r_{13.2} = \frac{.48 - .65\ (.38)}{.760 \times .925} = .33$$

$$r_{23.1} = \frac{.38 - .65\ (.48)}{.760 \times .877} = .10$$

Step 3: Computation of Partial S's.

$S_{1.23} = S_1 \quad 1 - r^2_{12} \qquad 1 - r^2_{13.2}$

$S_{2.13} = S_{2.31} = S_2 \quad 1 - r^2_{23} \qquad 1 - r^2_{12.3}$

$S_{3.12} = S_{3.21} = S_3 \quad 1 - r^2_{23} \qquad 1 - r^2_{13.2}$

$S_{1.23} = 9.5 \times .760 \times .944 = 6.8$

$S_{2.13} = 10.5 \times .925 \times .815 = 7.9$

$S_{3.12} = 4 \times .925 \times .944 = 3.5$

Step 4: Computation of partial regression coefficients and regression equations

$$b_{12.3} = r_{12.3} \frac{S_{1.23}}{S_{2.13}}$$

$b_{13.2} \quad = \quad .58 \times \frac{6.8}{7.9} \quad = \quad .50$

$b_{13.2} \quad = \quad .33 \times \frac{6.8}{3.5} \quad = \quad .64$

The regression equation is:

$X_1 \quad = \quad .50\,x2 + .63X_3$ in deviation form.

$(X_1 \_M_1) \quad = \quad .50(x_2 \_99.8) + .64(x_3 - 22)$ in score form

Given a student's intelligence score (x_2) and his age (x_3), we can estimate from this equation the most probable vocabulary score (x1) he will receive at the end of the term. Suppose the student has an intelligence score of 110 and has the age of 25 years, how many vocabulary scores should he have at the end of the term? Substituting $X_2 = 110$ and $X_3 = 25$ in the equation, we find that:

$(x_1 - M_1) \quad = \quad .50(110 - 99.8) + .64(25 - 22)$

$= \quad 7.02$

And the most likely vocabulary score he will receive as predicted from his intelligence score and age is 7.02

Step 5:

The final step in the solution of our 3 - variable problems is the computation of the coefficient of multiple correlations. Multiple R is defined as the correlation between score actually earned on the criterion (X_1) and scores predicted in the criterion from the multiple regression equation.

The formula for R when we have 3 variables is

$Rl_{(23)} = \quad 1 - \frac{S^2_{1.23}}{S_1^2}$

$R1_{(23)} = \quad 1 - \frac{(6.8)^2}{(9.5)^2} \quad = \quad .70$

Exercises

1. Given the distribution of scores below, calculate the Pearson product moment correlation with Deviation method and the Raw score method.

Eng (X)	15	10	12	16	20	14	14
Maths (Y)	20	18	18	19	18	15	12

2. Interpret the results of the computation done in No 1 above.

3a. Determine the correlation of the data below using Spearman rank ordered method

History	**Geography**
99	72
80	82
76	76
57	52
76	48
58	32
82	48
45	48
30	60

b. What type of relationship exists between the two subjects?

4. Plot scatter diagram with the data in No. 1 and No.3

5. The following data was collected by a researcher:

X	**Social Adjusted Students**	**Socially Maladjusted Students**
85 – 94	6	4
75 – 84	8	2
65 – 74	22	3
55 – 64	28	15
45 – 54	32	20
35 – 44	10	8
25 – 34	8	6
15 – 24	8	4

Complete biserial r for the data.

CHAPTER FIVE

REGRESSION ANALYSIS

Francis Galton first used the term “regression” with reference to children's inheritance of parental stature. Galton discovered that children of tall parents tend to be less tall than their parents while children of short parents tend to be less short. This means that the children's height tend to be "move back" towards the height of the general population which is the "mean height". Galton called this tendency of children's height moving back to the mean height, the principle of regression and called the line which describes the relationship of height in parent and children "regression line" (Garrett, 1966). In the course of time, the word regression became synonymous with the statistical study of relation among variables. This is called regression analysis.

The term regression analysis, according to Hamburg (1983) refers to "the methods by which estimates are made of the values of a variable from knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimate process". "Regression analysis concerns the study of relationships between variables with the object of identifying, estimating and validating the relationship. The estimated relationship can then be used to predict one variable from the value of another" (Johnson & Bhattacharya, 1996, P.461).

Regression analysis is a statistical procedure which provides the basis for predicting values of a variable from the values of one or more other variables (Hamburg 1983). Correlation analysis, on the' other hand, enables us to assess the strength of the relationships among the variables under study. In educational matters and social sciences, exact relationships are not generally observed among variables, but rather statistical relationships prevail. That is, certain average relationships may be observed among variables but these average relationships do not provide a basis for perfect predictions. The techniques of regression and correlation analysis are important statistical tools in this measurement and estimation process.

Regression analysis can be linear or multiple. Linear regression means an equation of a straight line of the form Y = A + BX where A and B are fixed numbers. This is used to describe the average relationship between the two variables and to carry the estimation process. The factor whose values we wish to estimate is referred to as the dependent variable and is denoted by the symbol Y. The factor from which these estimates are made is called the independent variable and is denoted by X.

Multiple regression analysis, on the other hand, refers to a situation where one wants to estimate a dependent variable from two independent variables. For example, when a teacher wants to predict students' achievements in a school subject from their family income and family size.

Purpose of Regression Analysis

a. The first purpose of regression analysis is to provide estimates of values of the dependent variable from values of the Independent variable. The device used to accomplish this estimation procedure is the sample regression line which provides estimates of mean values of Y for each value of X. That is, it describes the average relationship between the X and Y variables in the sample data.
b. Another purpose of regression analysis is to obtain measures of the error involved in using the regression line as a basis of estimation.

Conditions for using Regression Analysis

A study of the relation between two variables is often motivated by a need to predict one from the other. Guilford (1965) classified different types of prediction one may wish to make into four groups, namely:

i. Attributes from other attributes - as when we predict incidence of low male enrolment from social status or societal values.

ii. Attributes from measurements - as when we predict delinquency from achievement test scores or intelligence tests.
iii. Measurements from attributes - as when we predict problem test scores from sex, socioeconomic status, or marital stability.
iv. Measurements from other measurements - as when we predict academic achievement from intelligence test scores.

The administrator of a school may wish to study the relation between the duration of workshop training of his staff and the score of the staff on a subsequent skill test. A teacher may also wish to study the relation between the intelligence test score of his students and their achievement level in science subjects. In such context as these, the predictor or input variable is denoted by X, and the response or output variable is labeled Y. The object is to find the nature of relation between X and Y from collected data and use the relation to predict the response variable from the input variable X.

Formulae for Calculation of Regression Analysis

The first step in regression analysis is to plot and examine the scatter diagram. If the linear relation emerges, the calculation of the numerical value r will confirm the strength of the linear relation. Its value indicates how effectively Y can be predicted from X by fitting a straight line to the data.

A line is determined by two constants: its height above the origin (intercept) and the amount that Y increases whenever X is increased by one unit (slope). This leads to what is called regression equation and regression coefficient by Guilford (1965). In order to understand these regression analyses, we are going to use a combination of methods described by Guilford (1965) and Garrett (1966) to represent the old form and Johnson and Bhattacharya (1996) to represent a modification of old formula.

The equation for a straight line, in general form, is $Y = a + bX$. Such an equation completely describes a line when a and b are known, they are the data collected. Regression equation may be in deviation form or score form. The equations of the two regression lines in a correlation table represent the straight lines which "best fit" the means of the successive columns and rows in the table.

Deviation Form

1. $Y = r \frac{SY}{SX} \times X$

2. $X = r \frac{SX}{SY} \times Y$

Where;

r	=	Pearson product moment correlation coefficient
Y	=	mean of the variable Y
X	=	mean of the variable X
Sy	=	standard deviation of variable Y
Sx	=	standard deviation of variable X

Score Form:

1. $Y = r \frac{SY}{SX} (X - Mx) + My$

2. $X = r \frac{SX}{SY} (X – MY) + MX$

Suppose we have data like the following:

R	=	.60; SY = 2.62; SX = 15.55
MX	=	136.3 pounds representing a man's weight
MY	=	66.5 inches representing a man's height

If we want to predict X from Y and Y from X using the regression equation.

Deviation Form:

1. $Y = .60 \times \frac{2.62}{15.55} \quad X = .10X$

2. $X = .60 \times \frac{15.55}{2.62} \quad Y = 3.56Y$

Score Form:

1. $Y = .10X + 52.9$

2. $X = 3.56Y + 100.4$

The regression equation can be calculated as described by Johnson and Bhattacharya (1996) thus;

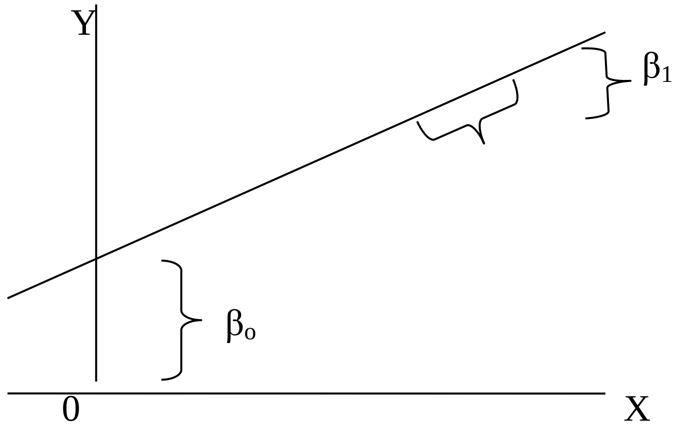
Method of Least Squares

$y = \beta o + \beta_1 x$

where slope $\beta_1 = \frac{Sxy}{Sxx} = \frac{\sum(X - X)\ (Y - Y)}{\sum(X - X)^2}$

Intercept $\beta_o = Y - \beta_1 X$

Line $y = \beta_o + \beta_1 x$



In order to illustrate the above calculation, a teacher wishes to study the relation between the performance of students in physics subjects and their aptitude test results in the same subject. In other words he Wishes to predict the performance of students (Y) given their aptitude in physics (X). To do so, we plot the data, calculate r, and obtain the fitted line. We will use the method of least squares.

$Y = \beta_o + \beta_{1x}$

Table 35: Calculation of Physics Aptitude Scores (x) and Physics Test Score (y).

Test Scores (y)

X	Y	X^2	Y^2	XY
0	1	0	1	0
1	5	1	25	5
2	3	4	9	6
3	9	9	81	27
4	7	16	49	28

Total	**10**	**25**	**30**	**165**	**66**

$\bar{X} = \frac{10}{5} = 2$ $\qquad \bar{Y} = \frac{25}{5} = 5$

$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 30 - \frac{(10)^2}{5} = 10$

$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 165 - \frac{(25)^2}{5} = 40$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 66 - \frac{(10 \times 25)}{5} = 16$

$r = \frac{S_{xy}}{(S_{yy})(S_{xx})} = \frac{16}{40 \times 10} = \frac{16}{20} = .8$

$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{16}{10} = 1.6$

$\beta_o = Y - (\beta 1) X \quad . \quad 5 - (1.6)2 = 1.8$

The equation of the fitted line is Y = 1.8 + 1.6x

If we are to predict the performance in physics y corresponding to the aptitude score 2.5, we substitute x = 2.5 in our regression equation and get the result.

At X 2.5, the predicted performance = 1.8 + 1.6 (2.5)

Y = 5.8

Exercises:

1a. What is Regression Analysis?

b. Write out the regression equations in score form for the data in Table 28.

c. What is the most probable grade in Y of a student whose X mark is 7.

2. Given the following data for two tests:

Mathematics (s) *Chemistry (y)*

MX = 50.00 My = 65.00

Sx = 7.50- Sy = 5.00

Rxy = .78.

a. Work out the regression equation in score form.

b. Predict the probable grade in chemistry of a student whose mathematics mark is 45.

c. Using the method of least squares, predict the performance of a student in mathematics (y) given his 1Q level (x) as 122 with the data presented below.

Students	**IQ Level**	**Mathematics Test Scores**
A	112	58
B	115	66
C	114	52
D	120	70
E	102	45
F	125	75
G	113	68
H	114	60
I	118	60
J	130	80

CHAPTER SIX
HYPOTHESIS TESTING

Hypotheses refer to statements of solutions to a problem which are tentative in nature. A hypothesis is therefore a tentative solution to a problem or what Nworgu (1992) termed an intelligent guess concerning solution to a research problem. This solution to the problem is relative to testing which makes it tentative in nature. In other words, when the proposal is made concerning the relationship between two or more variables, the researcher collects data and analyses the collected data in order to determine the truthfulness or otherwise of proposal made. The process of data analysis may necessitate rejection or acceptance of the hypothesis. This is what we call testing of a hypothesis or hypothesis testing.

