INTRODUCTION

The main purpose of this module is to introduce you to statistical method in research. In order to effectively handle this task, the module has been divided into eleven units. Unit one is an introduction to statistics. Unit two looks at the measures of central tendency while unit three focuses on measures of dispersion. Unit four gives an introduction to probability. Unit five discusses frequency distributions. Unit six is an introduction to estimates with samples. Unit seven focuses on hypotheses testing. Unit eight looks at analysis of variance (ANOVA). Unit 10 focuses on regression. Unit eleven presents the format for the PAS 2014 research report and gives guidance on how to write the research report.

AIM OF THE MODULE

The aim of this module is to introduce you to statistical methods in research.

OBJECTIVES OF THE MODULE

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

- Explain the major statistics concepts used in research.
- Demonstrate an understanding of descriptive statistics.
- Acquire knowledge of inferential statistics.
- Analyze data and write a research report

TIMES REQUIRED

It will take you approximately three months to undertake the requirements of this module. The time should be spent on studying the module, reading recommended material, doing self-help questions and writing the research report.

STUDY SKILLS

 As a distance student, your study skills will be different from students that are resident at the university. In your case, you will choose what to study, when to study and where to study from on your own. You will also need to fit your study activities around other professional and/or domestic responsilities. It is therefore important for you to allocate enough time to your studies and secure a conducive place to study from.

NEED HELP

 In cases where you need help, you should contact the Director, Institute of Distance Education, University of Zambia, P.O Box 32379, Lusaka. Cell 0978 772259 or 0979 772248, Email: direstor-ide@unza.zm

ASSESSMENTS

 The assessments under this module are in two categories. These are Continuous Assessment (CA) and the Final Examination as shown below.

ASSESSMENT

Continuous Assessment:	50%
Broken down as follows:	
a) Test 1	10%
b) Test 2	10%
c) Proposal	15%
d) Research Report	15%
Final Examination:	50

UNIT ONE: INTRODUCTION TO STATISTICS

1.0 Introduction

This unit gives an introduction to statistics. The unit begins by defining statistic and thereafter outlines the importance of statistics. We will then examine different types of frequency distributions such as ungrouped frequency distributions, grouped frequency distribution, absolute frequency distributions, relative frequency distributions, cumulative frequency distributions and decumulative frequency distributions. The last part of this unit will focus graphical presentation of data.

2.0 Aim of the Unit

The aim of this unit is to introduce you to statistics.

3.0 Objectives of the Unit

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

- Explain the meaning of statistics.
- Outline the importance of statistics.
- Distinguish between descriptive and inferential statistics.
- Identify different types of frequency distributions.
- Construct graphs used for data presentations.

4.0 Required Material

In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

5.0 Time Required

 It will take you approximately two hours thirty minutes to work through this unit.

6.0 Main Contents

6.1 Defining Statistics

Statistics is the plural of the word statistic. A statistic is a figure or a number. Statistics therefore refers to numbers, figures or numeric information. For example, the number of students studying PAS 2014 at the University of Zambia, the proportion of people who are unemployed in Zambia, the population of Zambia, the number of voters in a constituency, the number of people without access to clean drinking water and others. As a field of study, statistics is the study of the collection, analyzing and interpretation of data (numbers/information). In research, we are concerned with the study of methods of statistical analysis which involves descriptive and inferential statistics. Descriptive statistics is concerned with organizing, summarizing and describing of data. On the other hand, inferential statistics involves making generalizations for the entire population based on results obtained from a sample.

6.2 Importance of Statistics

Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study. Statistics help one to understand and evaluate quantitative information with minimum difficulties. For example, the following statements are statistical in nature and require knowledge of statistics to fully comprehend:

- 63% of people in Africa live in rural areas;
- The poverty level in Zambia is estimated at 68%;
- Almost 80% of youths in Zambia are unemployed;
- The average amount of time PAS 2014 students spend on studying per day is 30 minutes;
- The HIV/AIDS prevalence rate in Zambia is 14.3%;

- The voter turnout in the last Presidential and Parliamentary elections in Zambia was about 61%;
- 51% of the national budget in Zambia is spent on salaries for civil servants;
- Approximately 10% of Zambian Members of Parliament in Zambia are women and;
- The literacy level in rural areas is almost 52%.

Knowledge of statistics is important in solving practical problem as it enables one to know the severity of a problem. For instance, statistics on unemployment, prevalence of a disease, crime rate and others demonstrate the need for an intervention or practical solutions in solving the problem. This implies that statistics play a crucial role in policy formulation. Statistics is also important in development planning and decision making because statistics reinforce or lend credibility to an argument. Statistics is also important in business, particularly when it comes to market research. Statistics is also important when undertaking opinion polls or opinion research.

6.3 Descriptive Statistics

Descriptive statistics is a way of organizing, describing and summarizing data in way that is easy to understand. The organization of data takes several forms such as graphical techniques and numeric techniques. The simplest way of organizing data consists of listing observations in form of frequencies. In terms of the graphical techniques, data is presented graphically by way of, for instance, pie – charts, bar charts, histograms, scatter plots and other. In numerical presentation of data, we are interested in identifying some general characteristics by way of measures of central tendency such as the mean, mode and median. We are also interested in presenting data in terms of measures of dispersion such as the range, variance and standard deviation.

Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data is showing, especially if there is a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. For example, if we had test results for PAS 2014, we may be interested in the overall performance of these students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this.

6.4 Frequency Distributions and Tables

Frequency distributions help us to organize data by showing us the frequency (regularity) with which a particular variable or observation occurs. For example, suppose a sample of 20 civil servants were asked to indicate their level of job satisfaction on a 5-point scale ranging from (i) Very low; (ii) Low; (iii) Medium; High; and Very High.

Table 1.1: Level of satisfaction as indicated by 20 civil servants

Interviewee No.	Level of job satisfaction
1	Low
2	Very high
3	Medium
4	Medium
5	High
6	Medium
7	Low
8	Very low
9	Medium
10	Low
11	Medium
12	High
13	Low
14	High
15	Very low
16	High
17	Medium
18	Medium
19	Low
20	Medium

The above responses obtained do not give us a clear view as to whether civil servants are satisfied with their job or not. Therefore, the answers should be regrouped according to their content whereby the answer to each category is given and this leads to a more adequate description of the situation.

Table 1.2: Frequency distribution of the level of job satisfaction by 20 civil servants

Level of job satisfaction	Frequency	Interviewee No.
very high	1	2
High	4	5;12;14;16
Medium	8	3;4;6;9;11;17;18;20
Low	5	1;7;10;13;19
very low	2	8;15
	n = 20	

Table 1.2 above shows that civil servants who are neither satisfied nor dissatisfied with their jobs constitute the largest group (8) whereas there are more dissatisfied civil servants altogether (7) than those satisfied with their jobs. The total number of civil servants interviewed is equal to the sum of all the frequencies and is indicated by n = 20, n represents the sample size, that is, the total number of civil servants interviewed in this case.

1. Ungrouped Frequency Distribution

This is a type of frequency distribution which shows the number of times each observation appears separately on its own. For example, suppose you collected data on the age of distance students doing PAS 2014 at the University of Zambia. You sampled 25 (n = 25) students who gave the following responses about their age.

Table 1.3: Responses to the question on age by 25 distance students

48 36 44 32 35 43 27 40 31 62 50 36

51 41 55 56 46 47 41 28 33 36 22 53 49

The raw data on the age of 25 distance students presented on page 7 would be more confusing if hundreds of students were interviewed. The list would not make sense to most people. To make table 1.3 more meaningful, one has to arrange the years in their ascending order, starting from the lowest, which is 22 years to the highest which is 62 years. One can then find out the number of times each score (age) appears, the frequency of the score is represented by the letter (f).