A hypothesis may be inductive or deductive (statistical). For this chapter, our interest is on deductive or statistical hypothesis which is a hypothesis stated in statistical measurable terms. Statistical hypothesis is of two types viz: null hypothesis and alternative hypothesis.

Null Hypothesis (H_0)

Null hypothesis (H_0) is a statistical hypothesis which is chance explanation of relationship between two or more variables. In other words, it is a statement made on research problem which states that there is no significant relationship between two or more variables. Null means No or nothingness. H_0 will always affirm that the relationship or difference between two or more variables is not significant. Any relationship or difference noticed according to Ho, is a matter of chance.

Examples of H_0 are:

i. There is no significant difference between male and female students' performance in secondary school mathematics.
ii. The relationship between students' attitudes and achievements in Physics subject is not significant.

Note that the hypothesis that is tested is the null hypothesis. When HQ is rejected, the alternative hypothesis stands or is retained.

Alternative Hypothesis (H_a)

Alternative hypotheses are statements made on significance of relationships or differences between two or more variables. Where H_o states that there is no significant difference existing between variables, alternative hypothesis states that there is a significant difference between them. When the H_0 is rejected, H_a is accepted.

Examples of H_a

There is a significant difference between the performance of male and female students in secondary school mathematics.

There is a significant relationship between students' attitude and performances in secondary school Physics.

Directional and Non Directional Hypothesis (One Tailed and Two Tailed Tests)

An alternative hypothesis can be either directional or non-directional. Directional or non-directional hypotheses are determined from the way the alternative hypotheses are stated. Directional hypothesis refer to the hypothesis that states the direction of difference or relationship between variables of interest. In such hypothesis, the rejection or acceptance region is at one tail of the normal curve. This is why directional hypothesis leads to test called one tailed test. Example of directional hypothesis and one tailed test:

H_a: The performance of male students is significantly higher than that of female students in secondary school mathematics. This hypothesis indicates the direction of its acceptance or rejection which is on the high side or right tail of the normal curve.

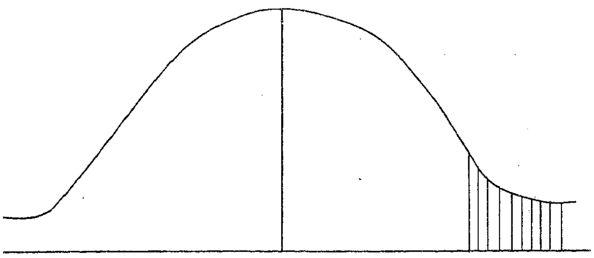


Fig 11: The One Tailed Test

Non-directional hypothesis, on the other hand, is a statement of relationship between two or more variables which does not indicate the direction of the relationship. In such situation, the hypothesis is rejected or accepted at both tails of a normal curve. This leads to what is called two tailed test. Example of non directional hypothesis and two tailed test:

H_a: The performances of male and female students in mathematics differ significantly.

This hypothesis does not indicate direction of rejection or acceptance of the hypothesis. Therefore the H_Q will either be accepted or rejected at both tails of a normal curve.

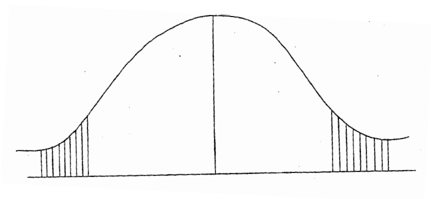


Fig 12: A Two Tailed Test

Decision Rule

In hypothesis testing, the calculated value of collected data with inferential statistical procedures is compared with the critical value or table value. The critical value becomes the criterion with which the calculated value is weighed in order to conclude whether the hypothesis will be accepted or rejected Each inferential statistics has a table of values which is copied by the tester. We will later see an example of how to go to the table for critical value of any test of interest.

The decision rule in hypothesis testing is:

1. Reject Ho when calculated value is greater than or equal to the critical value.
2. Retain Ho when calculated value is less than the critical value

While rejecting or retaining Ho, there is probability of committing some errors. Actually, four decisions may be taken namely:

1.	Rejection of False	Ho	
2.	Acceptance of a true	Ho	
3.	Rejection of a true	Ho	x
4.	Acceptance of a false	Ho	x

The first two decisions are correct while the last two are wrong and amount to errors known as type 1 and type II errors.

Type I and Type II Errors

Type I error refers to rejection of a true null hypotheses. This is very grievous error because it leads to a lot of expenditure and adjustments which are not necessary. Type II errors refer to acceptance of a false Ho. This is not grievous as Type I error because although it is incorrect decision, it does not lead to unnecessary adjustments and expenses.

In testing hypothesis, care is taken to avoid committing either of the two types of errors. This can be done through good sampling and use of appropriate statistical procedure. However, there is always the possibility of committing one of the error types no matter how careful a researcher is.

Level of Significance

Every statistical hypothesis stated indicates non-significance or significance of difference/relationship existing between variables. This significance indicates a level at which the hypothesis is measured. This is measured in relation to the risk the tester is running in committing Type I error. The amount of risk a researcher accords himself of committing Type 1 error is called level of significance or alpha level (α). This level differs depending on the level of risk every researcher gives himself of probably committing Type 1 error. Some researchers may give themselves a probability of committing error in 5 out of every 100 cases. This leads to α level of .05. Others may give themselves 1 error out of every 100 cases which leads to a α level of .01 etc.

The probability of committing Type II error is called Beta (β) level. As we mentioned earlier, Type II error is less grievous than Type 1. This is why probability of committing Type I error is used to test hypothesis. In doing so, however, caution will be taken not to limit the risk so much because as a researcher reduces the risk of committing Type I error, he runs a greater risk of committing Type II error.

Degree of Freedom *(df)*

Scores in a distribution may vary or change in line with the degrees of freedom they have to vary in that distribution. Degrees of freedom refer to the amount of freedom scores have to vary in a distribution. This normally amounts to N – 1 although the formula depends on the statistical method being applied in the analysis of data.

In hypothesis testing, one needs the calculated value, the level of significance (a level), the degrees of freedom of that particular statistics and the critical value. Actually the level of significance and the degrees of freedom are used to get the critical value of any test statistics. In any table of values, the degrees of freedom are written at the vertical axis while the α levels are written at the horizontal axis. With the particular degrees of freedom and a level, the researcher copies the value at the point where they intersect. This is the critical value of that particular test statistics.

Exercises

1. What is hypothesis testing?
2. Explain the meaning of these terms:
   - Levels of significance
   - Type I and Type II errors and
   - degrees of freedom.

3. How does a researcher reach a conclusion on whether to retain or not retain a stated null hypothesis?

CHAPTER SEVEN
INFERENTIAL STATISTICS

As has been mentioned earlier in chapter one, statistics may be either descriptive or inferential. Statistics procedures that are descriptive in nature have been discussed in the preceding chapters. There are occasions which necessitate other statistical procedures which are not descriptive. In such cases, inferential statistics are applied in data analysis. The situations that necessitate inferential statistics applications are:

i. When sample is selected from population and inferences of population characteristics are made from the sample.
ii. When we are testing hypotheses on significances of differences or relationships existing between two or more variables.
iii. When interest in research study is more than just describing the status of any phenomenon.

Inferential statistics therefore is statistics applied in analysis of data collected from a sample which is generalized to the population. Whenever we do such generalizations or inferences, some measure of error is involved which inferential statistics recognizes and makes provisions for. However, inferential statistics is mainly used for hypothesis testing.

Parametric and Non Parametric Statistics

Inferential statistics may be parametric or non parametric. Parametric statistics are the statistical methods used to test hypothesis when assumptions can be made that the data was collected from a normally distributed population. These statistical methods are used for data on the interval or ratio scales.

Non-parametric statistics are statistical procedures used to test hypotheses when one cannot assume that the data is collected from a normally distributed population. They are used for data on the nominal or ordinal scales. Parametric statistics are discussed in this chapter. Non-parametric statistics are discussed in the next chapter.

There are many statistical procedures that fall under parametric statistical category. These are:

t - test for correlated or non independent sample data;

t - test for non correlated or independent sample data;

t - test for Pearson correlation analysis.

Analysis of Variance (ANOVA);

Analysis of Covariance (ANCOVA).

t - Test Inferential Statistics

t- test is an inferential statistics which is used to test hypothesis concerning significance of difference between two means i.e. X_1, X_2.

This means that if interest of a research work does not involve testing hypothesis on the difference between two means only, t-test becomes inappropriate or else the researcher will be involved in the laborious task of hypothesis testing if the means are more than two.

This will imply testing two group combinations at a time till all the possible combinations are finished.

a. t-test for Non-Correlated or Independent Data

Whenever the researcher wants to make inferences concerning two population means that are different or independent of each other, t-test for independent data is used. Data is said to be independent or non-correlated when they are collected from two different samples. For example:

i. A researcher may collect data concerning students' test scores in physics and chemistry and wants to test significance of difference of students' performance in the two subjects. The sample is independent and non-correlated as they are different from each other.

ii. A researcher may also collect data from male and female students concerning their interest level in physics. The data is non-con-elated because they are collected from two different samples.

Procedure for Calculation of t-test for Independent Data

A researcher interested in finding out the difference in performance of students from two different socio-economic backgrounds in mathematics subject, takes the following steps:

Step 1: State the null and alternative hypotheses.

H_0 = There is no significant difference in performance of students from different socio-economic backgrounds in mathematics subject.

Ha = There is a significant difference in performance of students from different socio-economic backgrounds in mathematics subject.

Step 2:

Decide the a or significance level at which we are testing the H_0,

Given the following data

Group 1 (High Socio-Economic Status)	**Group 2 (Low Socio-Economic Status)**
X1 = *62*	X2 = 48
S^2_1 = 12	S^2_2 = 14
N_1 = 10	N_2 = 18

Step 3: State the formula

$$t = \frac{X_1 - X_2}{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$$

Step 4: Put the data into the formula and compute t.

$$t = \frac{62 - 84}{\frac{12^2}{10} + \frac{14^2}{18}} = 2.784$$

The result of the t calculated (t cal.) above is 2.784. In order to test the stated, null hypothesis, we need to consult the t table of values to copy the critical or table value (t crit.) which is used as a criterion to either reject or accept the null hypothesis. To do so we will first of all determine the degrees of freedom (df) of our data which in combination with the a level, will enable us to copy the table value.

df = N, + N_2 - 2 = 10 + 18 - 2 = 26

With 26 df and .05 α level, t critical value = 2.056.

Decision Rule

If t cal is < t crit, retain H_0

If t cal is > t crit, reject H_0

Decision: Because t cal of 2.784 is greater than t crit. of 2.056, the Ho is rejected at .05 *a* level.

Step 5: Build a summary, table of t. Table 36: Summary Table oft.

Subjects	N	SD	X	t-cal	df	a	t-crit
Students of High Socio-Economic Status	10	12	62	2.784	26	.05	2.056
Students of Low Socio-Economic Status	18	14	48				

Xi, Xa, S_1 & S_2 are known, There may be situations where only the raw scores are given. In order to determine that, we use the following formula and procedure.

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sum(X_1 - \overline{X_1})^2 + \sum(X_2 - \overline{X_2})^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where **X_i.** X_i = the means of the two sets of data

X_1X_2 = deviations of the scores from the means of the two sets of data

n_1,n_2 = the number of scores in groups 1 and 2 respectively

$\sum x_1^2$ $\sum x_2^2$ = sums of squared deviation scores for the two groups.

Supposing we have the data below, t-test becomes:

Table 37: Computation of t-test for Independent Data

Group 1 (x_1)	$(X_1 - \overline{\overline{X}})$	$(X - \bar{X})^2$	Group 2 (x_2)	$(X_2 - \bar{\bar{X}}_2)$	$(X_2 - \overline{X_2})^2$
19	4.64	21.5296	17	3.5500	12.6025
14	-0.36	0.1296	16	2.5500	6.5025
12	-2.36	5.5696	15	1.5500	2.4025
16	-1.64	2.6896	12	-1.4500	2.1025
17	-2.64	6.9696	10	-3.4500	11.9025
10	-4.36	19.0096	11	-2.4500	6.0025
13	-1.36	1.8496	12	-1.4500	2.1025
11	-3.36	11.2896	14	0.550	0.3025
14	-0.36	0.1296	11	-2.4500	6.0025
12	-2.36	5.5696	17	3.5500	12.6025
20	-5.64	31.8096	13	-0.4500	0.2025
158		106.5456	148		

$\bar{X}_1 = \frac{158}{11} = 14.36$

$\bar{X}_2 = \frac{148}{11} = 13.45$

Putting the data into formula

$$\frac{14.36 - 13.45}{\sqrt{\frac{106.545 + 62.7275}{11 + -2}\left(\frac{1}{11} + \frac{1}{11}\right)}}$$

$$= \frac{0.91}{\sqrt{\frac{169.2731}{20}(0.1818)}}$$

$$= \frac{0.91}{\sqrt{8.463655(0.1818)}}$$

$$t = \frac{0.91}{\sqrt{1.5387}}$$

$t = \frac{0.91}{1.2404}$
t = 0.7336
$t \cong 0.734$
Degree of freedom

$n_1 + n_2 - 2$
= 11 + 11 – 2
= 20

(a) α = 0.05, df = 20
ttab = 2.086

Recall: the decision rule says reject the H_o when the Tcal > Ttab
The Tcal = 0.734
Ttab = 2.086
On this note, the tab is greater than the tabulated; we therefore accept the null hypothesis that there is no significant difference in the twp method used for teaching the students.