Table 1.4: Ungrouped Frequency Distribution Constructed from the raw data in Table 1.3

Age	f	Age	f	Age	f
22	1	43	1	58	1
26	1	44	1	62	1
27	1	46	1		
31	1	47	1		
32	1	48	1		
33	1	49	1		
35	1	50	1		
36	3	51	1		
40	1	53	1		
41	2	55	1		

Although the ungrouped frequency distribution in table 1.4 gives a better description of the data than mere presentation of the raw data as in table 1.3, it is still a long and cumbersome list. Therefore, there is need to make it more meaningful and manageable by grouping raw data before counting their frequencies.

2. Grouped Frequency Distributions

The purpose of grouping data is to reduce the number of figures or scores presented in a distribution so as to enable the reader grasp the main features of the data and to present the information more comprehensively. However, the grouping must be done in accordance with some rules, without the distortion or loss of too many items of information contained in the data set.

Table 1.5: Grouped Frequency Distribution Constructed from the Ungrouped Frequency Distribution in table 1.4.

Age (Class – interval)	Frequency
20 – 24	1
25 – 29	2
30 – 34	3
35 – 39	4
40 – 44	5
45 – 49	4
50 – 54	3
55 – 59	2
60 – 64	1
	n = 25

This description of the age of 25 distance students in PAS 2014 shows clearly the pattern of the distribution. Most of the students are concentrated between 30 and 54 years whereas few students are aged below 30 and above 54 years.

3. Absolute Frequency Distributions

Absolutes frequency distributions are frequency distributions where there are only absolute values. Absolute frequency distributions can be applied to both grouped and ungrouped data. For example, the frequency table below is an example of an absolute frequency distribution because it does not show percentages which were observed at each score value.

Table 1.6: Absolute Frequency Distribution

Level of job satisfaction	Frequency
very high	1
High	4
Medium	8
Low	5
very low	2
	n = 20

4. Absolute Frequency Distributions

Relative frequency distributions are frequency distributions representing the percentage of cases which were observed at each score value or class – interval. Thus relative frequency distributions can be applied to both grouped and ungrouped data. The procedure is simply to express each frequency as its corresponding percentage. For example, table 1.6 when expressed as a relative frequency distribution will be like table 1.7.

Table 1.7: Relative Frequency Distribution

Level of job satisfaction	F	F %
very high	1	5
High	4	20
Medium	8	40
Low	5	25
very low	2	10
	n = 20	100%

5. Cumulative Frequency Distribution

This is a kind of frequency distribution used when you want to know the percentage of people/elements falling below a certain point. Cumulative frequency distribution is abbreviated

as (CF). For example, the cumulative frequency distribution from the data given in table 1.5 would be shown as follows:

Table 1.6 Cumulative Frequency Distribution

Age (Class – interval)	F	CF	CF %
20 – 24	1	1	4
25 – 29	2	3[1+2]	12
30 – 34	3	6[3+3]	24
35 – 39	4	10[6+4]	40
40 – 44	5	15[10+5]	60
45 – 49	4	19[15+4]	76
50 – 54	3	22[19+3]	88
55 – 59	2	24[22+2]	96
60 – 64	1	25[24+1]	100
	n = 25		

In interpreting the cumulative frequency, you are supposed to use the upper limit of the class interval. The upper limit is the number which shows where the class interval ends. For example, in a class interval of 45 – 49 years, the upper limit is 49 years. On the other hand, the lower limit is the number which shows where the class interval begins. In the class interval of 45 - 49 years, 45 is the lower limit. Therefore, the cumulative frequency above shows that 76% of distance students are below 49 years. In absolute terms, 19 distance students are below 49 years.

6. Decumulative Frequency Distribution

This is a kind of frequency distribution which is used when you want to know the percentage of observations located above a certain point. Decumulative Frequency Distribution is abbreviated as (DCF). For example, the decumulative frequency distribution for the data given in table 1.5 would look as follows:

Table 1.7: Decumulative Frequency Distribution

Age (Class – interval)	F	DCF	DCF %
20 – 24	1	25[24+1]	100
25 – 29	2	24[22+2]	96
30 – 34	3	22[19+3]	88
35 – 39	4	19[15+4]	76
40 – 44	5	15[10+5]	60
45 – 49	4	10[6+4]	40
50 – 54	3	6[3+3]	24
55 – 59	2	3[1+2]	12
60 – 64	1	1	4
	n = 25		

In interpreting the decumulative frequency distribution, you use the lower limit of a class interval. Therefore, the cumulative frequency distribution shows that 96% of distance students are above 25 years. In absolute terms, 24 distance students are above 25 years.

Note that to find the mid-point of a class interval, you add the lower and the upper limit and divide by 2. For example the mid-point the class interval 20 – 24 years is arrived at as follows: (20 +24) / 2 = 22

6.5 Graphical Presentation of Data

Once raw data has been organized into a frequency distribution, it can be visually presented by various types of graphs, bars charts, pie charts and other forms of pictorial presentations. Graphical representations have a great advantage of allowing one to grasp immediately the main characteristic of the information. Although a frequency distribution describes a set of data, it still remains a list of figures which has to be studied carefully to get information, for example the variations of the data. However, there are ways of depicting a frequency distribution in order to allow an immediate grasp of its characteristics, this is based on visual distributions where

numerical relationships are expressed in spatial relationships. The following are some of the ways in which data can be presented:

1. Pie Charts

Pie charts are diagrams that use a circle which is subdivided into sections. Each section is proportionate to the size of the figure represented. It is a very simple method of expressing various components and their relationship to the whole. Each figure must be transformed into an angle in degrees, the total corresponding to 360º.

For example, suppose are the analyzing the number of street kids in 4 major towns in Zambia and you have the following;

City	No. of Street Kids
Lusaka	450
Kitwe	300
Ndola	150
Livingstone	100
Total	1000

Since 1000 represents 360º,

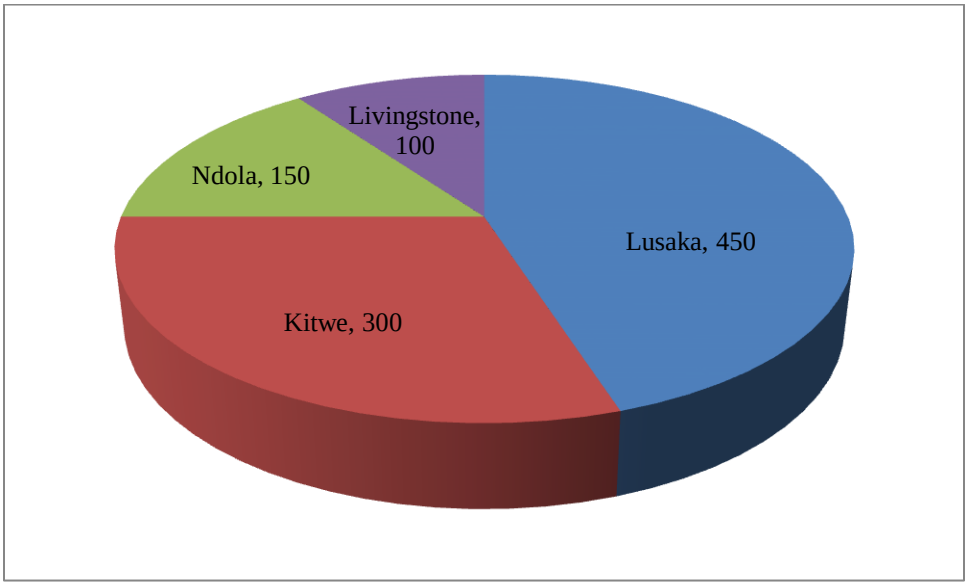
The number of street kids in Lusaka is equivalent to 450/1000 x 360 = 162º