When testing r, we use the small formula t which is $t = r\sqrt{\frac{n-2}{1-r^2}}$
r = 0.68
n = 2
Test the hypothesis that
Example 2
t- test for linear relationship
$t = \sqrt{\frac{n-2}{1-r^2}}$
r = 0.68
n = 2
$0.68\sqrt{21-2}$

Example 2
The scores below were generated from a test administered in Degree year 2 Biology department, Peaceland College of Education, Enugu.
Test the hypothesis at 0.05 alpha that the two methods are effective.

Group 1	60	38	45	70	25
Group 2	66	41	22	36	75

H_O: the two methods used are effective

Group x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$	Group x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
60	12.4	153.76	66	18	324
38	-9.6	92.16	41	-7	49
45	-2.6	6.76	22	-26	676

70	22.4	501.76	36	-12	144
25	-22.6	510.76	75	27	729
$\Sigma x_1 = 238$ $\bar{x}_1 = 47.6$		1265.2	$\Sigma x_2 = 240$ $\bar{x}_2 = 48$		1922

Mean $(\bar{x}_1)$

$\bar{x}_1 = \frac{\Sigma fx}{N}$

$\bar{x}_1 = \frac{238}{5}$

$\bar{x}_1 = 47.6$

Mean $(\bar{x}_2)$

$\bar{x}_2 = \frac{\Sigma fx}{N}$

$\bar{x}_2 = \frac{240}{5}$

$\bar{x}_2 = 48$

$S.D_1 = \sqrt{\frac{1265.2}{5}}$

$S.D_1 = \sqrt{253.04}$

$S.D_1 = 15.90$

$S.D_2 = \sqrt{\frac{1922}{5}}$

$S.D_2 = \sqrt{384.1}$

$S.D_2 = 19.59$

$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$T = \frac{47.6 - 48.0}{\sqrt{\frac{15.9}{5} + \frac{19.59}{5}}}$

$T = \frac{-0.4}{\sqrt{3.18+3.918}}$

$T = \frac{-0.4}{\sqrt{7.098}}$

$T = \frac{-0.4}{2.66}$

T = -0.15

$df = n_1 + n_2 - 2$

df = 5 + 5 – 2

df = 8

$\alpha = 0.05$

Tcrit = 1.960

Tcal = -0.15

In this situation, we do not reject the null hypothesis since the Tcrit is greater than Tcal. It is concluded therefore that the two methods used are effective

b. t test for Correlated or Non-independent Data

There are times or situations when two means to be compared are obtained from the same sample on different occasions or from matched pairs of subjects. When such situations arise, we say that the data is collected from correlated or non independent sample. Example of such data is when a research problem is on ascertaining whether a new method will enhance the performance of students in mathematics subject. The researcher gives the students a pretest and records the scores. He then administers the treatment which is teaching with the new method, after which he gives a post test and records .the data. The researcher now has pretest and posttest data collected from the same sample. In order to determine whether the difference in the two sets of data collected from the same sample is significant, he uses t test for correlated data.

Procedure for Calculation

Step 1: - State the H_0 and H_a.

H_0: The pretest mean scores of the students in mathematics do not differ significantly from the posttest scores after application of a new teaching method.

H_a: The pretest mean scores of the students in mathematics differ significantly from the posttest scores after application of a new teaching method.

Given the following data:

Pretest (X_1)	**Post test (x_2)**
56	66
50	48
50	60
72	62
74	76
40	46
61	82
50	42

Step 2: State the formula

$$t = \frac{D/N}{\frac{\sum D^2 \; - \; (\sum D)^2}{N(N \; - 1)}}$$

Where;

t = the t-value for correlated (non-independent) means
D = the difference between the paired scores
D = the mean of the differences
N = the number of the pairs

Step 3: Determine 'D', 'D2' and D as in the table below and apply the formula

Table 38: Computation of t-test for Correlated Sample

Number	**Pretest (X1)**	**Post test (X2)**	**D(X2 – X1)**	**D2**
1	**56**	**66**	**10**	**100**
2	**50**	**48**	**-2**	**4**
3	**50**	**60**		

the t-value for correlated (non-independent) means, the difference between the paired scores, the mean of the differences the number of the pairs.

Table 38: Computation of t-test for Correlated Sample

Number	**Pretest (X_2)**	**Posttest (X_2)**	**D(X_2- X_1).**	**D^2**
1	56	66	10	100
2	50	48	-2	4
3	50	60	10	100
4	72	62	-10	100
5	74	76	2	4
6	40	46	6	36
7	61	82	21	441
8	50	42	-8	64
			∑ = 29	**849**

$$= \frac{\frac{29}{8}}{\sqrt{\frac{849 - \frac{(29)^2}{8}}{8(8-1)}}}$$

$= \frac{3.625}{3.645} = 0.995$

df = N-l=7. Testing at .05α level

tcrit = 2.365, therefore Ho is retained

Note: To get the difference between the two sets of data: Pretest/Post test, the pretest is subtracted from posttest (X2 – X1)

Note: t-test is an inferential statistics used to test significance of difference between two population means when the sample size is small. However, when the sample size is large i.e. from 30 upwards, z test becomes more appropriate. The procedure for calculation of both the t test and z test are the same but they differ in the critical values, z test has constant critical values given below as:

CRITICAL VALUES OF Z

	.01	.05		.01	.05
One tailed test	2.326	1.645	Two tailed test	2.576	1.960

c. t-test for Correlation Analysis

If the interest of any research work is to ascertain the significance or relationship existing between two or more variables, correlation analysis becomes the appropriate statistical procedure which is used to determine the degree of the relationship between the variables. In testing for significance, t test for correlation analysis is applied. We will have to note here that before the application of t here, appropriate correlation coefficient of the variables will first be determined. If the data is on either the interval or ratio scale, Pearson r will be used initially and then t test for Pearson r will be .applied in testing for significance of relationship.

Given the Pearson r c.c as .76 and N = 25, in order to test the significance of this result concerning relationship between two variables, we apply the formula:

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

$$= .76\sqrt{\frac{25-2}{1-.76^2}} \qquad t = 5.608$$

Decision

Determine the degree of freedom using the formula,

df = N - ,2 = 23 and testing at .05 level,

If t cal. < t crit, do not reject H_0

If t cal. $\geq$ t crit, reject H_0

Decision: Because the t cal. 5.608.4s greater than t crit. of 2.069 at .05 a level, theH_0 is rejected at that level.

Analysis of Variance (ANOVA)

The t test discussed above is the statistical procedure used to test the significance of difference between two population means only. There are times however, when we need to test a difference between more than two means. In such situations, ANOVA becomes most appropriate. ANOVA, like the t-test-uses a ratio of observed differences/error-learn to test hypotheses which is called the T-ratio (Ary, Jacobs & Razavieh, 1979).

Iii ANOVA, the total variance-of all subjects can be analyzed into variance between groups which is the numerator of the F-ratio and variance within groups which is the denominator or error term.

ANOVA can be computed for a simple experiment which is used to check the relationship between one independent variable and two or more dependent variables. This is called simple or one way analysis of variance. However, a researcher may be interested in analyzing the combined effect of two or more independent variables. This leads to multifactor analysis of variance.

Computation of Simple Analysis of Variance

There are different formulae available for use by a researcher who is computing simple ANOVA. We will discuss only two of these.

Computation of ANOVA Using the ABC Method

An economics teacher collected the following data from a test given to three streams of students and computes the F-ratio like this:

Group 1		Group 2		Group 3	
X1	X_1^2	X2	$X^2{}_2$	X3	$X^3{}_2$
12 10	144 100	9 7	81 49	6 7	36 49
11 7	121 49	6 9	36 81	2 3	4 9
10	100	4	16	2	4
50	**514**	**35**	**263**	**20**	**102**

Step 1:

Calculate A = $\sum X^2$ which refers to summation of all the squares of the scores in the three groups.

= 514 + 263 + 102 = 879

Step 2: Calculate B = $(\sum X)^2$ where

X = Sum of all the scores in the three groups

N = Number of all the cases in the three groups.

$(50 + 35 + 20)^2/,_5$

Step 3: Calculate C = $\frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_2}$

$$= \frac{50^2}{5} + \frac{35^2}{5} + \frac{20^2}{5} = 82.5$$

Step 4: Calculate sum of squares between groups (SS_b) by obtaining the difference between C and B = C - B.

825-735 = 90

Step 5: Calculate sum of squares within groups (SSw) by obtaining the difference between A and C = A - C

879-825 =54

Step 6: Determine the degrees of freedom between groups which isK-1.

3 - 1 = 2

Step 7: Determine the degrees of freedom within groups using the formula N- K.

15-3 = 12

Step 8: Determine the mean sum of squares between groups (MS_b) by dividing SS_b with its $df_b = \frac{SS_h}{df_b}$

$$\frac{90}{2} = 4.5$$

Step 9: Determine the mean sum of squares within groups (MS_w)

$$= \frac{SS_w}{df_w}$$

$$= \frac{54}{12} = 4.5$$

Step 10: Compute the F - ratio using the formula

$$\frac{MS_B}{MS_w}$$

$$\frac{45}{4.5} = 10$$

Table 40: Summary Table of ANOVA

Source of Variance	SS	df	MSS	SD $(\sqrt{MSw})$	F
Between Group (True)	90 (C-B)	2 (K-1)	45 (C-B/K-1)	2.12	10
Within Group	54 (A-C)	12 (n-k)	4.5 (A-C/n-k)		

At this point, we now consult the F - table for the critical value which we will use to either reject or accept the null hypothesis stated.

The assumption underlying the analysis of variance procedure is that if the groups to be compared are truly random samples from the same population, then between - group mean square should not differ from the within-group mean square by more than the amount we would expect from chance alone. Thus under a true null hypothesis we would expect the F - ratio to be approximately equal to one. On the other hand, if the null hypothesis is false, the mean square between groups would exceed the mean square within groups. In such cases, the F-ratio which is the mean square between groups divided by the mean square within groups will have a value .greater than one. We then consult the Table of F-values to, determine whether-the ratio for our data is sufficiently greater 'than; 1.00 to enable us reject the null hypothesis' at our predetermined level. As the difference between these mean squares increases, the F-ratio increases and the probability of the null hypothesis being correct decreases.

When the null hypothesis is rejected as a result of this analysis-of-variance procedure, we cannot say more than that the measures obtained from the groups involved differ and the differences are greater than one would expect to exists by chance.

Computation of ANOVA Using Deviation Method

Suppose we conducted an experiment concerning the effect of three different teaching methods on students' performance in physics, and we wish to compare the performance of the students assigned to these three conditions in a physics quiz. To do this, we follow the procedure below:

Step 1: State the appropriate null hypothesis.
There is no significant difference in the performance of students taught with the three methods

Step 2: Find the sum of squared deviation of each of the student's scores from the grand mean. This is called the total sum of squares. To do this we construct the following table:

Table 41: Scores/Squared Scores of Students Taught with Three Different Methods

Problem Solving Method		***Discussion Method***		***Lecture Method***	
X,	x_1^2	x_2	X_2^2	X_3	X_3^2
20	400	19	361	15	225
20	400 .	18	324	14	196
19	361	18	324	13	165
18	324	18	324	13	165
17	289	16	256	12	144
16	256	14	196	10	100
16	256	12	144	8	64
14	196	12	144	8	64
140	**2482**	**727**	**2073**	**93**	**1123**

X_1 = 17.5; X_2 = 15.9 X_3 =11.6

Total sum of squares ($SS_{t)}$ is computed using the formula.