The number of street kids in Kitwe is equivalent to 300/1000 x 360º = 108º

The number of street kids in Ndola is equivalent to 150/1000 x 360 = 54º

The number of street kinds in Livingstone is equivalent to 100/1000 x 360 = 36º

Figure 1.1: Number of Street Kids in Four Selected Towns in Zambia



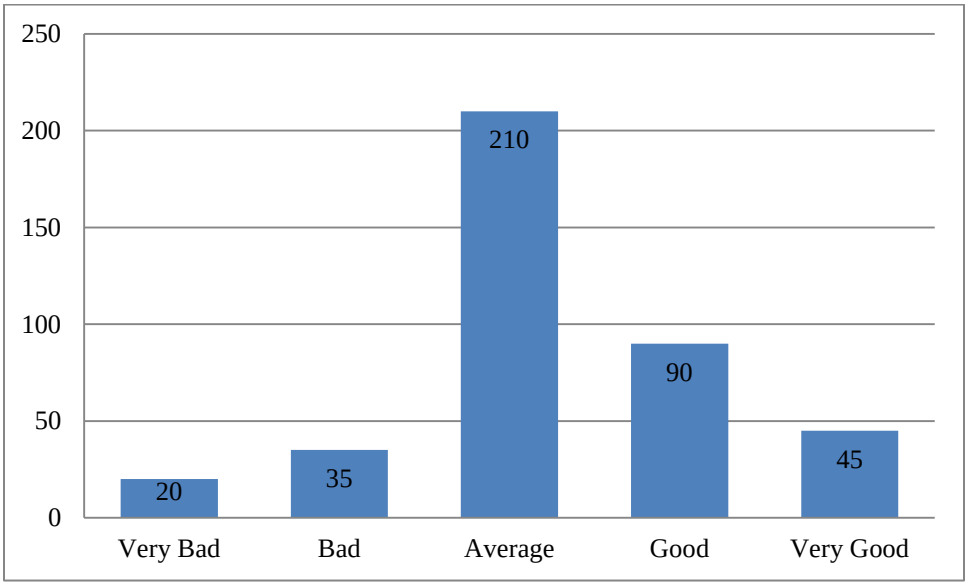
2. Bar charts

Simple bar charts represent data by a series of bars where the height of each bar indicates the size of the figure or the value represented. The width is the same for all the bars. To construct a bar chart, label the frequencies vertically, locate the categories of a variable on a horizontal scale and label the frequencies on the vertical axis by using whatever scale you consider appropriate.

For example, suppose 400 students at the University of Zambia were asked to rate the quality of internet connectivity at the institution and gave the following responses.

Response	Number of students
Very Bad	20
Bad	35
Average	210
Good	90
Very Good	45

Figure 1.2: Students perception of the quality of internet connectivity at UNZA

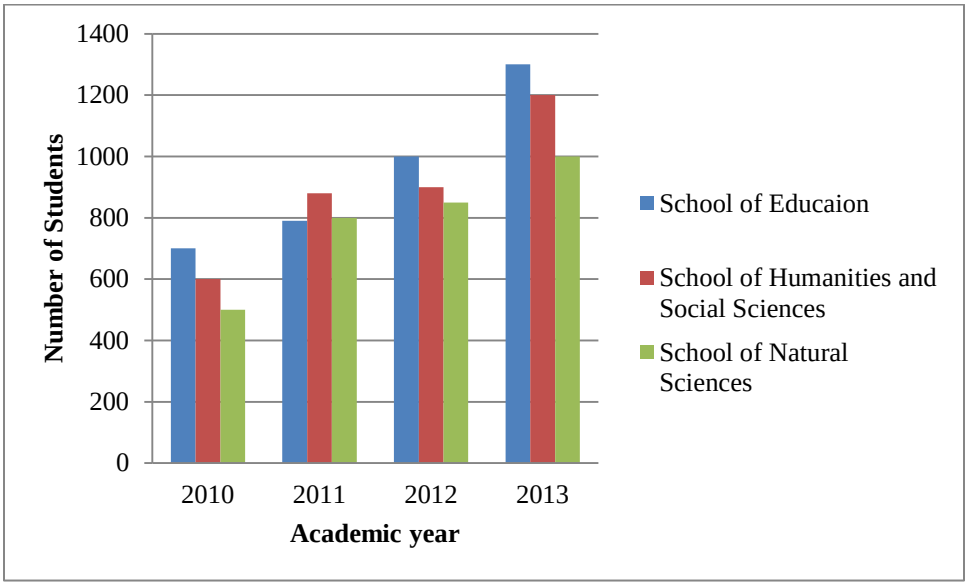


3. Multiple Bar Charts

A multiple bar chart allows the representation of many variables simultaneously. This is used when more than one relationship has to be depicted. For example, one might be interested in studying the students' population as it is spread over the various schools over a given period of time. This offers the possibility for comparing the students' population in different schools over a given period of time. For example, hypothetical data on the enrollment rate in three schools at the University of Zambia from 2010 to 2013 is presented in the table below.

	Academic Year			
School	**2010**	**2011**	**2012**	**2013**
Education	700	790	1000	1300
Humanities	600	880	900	1200
Natural Sciences	500	800	850	1000

Figure 1.3: Students Enrollment in three Schools at UNZA from 2010 - 2013

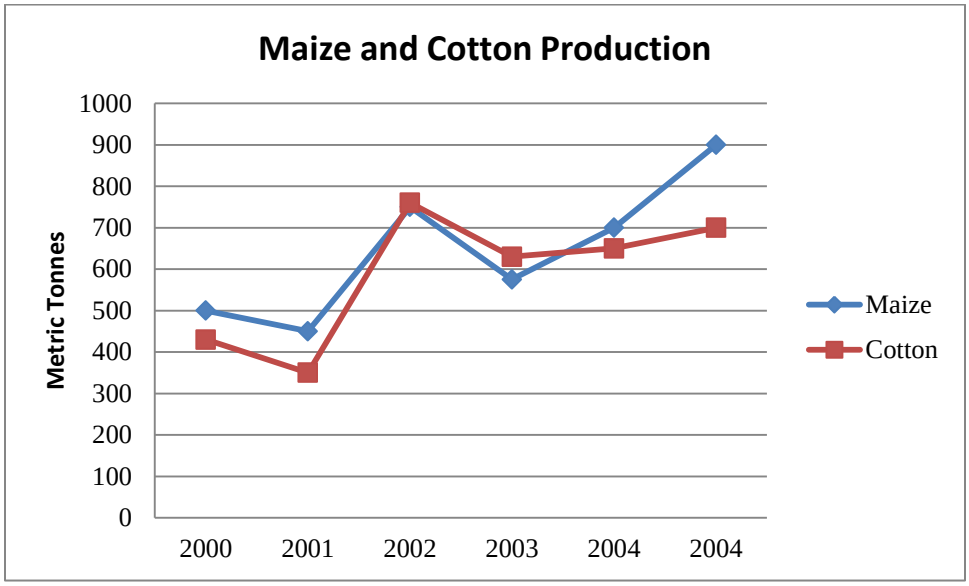


4. Graphs

A graph is a representation of data by a continuous line within the framework of two axes, the horizontal axis and the vertical axis. Social scientists tend to use a specific kind of graph where (y) denotes the frequency of the variable or variables and (x) denotes the periods under consideration. If the graph is about one variables or item, then there will be one continuous line in the graph. However, if two variables or items are considered, then there will be two continuous lines. For example, data on Maize and Cotton production in Gwembe district from 2000 to 2004 is shown below.

	Farming Season					
Crop	2000	2001	2002	2003	2004	2005
Maize	500	450	750	575	700	900
Cotton	430	350	760	630	650	700

Figure 1.4: Maize and Cotton Production in Gwembe, 2000 – 2004



7.0 Summary

You have been informed in that statistics refers to numbers, figures or numeric information. As a field of study, statistics is concerned with the analysis of data and this involves descriptive statistics and inferential statistics. Descriptive statistics is concerned with the organising, summarizing and describing of data while inferential focuses on making generalizations on results obtained from a sample for the whole population. Various techniques are used under descriptive statistics such as frequency distributions and graphs.

8.0 Self Assessment Questions

After reading through unit one, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit one again.

Given data below:

Class – interval	Frequency
12 – 18	2
19 – 25	1
26 – 32	5
33 – 39	1
40 – 46	4
47 – 53	4
54 – 60	2
61 – 67	2
68 – 74	3
75 – 81	2
82 – 88	2
89 – 95	2

Complete the frequency table by including the following;

a. The midpoint of each class interval

b. The relative frequency of each class interval

c. The cumulative frequency for the above data

d. The de-cumulative frequency for the above data

Distinguish between the following;

a. Ungrouped frequency distribution and grouped frequency distribution.

b. Cumulative frequency distribution and decumulative frequency distribution.

c. Pie chart and bar chart.

d. Inferential statistics and descriptive statistics.