$SS_t = \sum x^2 - \frac{(\sum x)^2}{N}$

$SS_t = 5678 - \frac{(360)^2}{24} = 278$

Step 3: Compute the sum of squares between (among) groups (SSb)

$SS_t = \frac{\sum x_1^2}{n_1} + \frac{(\sum x)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} - \frac{(\sum x)^2}{N}$

$SS_b = \frac{(140)^2}{8} + \frac{(127)^2}{8} + \frac{(93)^2}{8} - \frac{(360)^2}{24}$

Step 4: Compute the sum of squares within groups (SSw) by using the formula

$SS_w = X_1^2 - \frac{\sum x_1^2}{n_1} + X_2^2 \frac{(\sum x)^2}{n_2} + X^2_3 + \frac{(\sum x_3)^2}{n_3}$

or we can compute SS_w by simply attracting the SS_b from SS_t already computed above.

$SS_w = SS_t - SS_b = 278 – 147.2$

$SS_w = 130.8$

Step 5: Determine the degrees of freedom for between and within group variances. This is done by using the formula: Groups – 1 for between group variable (SS_b) and $N_1 – 1 + N_2 – 1 + N_3 – 1$ for within group variance (SS_w).

Step 6: Divide the SS_b and SS_w with the respective degrees of freedom to get between groups mean square and within-groups mean square (MS) = SS_b/dfb and SS_w/dfw respectively.

Step 7: Prepare a summary table of analysis of variance as in our example below.

Summary table of ANOVA

Source of Variance	ss	Df	MS	SD	F	Level of Significance
Between Groups	147.2	2	73.6			
Within Groups	130.8	21	6.23	2.50	11.81	.05
Total	**275**	***23***				

Step 8: Finally, determine the F-ratio by applying the formula

$$= \frac{MS_b}{MS_W}$$

The F-ratio for our example above becomes = 11.81.

Step 9:

Consult the table of F-ratio to determine whether the F-ratio obtained is statistically significant. In the table of values for F-ratio, there are two degrees of freedom on vertical and horizontal sides. The degree of freedom for greater mean square (MS) is on the horizontal side while that for the smaller mean square (MS) is on the vertical side. We will use the two degrees of freedom to obtain our critical F-ratio.

At the intersect of each of the two "degrees of freedom are-"two" values, one in Roman type arid "the other in Bold face type. 'The one in Roman type is the critical value when one is testing at .05 "level of significance while the one in Bold face-type is the critical value while testing at .01 level of significance.

For our example above, with 2 and -21 degrees of freedom and testing at .05 a level, the critical value is 3.47 and our obtained value is 11.81. This means that we will reject the H_0 at .05 a level since our obtained F-ratio is greater than the critical F-ratio. Our F-ratio therefore is significant.

A significant F-ratio does not necessarily mean that all 'groups' differ" significantly from each other. Group 1 maybe significantly different from Group 2 and Group 3 while Group 2 may not differ significantly from Group 3.

To actually find the location of the significant differences, we apply the standard deviation "pooled" formula, or Scheffe formula, etc to determine the significance between pairs of individual measures.

Comparison Between Pairs of Mean

Whenever one finishes computing the F-ratio and it is significant, this indicates that:

i. The sample group' means may probably not be from populations with identical means or
ii. The population means of at least two groups in the set of groups used for analysis may be significantly different.

The problem now is how to identify the pairs which are significantly between use of problem solving method and discussion method -one pair; or between discussion method and lecture method - another pair and so on.

A comparison between two means can be done by using different formulae namely Scheffe test, Tukey's test, SD pooled formula or the simple formula written below:

$$F\ \text{Comparison} = \frac{(\bar{X}_1 - \bar{X}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MS_w}$$

Where

X1 and X2 are the means of the two groups (a pair) being compared n_1 and n_2 are the sample sizes of the pair. MS_W is the mean square within groups from the ANOVA computed.

For our example above, taking Group 1 and Group 2, the F comparison

$$\frac{(17.5 - 15.9)^2}{(\frac{1}{8} + \frac{1}{8})6.23} = 1.644$$

We will then go to F-ratio table to determine the critical value using the degrees of freedom for between groups = 1 and within groups - 21. At .05 a level, the value is 4.32 which is greater than the obtained F-comp. of 1.644.

This indicates that the difference between Group: 1 and 2 is not significant at .05 level. The significance of difference may be between Groups 1 and 3, or 2 and 3. We can apply the same formula to determine which ones differ significantly.

Analysis of Covariance (ANCOVA)

In many experimental research designs, especially in behavioural studies and educational matters, researchers often desire to compare groups that are not initially alike in either the variable under study or some other related variables. In such situations, subjects are placed into experimental and control groups and comparison is made concerning the two groups after the experiment on the variable of interest. There is the need to ensure that the two groups were equivalent on the research variables before the experiment to avoid the intrusion of extraneous variable which confounds the result of a study. Many ways of ensuring the equivalence have been discovered. Some of these are:

i. Making the experimental and control groups equivalent by person - to - person matching. This method leads to sharp reduction of the size of the subjects and its application is not easy.
ii. Matching groups initially for mean and standard deviation in one or more related variables. This method also leads to sharp reduction of size of N

Analysis of covariance (ANCOVA) was developed to solve the problem of non equivalent groups. ANCOVA is an extension of analysis of variance (ANOVA) which makes room for the correction between initial and final scores. ANCOVA equates the control and experimental groups statistically by enabling researchers to effect adjustments in final or terminal scores while at the same time allowing for differences in some initial variable.

Researchers, who for some reasons, find it impossible or quite difficult to equate control and experimental groups at the start, can easily apply ANCOVA during their data analysis at the end of the experiment in order to achieve the much needed equivalence.

Calculation of ANCOVA Using ABC Method

Example: Suppose 12 subjects were given a pretest (X), four of them were then assigned randomly to three groups. A, B, C and an experiment conducted. After three weeks wherein groups A and B were subjected to different treatments while group C was the control group, the test was repeated. To test whether the differences in their performances in the pre and post tests were significant, we follow the procedure below:

Table 43: Data of Groups A, B and C Used for an Experimental Study

Group A					*Group B*					*Group C*				
X_1	X_1^2	Y_1	Y_1^2	X_1Y_1	X_2	X_2^2	Y_2	Y_2^2	X_2Y_2	X_3	X_3^2	Y_3	Y_3^2	X_3Y_3
20	400	40	1600	800	30	900	33	1089	990	10	100	15	225	150
15	225	30	900	450	15	225	17	289	225	15	225	20	400	300
30	900	35	1225	1050	20	400	25	625	500	25	625	25	625	625
10	100	20	400	200	20	400	15	225	300	10	100	15	225	150
15	225	25	625	375	15	225	15	225	225	15	225	15	225	225
$\sum=90$	1850	150	4750	2875	100	2150	105	2453	2270	75	1275	90	1700	1450
$\bar{X}_1=18$		$\bar{Y}_1=30$			$\bar{X}_2=20$		$\bar{Y}_2=21$			$\bar{X}_3=15$		$\bar{Y}_3=18$		

Step l:

Compute the sum of the scores and sum of squares for all groups.

For all 3 groups IX

$\sum X$ = 265

$\sum Y$ = 345

$\sum x2$ = 5275

$\sum Y2$ = 8903

$\sum XY$ = 6595

Step 2: Compute A for all variables

X = $\sum X^2$ = 5275

Y = $\sum Y^2$ - 8903

XY = $\sum XY$= 6595

Step 3: Compute B for all variables

X = $\frac{(\sum x)^2}{N}$ = $\frac{(265)^2}{15}$

Y = $\frac{(\sum Y)^2}{N}$ = $\frac{(345)^2}{15}$

XY = $\frac{(\sum X \sum Y)}{N}$ = $\frac{265 \times 345}{15}$ = 6095

Step 4: Find C for all variables

X = $\frac{(\sum x)^2}{n_1}$ + $\frac{(\sum x_2)^2}{n_2}$ + $\frac{(\sum x_3)^2}{n_3}$ = 4745

Y = $\frac{(\sum Y_1)^2}{n_1}$ + $\frac{(\sum Y_2)^2}{n_2}$+ $\frac{(\sum Y_3)^2}{n_3}$ = 8325

XY = $\frac{\sum X_1 Y_1}{n_1}$ + $\frac{\sum X_2 Y_2}{n_2}$ = $\frac{\sum X_3 Y_3}{n_3}$ = 6150

Step 5: Determine sum of squares between groups (SS_b**)** by using this method = C – B

For X = 4745 - 4681.7 = 63.3

Y = 8325 - 7935 = 390

XY = 6150 – 6095 = 55

Step 6: Determine sum of squares within groups (SS_w**)** by using the formula:

A - C

For X = 5275 – 4745 = 530

Y	=	8903 – 8325	=	578
XY	=	6595 – 6150	=	445

Step 7: Build a summary table of ANOVA before going on to ANCOVA

Table 44: Summary Table of ANOVA

Sources of Variation	***df***	***SSx***	***SSy***	***SSxy***	***MSSx***	***MSSxy***	***Fx***	***Fy***
Between groups	2	63.3	390	55	31.65	195		
Within groups	12	530	578	445	44.17	48.17	0.717	4.05
Total		**593.3**	**968**	**500**				

Note: To calculate mean sum of squares for x and y - MSSx and MSSy - respectively, we divide SSx and SS-y by their respective degrees of freedom. This is done for both variations i.e. between and within groups variations. To get Fx and Fy, we divide the mean sum of squares for between groups with that of within groups for both x and y variables respectively.

Step 8: Using the degrees of freedom for between and within groups variables which are 2 and 12, in our example above, we go to the F table of values and determine the critical value of F for our analysis. At .05 level of significance, the F critical value is 3.88, Comparing this value with our calculated F ratios, we find out that the calculated Fx value of .717 is smaller than the critical value of 3,88. This shows that the experiment was quite successful in getting random samples in Group A, B,C.

Step 9: Graduate to ANCOVA by computing adjusted sum of squares for y i.e. SSy.x using the formula

$$SS_y - \frac{(SSxy)^2}{SSx}$$

For total sum of squares (SStT), SSy.x becomes

$$968 - \frac{(500)^2}{593.3} = 546.6$$

For sum of squares within groups (SS_w), SSy.x is

$$578 - \frac{(445)^2}{530} = 204.4$$

For sum of squares between groups (SS_b), SSy.x is

$$SS_T - SS_w = 546.6 - 204.4 = 342.2$$

Step 10: Prepare summary of ANCOVA

Table 45: Summary Table of ANCOVA

Source of Variation	***df***	**SSy.x**	**MSSy,x**	**SD**	**Fy.x**
Between groups	2	342.2	171.1	4.13	
Within groups	12	204.4	17.03		10.05
Total	**14**	**546.6**			

Note: The standard deviation is determined by getting the square root of the mean sum of squares of adjusted y.x for within groups variance i.e. 17.03. The F table value at .05 level is 3.88. This is lower than the calculated value of 10.05 which shows that there is a significant difference between the groups. We now need to do multiple comparison in order to determine where the significant difference lies. To do this, we need to use the means of the pre and post tests. For ANCOVA, we have to adjust these means. This is done in step 12 below.

Step 11. Calculate the regression or beta coefficient = SSxy/SSx

For total sum of squares = b_{tot}ai 500

593.3
.84

between groups
SS = between group 55
.63.3
.87

within groups SS = within groups 445 .

530
*- = . .84

Step 12: Calculate adjusted y means

Groups	**N**	**MX**	**My**	**My.x (adjusted means)**
A	*5*	*18*	*30*	*29.7*
B	5	20	21	19.1
C	*5*	*15*	18	20.3
General Means		**77.7**	**23'**	

To get My.x, use the formula

My - b (Mx - GMx)

Note: b is the regression coefficient of within groups.

For Group A My - bx = 30-.84. (18-17.7)

= 29.7

B My – bx = 21-.84 (20-17.7)

= 19.1

C My – bx = 15 - .84(15-17.7)

= 20.3

To do multiple comparison, you can use any of the formula mentioned above in the section on multiple comparison under ANOVA. For our example here, we will use standard deviation pooled formula (SDP).

Step 13:

Calculate significance of differences among adjusted y means.

To do this, pool the standard deviation from ANCOVA table. Then apply the formula.

SEmy.x = $\frac{4.13}{\sqrt{5}}$

= 1.85

Calculate SDP

SDP between any two adjusted means

$$= SDy.x \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$= 4.13 \sqrt{\frac{1}{5} + \frac{1}{5}}$$

= 4.13 x .63

= 2.602

Go to the table value of t - because in multiple comparison, you have changed from F ratio to t test where you are determining the significance of differences between two means - and copy the critical value of your test. For our example, t .05 = 2.18. Significant difference at .05 level

= 2.18 x 2.602

= 5.67

Use the result above to compare, the adjusted mean differences between A and B, B and C, A and C - in order to determine which group difference is significant. For example, the adjusted mean differences of A and B = 29.7 - 19.1 = 10.6. This indicates that the mean difference between A and B is significant, because 10.6 is greater than the critical value of 5.67.

Calculation of ANCOVA Using Correction Term Method

Given the data in our Table 43 above, the following procedure will be followed to compute ANCOVA with correction term method.