UNIT TWO: MEASURES OF CENTRAL TENDENCY

1.0 Introduction

This unit discusses the measures of central tendency. The unit begins with explaining the mode and thereafter discusses the median. We will then focus on the mean for both ungrouped and grouped data.

2.0 Aim of the Unit

The aim of this unit is to introduce you to the measures of central tendency.

3.0 Objectives of the Unit

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

- Interpret the measures of central tendency
- Compute the measures of central tendency

4.0 Required Material

In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

5.0 Time Required

It will take you approximately forty minutes to work through this unit.

6.0 Main Contents

Measures of central tendency are ways in which a set of figures can be summarized by using only one number which expresses the central feature of the whole set. These measures are used in situations where graphical methods may be inappropriate such as when making statistical inference. These methods are also used for reasons of expediency because it is sometimes easier to verbally communicate than to use graphs. The measures of central tendency used in research are the mode, median and the mean.

6.1 The Mode

The mode of a set of measures is defined as the measurement or observation that occurs most frequently in that distribution or has the highest frequency in that distribution. For example, n=25 (n stands for sample size or number of observations), number of hours PAS 2014 students spend studying per week.

7, 10,8, 11,9, 9, 9, 8, 9, 8, 9, 9, 9, 8, 9, 8, 8, 9, 10, 11, 10, 7, 10, 9, 8

Arrange the data from the smallest to the highest according to the number of times (frequency) they occur.

7, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 11, 11

The number whose observation occurs most frequently is 9. Therefore, the mode is 9, which means that most PAS 2014 students spend 9 hours per week studying. Sometimes it is possible to have more than one mode, as is the case in bimodal or trimodal distributions.

The mode is an important measure of central tendency insofar as it enables one to easily determine what is typical or common among a group. However, the shortcomings of the mode are that it does not take into account all the observations or values in its computation.

6.2 The Median

The median of a set of measurements or observations is the middle value when these measurements or observations are arranged in order of magnitude. Where there is an even number of observations, the median is simply the average of the two midpoint values. In situations where we have an old number of observations, the midpoint value becomes the median. Suppose you had the following number of observations where n=10;

95, 86, 78, 90, 62, 73, 89, 92, 84, 76

Rearrange the numbers in order of magnitude

62, 73, 76, 78, 84, 86, 89, 90, 92, 95

Median = (84 + 86) /2 = 170/2 = 85

If we include 60 to make n=11, we will have;

60, 62, 73, 76, 78, 84, 86, 89, 90, 92, 95

Median = 84

6.3 The Mean

The mean is often referred to as the average. It is the sum of measurements divided by the total number of measurements. The population mean is denoted by μ while the sample mean is denoted by $\bar{x}$.

a. The Mean for Ungrouped Data

For example in class of 10 students, the marks obtained by the students in a test were as follows;

50, 59, 65, 63, 85, 76, 39, 41, 51, 51

Mean ($\bar{x}$) = $\frac{50 + 59 + 65 + 63 + 85 + 76 + 39 + 41 + 51 + 51}{10}$

= $\frac{580}{10}$

= 58

The interpretation of the mean mark we have computed is that on average, each student is expected to have scored 58 marks in the test.

b. The Mean for Grouped Data

When calculating the mean for grouped data, you need to have the midpoints of the class intervals. The midpoint of a class interval is denoted by xi and is computed by adding the upper and lower limit of the class interval and dividing the answer by 2. For example, for the class interval 20 – 24, the midpoint will be (20 + 24) / 2 = 22. The formula for grouped data mean is:

$$\bar{X} = \frac{\sum f_i x_i}{n}$$

Where $\bar{x}$ = mean

Σ = is a summation symbol

x_i = is the mid - point of a given class interval

f_i = is the frequency for a given class interval

$f_i x_i$ = is the product of the frequency of a class interval and its midpoint

Suppose you have the following grouped data.