Step 1: Calculate the sum of the scores and sum of squares for all groups.

For all 3 groups:

$\sum x = 265$

$\sum y = 345$

$\sum x2 = 5275$

$\sum y2 = 8903$

$\sum xy = 6595$

Step 2: Calculate corrections terms (C) for x, y and xy respectively. The correction term is needed in order to computer sum of squares (SS's) for x, y and xy.

$$Cx = \frac{(\sum X)^2}{N} = \frac{(265)^2}{15} = 4681.7$$

$$Cy = \frac{(\sum Y)^2}{N} = \frac{(345)^2}{15} = 7935$$

$$Cxy = \frac{(\sum X \sum Y)}{N} = \frac{265 \text{ x } 345}{15} = 6095$$

Step 3: Calculate total sum of squares (SST) for X, Y and XY

For X $= \sum X^2 - CX = 5275 - 4681.7 = 593.3$

Y $= \sum Y^2 - CY = 8903 - 7935 = 968$

XY $= \sum XY - CXY = 6595 - 6095 = 500$

Step 4: Calculate sum of squares between groups (SSB) for all variables.

$$X = \frac{\sum X_1^2 + \sum X_2^2 + \sum X_2^3}{n} - CX$$

$$= \frac{(90)^2 + (100)^2 + (75)^2}{5} - 4681.7$$

$$= 63.3$$

$$Y = \frac{\sum Y_1 2 + \sum Y_2 2 + \sum Y_3 2}{n} - CY$$

$$= \frac{(150)^2 + (105)^2 + (90)^2}{5} - 7935$$

= 390

$$XY = \frac{\sum X_1 \sum Y_1 + \sum X_2 Y_2 + \sum X_3 \sum Y_3}{n} - CXY$$

$$= \frac{90 \times 150 + 100 \times 105 + 75 \times 90}{5} - 6095$$

= 55

Step 5: Compute sum of squares within groups (SS_w) for all variables. This can be done by subtracting SSB from SST for all the three variables.

For X = 593.3 – 63.3 = 530

Y = 968 – 390 = 578

XY = 500 – 55 = 445

Step 6: Make a summary of ANOVA table of the X and Y trials taken separately

Table 46: Summary Table of ANOVA

Sources of Variation	***df***	***SSx***	***SSy***	***SSxy***	***MSSx***	***MSSxy***	***Fx***	***Fy***
Between groups	2	63.3	390	55	31.65	195		
Within groups	12	530	578	445	44.17	48.17	0.717	4.05
Total	**14**	**593.3**	**968**	**500**				

f table value at .05 level = 3.88

Note: Mean sum of squares (MSS) for X and Y respectively is determined by dividing the sum of squares with their respective degrees of freedom = SSx/dfx and SSv/dfy. F ratio is calculated by dividing MSSg with MSS_H, for X and Y respectively.

The F test applied to the initial (X) scores (Fx = .72) falls far short of significances at the .05 level, from which it is clear that the x means do not differ significantly and that random assignment of subjects to the three groups was quite successful. The F test applied to the final (y) scores (Fy = 4.05) is significant at .05 level.

Step 7: Compute Adjusted SS for y ie Ssy.x

The computations carried out in this step are for the purpose of correcting the final (y) scores for differences in initial (x) scores.

The symbol SSy.x means that the SSy have been "adjusted" for any variability in y contributed by X, or that the variability in X is held constant.

The general formula is $SSy.x = SSy - \frac{(SSxy)^2}{SSx}$

For $(SS_T) = 968 - \frac{(500)^2}{593.3} = 546.6$

$(SS_W) = 578 - \frac{(445)^2}{530} = 204.4$

$(SS_B) = SS_T - SS_w$

$= 546.6 - 204.4 = 342.2$

From the various adjusted sums of squares the variances (Msy.x) can now be computed by dividing each SS by its appropriate df. The value of analysis of covariance becomes apparent when the F test is applied to the adjusted between and within variance (Fy.x).

Step 8: Make a Summary Table of ANCOVA

Source of Variation	*Df*	*SSy.x*	*MSSy,x*	*SD*	*Fy.x*
Between groups	*2*	342.2	171.1	4.13	10.05
Within groups	12	204.4	17,03		
Total	***14***	***546.6***			

The F table value at. 05 level is 3.88.

Comparing the calculated Fy.x of 10.05 with the F critical value of 3.88, we find that the calculated F is highly significant. To find which of the three possible differences is significant or whether all are significant (i.e. multiple comparison), we must apply the t test. To do so, we will follow the same procedure in steps 11 - 13 of the first formula above.

Exercises

1. What assumptions have to be met before a researcher can use parametric or non-parametric statistics?

2a. Differentiate between t test inferential statistics and ANOVA,

b. Given the data below, test an appropriate hypothesis using t test statistics at .01α level.

X_1 - 34; S_1 = 4.08 N_1 = 14

X_2= 28; S_2= 3.12 N_2 = .16

3. A researcher wanted to investigate the relative changes in student's performance in mathematics after introduction of a new instructional mode. He collected pretest result of the students, introduced the change and collected post test result as follows.

Pretest:	*48*	*52*	*61*	*34*	*38*	*72*	*55*
Posttest:	*'60*	*60*	*72*	*58*	*64*	*74*	*65*

Test an appropriate hypothesis using t test statistics at .05 a level.

4a. Analysis of variance is more versatile than t test, discuss. Given three sets of results collected from year three students in three different courses, determine the significance of difference between the performance of students in the courses using ANOVA statistics. The results are as follows:

A:	17	15	16	16	14	12	13	12	13	17	17
B:	20	19	17	17	18	14	14	15	16	19	12
C:	13	12	15	17	20	14	15	16	16	17	12

5. In an experiment, the researcher collected the data below:

Group 1		***Group 2***		***Group 3***	
Pretest	*Posttest*	*Pretest*	*Posttest*	*Pretest*	*Posttest*
14	35	20	*26*	6	*12*
12	25	12	14	12	*11*
20	22	13	18	22	22
6	14	14	10	6	12
12	25	10	8	10	10

Using the analysis of covariance, test the significance of the differences among the groups, setting your alpha level at .05.

6. The final examination scores of eight students in statistics 1 and statistics II are:

Statistics I	***Statistics II***
86	80
72	58
78	67
66	65
92	90
56	52
54	*45*
65	40

Using a 5% level of significance, test the null hypothesis that there is no correlation between the statistics I final exam grade and the statistics n final exam grade.

CHAPTER EIGHT

NON PARAMETRIC STATISTICS AND TESTS

Statistical procedures are described as non parametric statistic or distribution free statistics if no useful reference is required about the population parameters. Also, the distribution describing the sampled population is not known to exist. Statistical tests carried out when the above fact(s) prevail are referred to as non parametric tests or distribution free tests (Freud, 1992).

Examples of Non Parametric Tests are

1. Mann Whitney Utest
2. Sign test
3. Wilcoxon Signed Rank test
4. Run's test
5. Chi Square test of independence, etc,

The above mentioned tests will be discussed in this chapter.

Other Assumptions of non Parametric Tests are:

1. In situations where a researcher wants to analyze data on the ordinal or nominal scale, non parametric tests are used because parametric tests are not very appropriate in such situations.
2. Where departures from normality or homogeneity of variance are severe, non parametric tests are more appropriate.

Many researchers prefer parametric tests to non parametric which explains its unpopularity and non frequency of use in analysis of data. Many researchers also do not know much about the computational procedures of non parametric tests or the particular tests that are non-parametric. At times one finds researchers using-a non parametric statistics for data that require parametric test and vice versa. However, two of the non parametric tests have been commonly applied by researchers over the years in. data analysis, though many may not know that they are non parametric. These are Pearson Chi Square test and Spearman Rank c.c. test,

Mann Whitney U Test

Mann Whitney U test is used to test the null hypothesis that two samples come from the same population.

The rank of the observations forms the basis of the-test. It makes use of more information contained in the data than Median test which focuses on the median rather than the mean (Zuwaylif, 1979). Whenever each of the sample size is at least.8, the distribution of U approximates normal.

Hence, $$Z = \frac{u - \mu u}{6u}$$

Is normally distributed with mean = 0 and variance = 1.

Where z is the standardized value

$$U_f = N_f N_s + \frac{N_f (N_f + 1)}{2} - R_f$$

or

$$U_s = N_f N_s + \frac{N_s (N_s + 1)}{2} - R_s$$

$$\mu u = \frac{N_f N_s}{2} \quad \ldots (3)$$

$$6u = \frac{N_f N_s (N_f + N_s + 1)}{12} \quad \ldots (4)$$

Where

N_f is the size of the first sample.

N_s is the size of the second sample.

Rf. is the ranksum of second sample after joint ranking.

U = U_f, or U = U_s = the computed Mann-Whitney Statistic.

Example 1:

Mrs. Uju has two classes in Social Studies, JS IA of 11 students and JS IB of 13 students. She scheduled examination at the same time for all the students. Their scores were as shown in Table 48. Can one conclude at 5% significant level that the students were drawn from the same population?

Table 48: Scores of JS I Students in Social Studies

JS 1A	13	24	32	17	90	33	34	10	45	44	47		
JS 1B	19	25	15	29	40	84	23	29	75	11	45	27	70

Steps to be followed:

Step 1: Arrange the scores of the two classes jointly in ascending order or descending order and rank them accordingly

Score	10	11	13	15	17	19	23	24
Rank	1	2	3	4	5	6	7	8

Score	25	27	29	29	32	33	34	40
Rank	9	10	11.5	11.5	13	14	15	16

Score	44	45	45	47	70	75	84	90
Rank	17	18.5	18.5	20	21	22	23	24

Step 2:

Regroup the scores according to their former classes attaching to each score its rank number as shown on Table 49.

NB: When a score occurs more than once, the rank of each is the average of what their individual ranks should have been, if the scores were different.

Table 49: Ranks of Students' Scores in Social Studies

First Sample (JS 1A)		***Second Sample (JS IB)***	
Score	***Rank***	***Score***	***Rank***
13	3	19	6
24	8	25	9
32	13	15	4
17	5	*29*	11.5

90	24	40	16
33	14	84	23
34	15	23	7
10	1	29	11.5
45	18.5	75	22
44	17	11	2
47	20	45	18.5
		27	10
		70	21
	R_f = 138.5		***R_s = 161.5***
	Nf = 11		***N_s = 13***

$$= \frac{143}{2}$$

$$= 71.5$$

$$6u = \frac{N_f N_s + N_f (N_s + 1)}{12}$$

$$= \frac{11(13)(11 + 13 + 1)}{12}$$

$$= 17.26$$

Step 7: Computation of Z by the aid of Mann Whitney U Statistics from First Sample

$$Z = \frac{U_f - \mu u}{6_u}$$

$$= \frac{70.5 - 71.5}{17.26}$$

$$= \frac{-1}{17.26}$$

$$= -0.06$$

From the standard normal table, we have:

$Z \leq 1.96$ or $Z \geq -1.96$

$-1.96 \leq Z \leq 1.96$

Since the computed Z falls within the interval defining the area under normal standard curve, we accept H_o and conclude that the students are selected from the same population.

Suppose we apply the Mann Whitney U Statistics of second sample, we have:

$$Z = \frac{U_s - \mu u}{6_u}$$

$$= \frac{72.5 - 71.5}{17.26}$$

$$= \frac{1}{17.26}$$

$$= 0.06$$

Again, the computed Z = 0.06 falls within the interval, $-1.96 \leq Z \leq 1.96$. Hence, we accept H_o and conclude that the students come from the same population.

Sign Test

Sign test is a non-parametric alternative to one sample t-test. It is used as a test tool where t-test cannot be applied. For instance, when we cannot assume that the data we have are normally distributed, t-test may not be used. Also sign test emphasizes median rather than mean and its computation in based on signs (+ and -_) instead of numerical values. It is very suitable to tests involving match (paired) samples, but can also be used on equal independent samples (Frend, 1992).

To apply or conduct Sign Test, we note that:

1. the population is assumed to be continuous to make room for zero probability
2. the hypothesis that $\mu = H_o$ where μ is the population mean, and μ_o is the sample mean implies that sample value greater than μ_o takes +, and – for each sample value less than μ_o. If sample value equals μ_o it is discarded.
3. the test statistic X is the number of plus or minus depending on whether Ha bears > or < (greater than or less than symbol).

4. small sample size necessitates the use of binomial approach
5. if n ≥ 10, the sample size may be considered large, then normal approximation of binomial distribution.

Example:

A teacher was interested in determining whether a new teaching method would increase students' achievement in mathematics. The old and new methods were used on the same sample of 12 students at different time intervals. The test scores under the two methods are listed below.

Teaching Methods

Students	**New Method**	**Old Method**
1	3.7	4.1
2	4.2	3.8
3	5.1	6.7
4	6.3	6.3
5	5.3	5.4
6	4.7	5.2
7	6.2	5.3
8	6.3	5.7
9	6.7	4.6
10	7.1	7.6
11	4.2	4.2
12	4.8	3.5

She used the sign test to test at 5% significance level the null hypothesis that the new method yields achievement scores that are at least as much as those yielded by the old method. In conducting the test, she followed the procedure below:

Step 1

Each of the 12 students' scores under old and new methods are compared. Old method score is removed from the new method score. Plus sign (+) is indicated for positive difference and minus (-) for negative difference.

Step 2: Hypothesis Ho N1 = N2

Ha N1 > N2

A = 5%

Assumption:

There are equal number of plus (+) and minus (-). Hence if P is the proportion of plus signs, we would expect P to be around 0.5.

Table 50: Students' Scores under New and Old Teaching Methods

Students	New Method	Old Method	Sign of difference
1	3.7	4.1	-
2	4.2	3.8	+
3	5.1	6.7	-
4	6.3	6.3	
5	5.3	5.4	-
6	4.7	5.2	-
7	6.2	5.3	+
8	6.3	5.7	+
9	6.7	4.6	+
10	7.1	7.6	-
11	4.2	4.2	
12	4.8	3.5	+

Step 3: The test statistics for testing proportion is defined below:

$$Z = \frac{P - p}{\frac{p(1 - p)}{n}} \quad \ldots 4.1$$

From table 50, we have n = 10, where n = Total no. of signs. p = total proportion of plus (+) signs as specified in the null hypothesis. This means that if the new method does not affect a students' score we should expect as many plus signs as minus signs. Therefore p = 0.50. We take the minus or plus signs and ignore the cases that have no change.

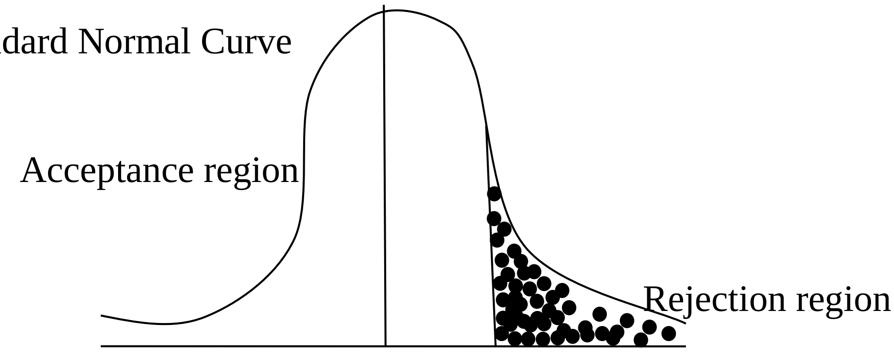
$$p = \frac{\text{No. of plus}}{\text{Total no of signs}} \qquad p = \frac{5}{10} = 0.5$$

Step 4: The computation of z becomes

$$Z = \frac{0.5 - 0.5}{\frac{0.5\,(1 - 0.5)}{10}} = 0$$

Step 5: Decision

A standard Normal Curve



From the standard normal curve we can see that Z = 0 falls in the acceptance region for the one tailed test. Since the computed Z = 0 < table value Z at .05 level which is 1.64, we accept Ho and conclude that the new teaching method yields the same achievement score as the old teaching method.

Wilcoxon Signed – Rank Test

In sign test, only signs are considered in tests conducted on both single sample and paired samples. Signs define direction. Wilcoxon signed rank test utilize direction and magnitude in the performance of test. Hence, it is an improvement on sign test and relies on the difference between matched pairs.

Test Procedure

Step 1: Obtain the differences between all sample pairs.

Step 2: Rank the absolute values of the differences

Step 3: Affix the sign of each difference to the corresponding rank; discarding zero difference

Step 4: The sum, W of the ranks with the less frequently occurring sign is calculated for use as the test statistic to test the same null hypothesis associated with the sign test

Step 5: The normal approximates of the sampling distribution of T is defined at:

$$Z = \frac{T - \mu T}{6_T}$$

Where $\mu_T = \frac{n(n + 1)}{4}$ i.e. the mean

$$\sigma_T = \frac{n(n + 1)(2n + 1)}{24}$$

Example 1:

Eight students were told to rate two of their lecturers in terms of responsibility to duty. Each students was told to rate these lecturers on a scale of 1 to 9 with higher rating implying more dutifulness. The above rating involves paired appraisal. This leads to an analysis that requires computing differences between measurements. Using Wilcoxon Signed rank test, we follow the procedure below:

Table 51: Ratings of Students Concerning their Lecturers' Work Responsibility

Students	**Lecturers**		**Differences**	**Absolute value of Difference**	**Rank of Absolute Value of Difference**	**Signed Rank**
	A	B	(A – B)	(D)	(D)	
1	5	6	-1	1	1	-1
2	6	4	+2	2	4	+4
3	4	5	-1	1	1	-1
4	7	9	-2	2	4	-4

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

T+ = Sum of Positive Ranks = 23

T- = Sum of Negative Ranks = 6

Procedure for Computation

Step 1: Determine the difference between the students' ratings of the two lecturers by subtracting the rating of lecturer B from that of lecturer A. These differences may be positive, negative or zero.

Step 2: Determine the absolute values.

Step 3: Rank these absolute values in order from the lowest to the highest

Step 4: Give each rank a plus (+) sign or minus (-) sign in correspondence to the sign of the original measurement

Step 5: Sum the positive ranks and the negative ranks separately

Step 6: State the null and alternative hypothesis

In our example above:

H_o: The ratings of students for the two lecturers are not significantly different

H_a: The ratings of students for the two lecturers differ significantly

(The alternative hypothesis above is two sided which implies a two tailed test)

Step 7: The test statistics is the sum of the ranks with the smaller sum or with less frequent sign. If the H_o is correct, the sum of the positive ranks and the sum of the negative ranks are expected to balance. If otherwise, then there is a likelihood that the Ho will be rejected.

In our example, the sum of the ranks is +23 for positive sign ranks and -6 for negative sign ranks

Step 8: Select the smaller sum of the ranks or ranks with less frequent sign and in our example above, this is -6. Compare this test statistic value with the critical value for Wilcoxon signed Rank test in the Appendix. Using $n = 8$ and $\alpha = 0.05$, if the critical value is larger than or equal to the obtained value we reject the null hypothesis. On the other hand, if the critical value is less than the obtained value we do not reject the null hypothesis.

Note: Where there is a difference of zero, this value is not used in performing the calculations of Wilcoxon signed rank test as it cannot be considered as positive or negative

Note: Whenever the n of the sample of either of the paired observations is at least 8, the distribution of T approximates normal. This necessitates the calculation of the Z which is $\frac{T - \mu T}{6_T}$

Where $\mu r = \frac{n(n+1)}{4}$ i.e. the mean

ƪT is the standard deviation of the signed rank test

$$= \frac{n(n+1)(2n+1)}{24}$$

$$Z = \frac{T - \mu T}{ƪT}$$

for the example above:

$$\mu T = \frac{8(8+1)}{4}$$

$$= \frac{72}{4} = 18$$

$$6T = \frac{8(8+1)(2 \times 8 + 1)}{24} = 7.14$$

$$Z = \frac{23 - 18}{7.14} = 0.42$$

α at .05 is Z ± 1.96

Discussion: Where the computed Z is less than the critical z, we accept the H_o as in our example above. If the reverse is the case, we reject the H_o.

Example 2:

Mr. Aja gave end of term examination to students in JS2 classes of St. Charles College, Ankpa on English Language. He selected two classes randomly from the stream. The students in the chosen classes JS 2A and 2B had the scores presented below.

JS 2A	13	40	15	60	70	83	05	44	50	76	66	26	62
JS 2B	28	41	34	8	32	45	45	65	53	95	11	30	32

Using Wilcoxon sign-rank test to test null hypothesis that there is no significant difference in the performance of the students at 5% level of significance.

Step 1:

Preparation of table showing the scores, difference, ranking in absolute term, signed ranks.

Table 52: Scores/Ranks of Students' Scores in English Language

S/N	**JS 2A (U_1)**	**JS 2B (U_2)**	**Difference ($U_1 – U_2$)**	**Rank**	**Signed Rank**
1	13	28	-15	4	-4
2	40	41	-1	1	-1
3	15	34	-19	5.5	-5.5
4	60	08	+52	12	+12
5	70	32	+38	9.5	+9.5
6	83	45	+38	9.5	+9.5
7	05	45	-40	11	-11

8	44	65	-21	7	-7
9	50	53	-3	2	-2
10	76	95	-19	5.5	-5.5
11	66	11	+55	13	+13
12	26	30	-4	3	-3
13	62	32	+30	8	+8

From Table 50, No. of + = 5

$\sum T^{+} = 12 + 9.5 + 9.5 + 13 + 8 = 52$, $n = 13$

Step 2: Hypothesis: $H_o: \mu_1 = \mu_2$

Ha: $\mu_1 \neq \mu_2$

$\alpha = 0.05$

Step 3: $\mu_T = \frac{n(n+1)}{4}$

$$= \frac{13(13+1)}{4}$$

$$= \frac{182}{4}$$

$$= 45.5$$

$$\sigma_T = \frac{n(n+1)(2n+1)}{24}$$

$$= \frac{13(13+1)(2 \times 13+1)}{24}$$

$$= \frac{(182)(27)}{24}$$

$$= 14.31$$

$$Z = \frac{T - \mu_T}{\sigma_T}$$

$$= 52 - 45.5$$

14.31

= 0.45

But Z 0.025 = 1.96

Computed Z, Zc = 0.45

Step 4: Decision

Since the computed Z, Zc = 0.45 < Z tabulated, Z 0.025 = 1.96, we accept Ho and conclude that the students performed equally well. In other words, there is no significant difference between JS 2A and JS 2B students.

Runs Test of Randomness

Random is a common word often used to refer to an arrangement or situation or which event evolves without favour or partiality or influence to the order or part of the arrangement, situation or event his data? A test statistics called Runs test has been developed for determining whether a sample is randomly selected or not.

Run is viewed from the angle of randomness. A run can be defined as a succession of identical letters (or other kinds of symbols) which is preceded and followed by different letters or no letters. It is a sequence which is considered non-random if there are either too many or too few runs and random otherwise (Spiegel, Schiller & Scrimvasan, 1980). For example, the sequence below may be taken as a run, composed of b's and a's.

bbb	a a	b b	a	b b b	a a a a a	b	a a a a	b b b
1	2	3	4	5	6	7	8	9

The sequence is made up of 9 runs described by the braces numbering 1 to 9. The sequence can be summarized as being made up of N_1 a's and N_2 b's making up N symbols i.e. $N_1 + N_2 = N$ in the whole arrangement. We have to note that when using the run's test, the length of each individual run is not important. What is important is the number of times that each letter appears in the entire sequence of letters. In our example above, there are 24 cases made up of 12 b's and 12 a's. In order to test whether the sample is random, we go to the critical values for the total number of runs in Appendix A.

To be able to determine the larger of the n_1 and n_2 samples where n1 represents the samples of one kind and n_2 represents the samples of another kind we consult the table of critical values. On the table of critical values, the larger of n_1 and n_2 is on the vertical axis. At the intersect of the two are two numbers. These represent the critical intersect of the two are two numbers. These represent the critical values. However, where n1 and n2 are exactly equal in our own case, we take the intersect of the two figures not minding which one is on the horizontal or vertical axis.

If the number of runs is between the two numbers at the intersect, we do not reject the null hypothesis. If the number of runs is less than the small number or more than the large number at the intersect, we reject the null hypothesis. This means that we do not have a random sample. In our example, the critical values are 7 and 19. Because the number of our runs is between 7 and 19, we do not reject the null hypothesis. The sequence in our example has an associated number of runs denoted by the statistic V whose sampling distribution has mean and variance defined as:

$$\mu_v = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2N_1N_2}{N} + 1$$

Where μ_1 is the mean

$$\text{l}v^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2 (N_1 + N_2 - 1)}$$

$$= \frac{2N_1N_2 (2N_1N_2 - N_1 - N_2)}{(N)^2(N - 1)}$$

Where $\text{l}v^2$ = Variance

$$\sigma v^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N)^2\,(N - 1)}$$

σv^2 = Standard deviation

The sample size is large if N_1 and N_2 are respectively at least equal to 8. When the sample size is large, the sampling distribution of V is closely approximate by normal distribution and defined by:

$$Z = \frac{V - \mu v}{6v}$$

Example 1: The following is the order in which a teacher obtained Pass (P) and Fail (F) in a test he gave to his students.

P P F F F F F F F F P P P P F F F F P F P

Test for randomness at 5% level of significance

Solution:

Step 1: Identify the number of runs (v), N_1, N_2 and N

Step 2: State Hypothesis

Step 3: Compute the mean (μv), standard deviation (σv) and Z – computed (Zc)

Step 4: Make decision and draw conclusion

Hence:

Step 1: V = 7, N1 = 8, N2 = 13, N = 21

Step 2: Hypothesis: Ho: Arrangement is random

Ha: Arrangement is not random

α = 5%

Step 3:

$$\mu v = \frac{2N_1N_2}{N} + 1$$

$$= \frac{2.8.13}{21} + 1$$

$$= 10.9$$

$$\text{ʅ}v = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N)^2 \ (N - 1)}$$

$$= \frac{2.8.13\ (2.8.13\ - 8\ - 13)}{(21)2 \quad (21 - 1)}$$

$$= \frac{38896}{8820}$$

$$= 2.1$$

$$Zc = \frac{V - \mu v}{\text{ʅ}v}$$

$$= \frac{7 - 10.9}{2.1}$$

$$= \frac{-3.9}{2.1}$$

$$= -1.86$$

$$Z_{0.05} = \pm 1.96$$

$$-1.96 \le Z \le 1.96$$

Since Zc = -1.86 falls within the interval, we accept H_o and conclude that the order was random

Example 2:

The loading times, in hours, of passengers at the students' bus stop in Temporary Site of Nnamdi Azikiwe University (NAU) are obtained as follows:

3 2 5 7 4 3 7 4 2 5 2 1 2 7

Test for randomness at 5% level of significance.

Solution:

Step 1: Determine the median of the loading time.

Step 2: Represent numbers greater than the median with a and b for numbers less than the median, by so doing form a sequence of 'a's and 'b's.

Step 3: Identify the number of runs (v), N1, N2 and N.

Step 4: State Hypothesis

Step 5: Compute the mean (μv), standard deviation (ʅv) and Z – computed (Zc)

Step 6: Take decision and draw conclusion.

Given: 3 2 5 7 4 3 7 4 5 2 1 1 1 2 1 2 7

Step 1: To determine median, arrange the numbers orderly

1 1 1 1 2 2 2 2 2 3 3 4 4 5 5 7 7 7

$$\text{Median} = \frac{2 + 3}{2} = 2.5$$

a b a a a a a a b a b b b b b b b a

Step 3: V = 7; $N_1 = 9$, $N_2 = 9$, $N = 18$

Step 4: Hypothesis: = H_o: The arrangement is random

= H_a: The arrangement is not random

= H_o: The loading time is random

= H_a: The loading time is not random

Step 5: Computation

$$\mu v = \frac{2N_1N_2}{N} + 1$$

$$= \frac{2.9.9}{18} + 1$$

$$= 9 + 1$$

$$= 10$$

$$ʅv = \frac{2N1N2(2N1N2 - N1 - N2)}{(N)2\ (N - 1)}$$

$$= \frac{2.9.9(2.9.9 - 9 - 9)}{(9)2(8)}$$

$$= \frac{2.9.9\ (162 - 18)}{81(8)}$$

$$= \underline{23328}$$

$$= \frac{648}{6}$$

$$Zc = \frac{V - \mu v}{6v}$$

$$= \frac{7 - 10}{6}$$

$$= -0.5$$

$$Z0.05 = \pm 1.96$$

$$-1.96 \leq Z \leq 1.96$$

Step 6: Since Zc falls within the interval, we accept Ho and conclude that the loading time (arrangement) is random

Chi-Square Test

Nature of data determines to a great extent the type of statistical test that would apply to the test of a given hypothesis. Two major divisions of data are qualitative data and quantitative or count data.

Statistical tests that are of parametric nature are very suitable to test of hypotheses involving quantitative data. Such statistical tests include normal-student t and F – tests. Qualitative data arise from assignment of values to abstract situation. They can be feeling, opinions or state of health of a person or group of persons expressed in categories. Color can be used to distinguish among states of objects or things that cannot be subjected to quantitative measurement. Data arising from the color categorization are also qualitative data.

Qualitative data are measured mainly on nominal scale. Nominal scale implies the assignment of values on the attributes to categorize them into two or more classes. In this case, the procedure for hypothesis testing that would be suitable follow chi-square distribution among other non-parametric procedures. Thus chi-square test is very important for test of hypothesis on qualitative or count data or data consisting of attributes.

Evaluation of Chi-square Test

Emergence of chi-square test as a veritable test procedure is traced to the works of Karl Pearson. Karl Pearson developed chi-square test in 1900 from his knowledge of Z-statistics and its relationship with standardized binomial. He discovered that the square of a single standardized binomial could lead to a chi-square statistic for number of categories, k equal to two. He arrived at the general chi-square statistic by approximating it with the sum of K-1 squared, standardized normal distributions. Karl applied approximation because of his conviction that standardized binomial could be approximated by the standard normal distribution for a large sample (Oyeka, 1992).

Karl Pearson's inquiry commenced with experimental throwing of dice. Then he considered the discrepancies between the observed outcomes and expected values, given the event that a particular side turned up. Pearson noted that adding deviations between observed data and expected values could not yield the extent to which observed values differ from their expected values, as the sum of the scaled deviation yielded zero (Oyeka, 1992).

Therefore, he introduced scaled squared deviations between the observed and expected values. Consequently, Karl Pearson inferred that the same scaled squared deviations could be calculated for any number of possible outcomes.

The sequence of Karl Pearson's argument can be summarized as shown below:

Step 1: Determine the deviation, Oi – Ei for each possible outcome. This is the deviation between the observed (Oi) and the expected (Ei) occurrences. Sum of the deviations is often zero.

Step 2: Each of the deviation is squared, $(Oi - Ei)^2$. The effect of this squaring is that positive and negative deviations do not cancel out as would be the case in step 1. Rather both positive and negative deviations count equally.

Step 3: After squaring each deviations, it is divided by the corresponding expected number of occurrence Ei i.e. $(O_i - E_i)^2/E_i$. This scaling of the squared deviations makes different studies comparable.

Step 4: The scaled squared deviations are added together:

i.e. $\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$ etc

Karl Pearson named the end result of the above procedure a chi-square statistic and represented it as follows:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots\ldots + \frac{(O_k - E_k)^2}{E_k} \quad \ldots\ldots (1)$$

Definition

Chi-square statistic can be defined as the sum of the squared discrepancies between observed and expected frequencies over the expected frequencies for k possible outcomes.

Pearson defined chi-square as a single test that simultaneously compares the observed and expected values for each of the possible outcomes of a certain event. He arrived at the general chi-square (x^2) statistic (equation 1) by squaring and rearranging the Z-statistic in an attempt to test a single outcome as follows:

$$Z = \frac{(X - \pi n)}{(\pi(1 - \pi)\, n^{1/2}}$$

Where X is the observation corresponding to O_{ii} π is the proportion, πn is the mean corresponding to the expected E_{ii}.

Equation 2 tends to show by how many deviations the number of observation X is away from its expected value πn.

When both sides of equation 2 were squared Pearson obtained:

$$Z^2 = \frac{(X - \pi n)^2}{\pi (1 - \pi) n}$$

$$(X - \pi n)^2 \left({}^{1}/\pi n + {}^{1}/_{(1-x)} n \right)$$

$$\frac{(X - \pi n)^2}{\pi n} \qquad \frac{(n - x) - (1 - \pi)n)^2}{(1 - \pi)n}$$

substituting X for O and πn for E and Z^2 for x^2 we have:

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad ..... \quad (3)$$

Hence the sum in equation 3 has a chi-square distribution with 2 degrees of freedom corresponding to the number of independent squared terms added together. The generalization of the equation is as shown in equation 1.

Chi-Square Test's Applications

Chi-square test statistic, x^2, is widely applied in the tests for goodness of fit, independence, homogeneity and in comparison of several proportions.

Chi-Square Distribution and Degree of Freedom

Degree of freedom tends to indicate the number of observations that is at liberty to change when the data are under given rules. In other words, the size of degree of freedom shows the number of observations that is free to vary after certain restrictions are placed on the data. The greater the degree of freedom, the more precisely the chi-square distribution would approximate standard normal distribution.

Chi-Square Test in One Sample Case

One sample case prevails when the number of possible outcomes (N) of a given event is represented in K number of classifications consisting of K cells in a single row or column. Then chi-square statistic (x2) is defined as:

$$X^2 = \sum \frac{(Of - Ef)2}{Ef}$$

Worked Example:

Example 1:

1. The number of students that cheat in a semester examination, who appeared before Examination Malpractice Committee is shown on the table below. Test the null hypothesis that cheating is independent of nature of programme a student runs in a university, at level of significance (α) = 5%.

Table 53: Number of Cheats in Different University Programmes

	CEP	**SANDWICH**	**DIPLOMA**	**REGULAR**
Cheats	25	22	21	24

Steps to be followed:

1. Note that it is a case of one sample having one row and many columns.
2. State the hypothesis symbolically or in words whichever is most convenient.
3. Put the observed frequencies into k categories.
4. Given, Ho and Ha, obtain the expected frequencies (Ef's) for each of the k cells. In one sample case, Ef = $^N/_K$.
5. Using equation 4, compute the value of x^2.
6. Obtain the degree of freedom (df), with K-1.
7. Prepare a table showing observed and expected frequencies.

Solution

Hypothesis: Null hypothesis (Ho) and Alternative hypothesis (Ha).

Step 1:

Ho.: Cheating in examination is independent of nature of programme students run in a university.

Ha: Cheating in examination is dependent on nature of programme students run in a university.

Step 2:

Table 54: Observed and Expected Frequencies of Cheats

	CEP	**SANDWICH**	**DIPLOMA**	**REGULAR**	**TOTAL (n)**
Cheats	25 (23)	22 (23)	21 (23)	24 (23)	92

Step 3: Expected frequency Ef = $^{N}/K$

Where;

N = total sum of the observed frequencies

K = No. of categories (cells)

Ef = $\frac{92}{4}$ = 23

See Ef in bracket in each cell in table (54).

Step 4: $x^2 = \frac{\sum(O_f - E_f)^2}{E_f}$

O_f	E_f	$O_f - E_f$	$(O_f - E_f)^2$	$(O_f - E_f)^2/E_f$
25	23	2	4	0.1739
22	23	-1	1	0.0434
21	23	-2	4	0.1739
24	23	1	1	0.0434
			∑	**0.4346**

x^2 = **0.4346**

df = K - 1 = 4 – 1

= 3

$X^2_{0.05}3$ = 7.815

Decision Rules

1. Whenever calculated chi-square (x2) is greater than tabulated chi-square reject Ho and conclude in favour of H1.

2. Whenever calculated chi-square (x2) is less than tabulated chi-square accept Ho and conclude in favour of Ho.

3. Since x2 = 0.4346 < $x^2_{0.05}3$ = 7.815, we accept Ho and conclude that cheating in examination is independent of nature of programme students run.

   (**NB: <** means less than).

Chi-Square Test in Two or More Sample Cases

Two or more sample cases prevail when the number of possible outcomes of a given event consists of K cells in two or more rows. In such cases, two or more different groups are involved in the test. The chi-square table may then be called two or three, etc, row chi-square table.

To determine expected frequency (Ef) in each cell of the contingency table, we take the following steps

Let rows total be ni.., and columns total be n.j., Ef should be given as:

$$Ef = \frac{ni \times n.j}{N} \quad \ldots \quad (5)$$

Where n is the grand total

Example 2:

In a College of Education Lecture Hall, seats are arranged under different colours. Male and female students make a choice of color of seats during lectures. This choice is shown below in table 55. Use the data to test the null hypothesis that male and female students' choice of seat is independent of the colour of the seat.

Table 55: Students' Choice of Seats

	Colour of Seats				
Gender	**Green**	**Blue**	**Red**	**Yellow**	**n_1**
Male	8	7	2	3	20
Female	40	60	2	2	104
n.j.	48	67	4	5	124

Statement of Hypothesis

H_o: Male and female students' choice of seat is independent of color of the seat.

H_1: Male and female students' choice of seat is independent on colour of the seat.

Table 56: Observed and Expected Frequencies of Students' Choice of Seats

	Colour of Seats				
Gender	**Green**	**Blue**	**Red**	**Yellow**	**n_1**
Male Female	1). 8(7.74) 2). 40(40.26)	2). 7(10.81) 6). 60.(56.19)	3). 2(0.65) 7). 2(3.35)	4). 3(0.81) 8). 2(4.19)	20 104
n.j.	48	67	4	5	124

Expected frequencies, $Ef = \frac{ni \times n.j}{N}$

Where;

ni = row total i.e. 20 and 104

n.j. = column totals ie. 48, 67, 4 & 5

n = grand total i.e. 124

$Ef._1 = \frac{20 \times 48}{124}$

$= 7.74$

$Ef._2 = \frac{20. \times 67}{124}$

$= 10.81$

$Ef._3 = \frac{20 \times 4}{124}$

$= 0.65$

$Ef._4 = \frac{20 \times 5}{124}$

$= 0.81$

$Ef._5 = \frac{104 \times 48}{124}$

$= 40.26$

$Ef._6 = \frac{104 \times 67}{124}$

$= 56.19$

$Ef._7 = \frac{104 \times 4}{124}$

$= 3.35$

$Ef._8 = \frac{104 \times 5}{124}$

$= 4.19$

$x^2 = \frac{\sum(O_f - E_f)^2}{E_f}$

O_f	***E_f***	***$O_f - E_f$***	***$(O_f - E_f)^2$***	***$(O_f - E_f)^2/E_f$***
8	7.74	0.26	0.0676	0.0087
7	10.81	-3.81	14.5161	1.3428
2	0.65	1.35	1.8225	2.8038
3	0.81	2.19	4.7961	5.9211
40	40.26	-0.26	0.0676	0.0016
60	56.19	3.81	14.5161	0.2583
2	3.35	-1.35	1.8225	0.5440
2	4.19	-2.19	4.7961	1.1446
			∑ =	**12.0249**

$x^2 = 12.0249$

df = (r – 1) (c-1)

where:

r = number of rows

c = number of columns

= (2 – 1) (4 – 1)

= (1)(3)

= 3

$X^2_{0.05}3$

= 7.815

Let x^2_c be used as chi-square calculated. Applying the decision rules stated above:

Since x^2_c = 12.0249 > $X^2_{0.05}3$ = 7.815 we reject Ho, and conclude that male and female students' choice of seats is dependent on colour of seats.

Problems of Small Expected Frequency

Example two above highlighted the possibility of encountering very small expected frequencies in the process of estimating x2 – statistics. Presence of many small expected frequencies in the contingency table may cause the calculated x2 – statistic to be biased. That is the theoretical (calculated) x2 – statistic would be over stated in value (what is obtained from calculation is more than what it should actually be under normal case). The effect of the exaggeration in the value of x^2_c is the rejection of Ho when Ho should have been accepted, thus leading to wrong decision.

To remedy such problem in a chi-square goodness of-fit test, Cochran (1954) suggested that no expected frequency should be less than 1, and that those less than 5 must not be up to 20% of the whole expected frequencies in the contingency table. He recommended that adjacent categories should be pooled together along column or row in the event of any of the two conditions prevailing among the expected frequencies.

Example 3:

Refer to the Table 56 above, let us pool Red and Yellow columns together by joining (adding) values in their cells together to obtain Table 57 below.

Table 57: Observed and Expected Frequencies of Students' Choice of Seats

Gender	Green	Blue	Red	Yellow	n_1
Male	8(7.74)	7(10.81)	2(0.65)	3(0.81)	20
Female	40(40.26)	60.(56.19)	2(3.35)	2(4.19)	104
n.j.	48	67	4	5	124

$$x^2_c = \frac{\sum(O_f - E_f)^2}{E_f}$$

O_f	E_f	$O_f – E_f$	$(O_f – E_f)^2$	$(O_f – E_f)^2/E_f$
8	7.74	0.26	0.0676	0.0087
7	10.81	-3.81	14.5161	1.3428
5	1.46	3.54	12.5316	8.5832
40	40.26	-0.26	0.0676	0.0016
60	56.19	3.81	14.5161	0.2583
4	7.54	-3.54	12.5316	1.6620
			$\sum$ =	**11.8566**

$x^2_c = 11.8566$

$$\begin{aligned} df &= (r – 1) (c – 1) \\ &= (2 – 1) (3 – 1) \\ &= (1) (2) \\ &= 2 \end{aligned}$$

$X^2_{0.05}2 = 5.991$

Since $x^2_c = 11.8566 > X^2_{0.05}2 = 5.991$, we reject Ho, and conclude that male and female students' choice of seats in dependent on colour of seats.

The above result reinforces our findings in example 2. Though X^2_c is overstated, the extent to which it is inflated is insignificant.

Example 4:

A random sample of male and female sandwich students of a certain session were asked to respond to the statement below, by choosing from the options provided, the most appropriate response.

Children restricted to breast feeding in the first twelve months of their lives are very strong.

a). Strongly agree (b). Agree (c). Undecided

d). Disagree (e). Strongly disagree

The data obtained from their responses were as presented in the table below.

Table 58: Responses of Male and Female Sandwich Students to Breastfeeding of Children

Number of respondents		
Responses	Male	Female
Strongly Agree (SA)	150	148
Agree (A)	125	132
Undecided (UD)	55	40
Disagree (DA)	34	56
Strongly Disagree (SDA)	36	24

Do their responses differ at 5% significant level?

Solution

1. **Hypothesis:**

H_o.: There is no significant difference between the male and female responses.

H_1: There is significant difference between the male and female responses

Table 59: Observed and Expected Frequencies of Sandwich Students' Responses on Breastfeeding

Number of respondents			
Responses	**Male**	**Female**	**Ni**
SA	150(149)	148(149)	298
A	125(128.5)	132(128.5)	257
UD	55(47.5)	40(47.5)	95
DA	34(45)	56(45)	90
SDA	36(30)	24(30)	60
n.j.	400	400	800

3. Expected frequencies:

$$Ef = \frac{ni. \times n.j}{n}$$

$$E_f.1 = \frac{298 \times 400}{800}$$

$$= 149$$

$$E_f.2 = \frac{298 \times 400}{800}$$

$$= 149$$

$$E_f.3 = \frac{257 \times 400}{800}$$

$$= 128.5$$

$$E_f.4 = \frac{257 \times 400}{800}$$

$$= 128.5$$

$$E_f.5 = \frac{95 \times 400}{800}$$

$$= 47.5$$

$$E_f.6 = \frac{95 \times 400}{800}$$

$$= 47.5$$

$$E_f.7 = \frac{90 \times 400}{800}$$

$$= 45$$

$$E_f.8 = \frac{90 \times 400}{800}$$

$$= 45$$

$$E_f.9 = \frac{60 \times 400}{800}$$

$$= 30$$

$$E_f.10 = \frac{60 \times 400}{800}$$

= 30

x^2_c = $\frac{\sum(O_f - E_f)^2}{E_f}$

O_f	E_f	$O_f - E_f$	$(O_f - E_f)^2$	$(O_f - E_f)^2/E_f$
150	149	1	1.00	0.0067
148	149	-1	1.00	0.0067
125	128.5	-3.5	12.25	0.0953
132	128.5	3.5	12.25	0.0953
55	47.5	7.5	56.25	1.1842
40	47.5	-7.5	56.25	1.1842
34	45.0	-11	121.00	2.6888
56	45.0	11	121.00	2.6888
36	30	6	36.00	1.2000
24	30	-6	36.00	1.2000
			∑ =	**10.3500**

x^2_c = 10.35
df = (r – 1)(c – 1)
= (5 – 1) (2 – 1)
= (4) (1)
= 4
$X^2_{0.05}4$ = 9.488

5. Since $x^2_{cal} = 10.35 > x^2_{crit.\ 0.05}4 = 9.488$, we reject Ho, and conclude that there is significant difference between the male and female students' responses on the assertion or claim.

Exercises:

1. A random sample of members of four university based trade unions in attendance to a nationwide rally were required to indicated political system suitable to Nigeria. The selection they made were shown in the table below.

Table 60: Responses of Suitable Political System for Nigeria

Political Systems	Trade Unions			
	ASUU	**SSANU**	**NASUU**	**ASUTON**
Military	60	30	30	40
Democracy	100	30	40	70
Autocracy	80	30	50	40

Did they select the same political system at 5% level of significance?

2. Three oranges were randomly selected from three species of orange tree. The number of seeds in each specie is as presented in the tables below:

Table 61: Random Selection of Orange Seeds

Specie	**Hybrid**	**Local**	**Mixbrid**
Number of Seeds	4	11	6

Do the species have equal number of seeds at 5% level of significant?

3. The following is the order in which savings deposits (D) and withdrawals (W) from savings were received.

   D D D W D W W D D W W D W W D

   Test for randomness at 5% level of significance

4. The number of students arriving at students' but stop at Temporary site of NAU by the afternoon period were as given below:

   3 1 5 6 4 5 6 1 5 8 3 12 1 2 9 7

   Would you say that arrangement is random at 5% level significance?

References

Aduojo S.O. (2003) *Basic statistics in behavioural research*. Kogi: FIsher Pub. Ltd

Agu, N. *Basic statistics for behavioural sciences*. Awka: Madonna Pub Ltd

Ayuba M.M (2010). Elementary *statistics*. Maiduguri: University Printing Press

Ajanyi E.A (2010). *Educational measurement and evaluation*. Lagos: University Press Ltd

Ajoni P.A (2011). *Business statistics*. Ife: Odua Ajojo Pub ltd

Bright N.k (2011). Uses of statistics: *American Psychologist*, 38(5), 1045-1050

Bakwo (2015). *Introduction to applied statistics in science*. Kaduna: Kwabo African Press Ltd

Berk H. (2000). *Statistical reasoning*. (2nd Ed). New York: MacMillan Pub. Co.

Bester P.O (2009). *Causes of difference between aptitude and academic achievement*. (3rd Ed). New York: John Wiley & Sons Inc.

Coetzee K. (2001). Theories and problems of probability in modern life. *West African Journal of Education*, 18(2) 300-350.

Egwa O.A (2011, June, 8). Strategies for improving universities student academic achievement in statistics. *Sunnews*, P.3

Eno & Seldon (nd). *General applied statistics*. New York: McGraw-Hill

Egbo O.A (2015). Rules of statistical testing. Enugu: Celex Pub. Ltd

Encarta Encyclopedia. (2015). *Statistics*. Encarta: Oxford University Press.

Eze N.B (2009). *Basic business statistics*. Awka: Famous Pub. Ltd.

Ejeh et al (2009). Multiple admission in Nigerian universities. In P. Edo(ed) *Handbook of Statistics.* vol 12, Oxford, UK: Trikles Ltd.

Founche and Verwey (2013). Consequencies of sampling and its implications. *American Psychologist*, 32(2), 200-215

Ghiselli S. (2000). Inferential statisitcs. *Measurement in Education*, 2, 3-9

Joela S.K (2004). Numerical tabulation of Kolmongorov's statisitics for sample size. *Journal of Educational measurement.* 12, 121- 127

Kama T.O. (2010). *Introduction to statistical analysis*. Zaria: Sambo and Kofas

Kaka M.A (2010). *Applied regression analysis*. Anyigba: Agaba Pub Ltd

Kelly H. (2014). *Statisitcal analysis for rates and proportion*. Enugu: Holand Ltd

Masel N.P. (2010). *Non parametric statistical method*. Enugu: Celex Ltd

Maris, M. (2012). *Advanced theory of statistics*. Handbook I: *Cognitive domain.* New York: Greg Int.

Mussen, Conger & Huston (2004). *Statistics for management and economics*. Ghana:University of Ghana Press Ltds.

Ngozi, A. (2009). *Basic Statistics for Education and Behavioral Science*. Awka. Valouxy Press Ltd.

Oxford, A. (2013). Oxford Advanced Learner's Dictionary. (2nd Ed). London: Oxford University Press

Okonkwo, N. (2013). *Another view of statistics*. *Abia*: Ukah Pub. Co

Okoye, R.O (2015). *Educational and psychological statistics*. (2nd Ed). Awka: Erudition Pub.

Pearson, K. (nd). *Applied linear statistical models*. New York: Winston and Brass.

Port P.P & Digresia G. (2010). *Non parametric statistics*. Georgia: University of American Pub.Co.

Prediager, D., Waple, H & Nusbaum (2008). *Analysis of variance test for normality*. New York: John Willy Press

Reber O.O (2010). *Simultaneous statistical inference*: Alpha Book Press

Stumph, T., Stanley, D. (2002). Analysis of variance: An overview. *Theory and practice*, 30(3), 112-130

Steyn, R.B. (2008). Academic aptitude and its effects on the learner's academic performance. *A collection of papers*. New York: Winston Press.

Tko, O., & Tolu, O. (2012). *Measurement and evaluation in education.* Lagos. University Press.

Taylor, O.A. (2014). *Predicting academic achievement of students, using scholastic aptitude test.* A paper presented at the International Conference on measurement in education, Arizona II

Ugboduma, U. (2011). *Effectiveness of aptitude test in predicting academic achievement in Health Sciences*. Oxford: Peg-mound Pub.Co

Vosloo, R.O., Coetzee & Classen (2000). Test validity and ethics of assessment. *American Psychologist:* 30,1000-1124