Table 2.1: Grouped Data on Age

Age group	X_i	F_i	F_iX_i
20 – 24	22	3	66
25 – 29	27	4	108
30 – 34	32	5	160
35 – 39	37	6	222
40 – 44	42	5	210
45 – 49	47	4	188
50 – 54	52	3	156
		n = 30	$\Sigma f_i x_i$ =1110

Therefore, the mean age for the grouped data above is:

$$\overline{X} = \frac{\Sigma f_i x_i}{n}$$

$$= \frac{1110}{30}$$

= 37 years

7.0 Summary

You have been informed in this unit that measures of central tendency are ways in which a set of figures can be summarized by using only one number which expresses the central feature of the whole set. These methods are also used for reasons of expediency because it is sometimes easier to verbally communicate than to use graphs. The measures of central tendency used in research are the mode, median and the mean.

8.0 Self Assessment Questions

After reading through unit one, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit two again.

The following are the marks scored by students in a test:

40 28 38 78 82 56 74 20 62 46

80 90 52 30 58 82 72 24 48 70

Using the data above **compute** the following and give an **interpretation** of each: (65 Marks)

a. Mean
b. Median
c. Mode

The data below shows the number of 50 kg bags of maize harvested by 1, 000 small scale farmers. The number of bags is grouped in intervals of 5.

Table 2.1: Grouped Data on Bags of Maize Harvested

Class-interval	Frequency
10 – 14	8
15 – 19	30
20 – 24	40
25 – 29	94
30 – 34	126
35 – 39	156
40 – 44	176
45 – 49	138
50 – 54	118
55 – 59	70
60 – 64	20
65 – 69	16
70 – 74	8

a. Compute the mean and interpret the answer.

UNIT THREE: MEASURES OF DISPERSION

1.0 Introduction

This unit discusses measures of dispersion. The unit begins by explaining the ranges as a measure of dispersion. Thereafter, we shall look at the average deviation from the mean as a measure of central tendency. We will then examine the variance as a measure of central tendency and the last part in this unit will discuss standard deviation.

2.0 Aim of the Unit

The aim of this unit is to explain measures of dispersion.

3.0 Objectives of the Unit

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

- Identify the measures of dispersions.
- Compute the measures of standard deviation.

4.0 Required Material

In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

It will take you approximately one hour to work through this unit.

6.0 Main Contents

When dealing with quantitative data, the mode, median and the mean can be computed but variability is equally important. Variability is concerned knowing the extent to which values in the distribution deviate from the central value. Measures of variability are also called measures of dispersion; they constitute another type of statistic which summarizes a distribution. Just like measures of central tendency, there are different types of measures of dispersion.

6.1 The range

The range is the difference between the highest and the lowest score in a distribution. Its computation however, depends on whether the data is grouped or ungrouped. The range for ungrouped data is given by the formula; Range = highest score –lowest score. For example, if a distribution has the following scores, 4, 8, 9, 7, 5, 3. Then range is 9-3 = 6

For grouped data in class intervals, the range is defined as the difference between the real upper limit of the highest class interval and the real lower limit of the lowest class interval. Therefore, Range = real upper limit of the highest class interval – the real lower limit of the lowest class interval.

Table 3.1: Grouped data on Age

Age (Class – interval)	F	CF	DCF
20 – 24	1	1	25
25 – 29	2	3	24
30 – 34	3	6	22
35 – 39	4	10	19
40 – 44	5	15	15
45 – 49	4	19	10
50 – 54	3	22	6
55 – 59	2	24	3
60 – 64	1	25	1
	n = 25		

Referring to the data above, the real upper limit of the highest class interval, 60 – 64 years is 64.5 and the real lower limit of the lowest class interval 20 – 24 years is 19.5 years. Hence the range of this distribution is 64.5 – 19.5 = 45. The frequency of the class interval does not affect the range. Note that the upper and lower limits of the class interval you have in the table above are stated limits. To find a true upper limit, you add 0.5 to the stated upper limit. Similarly, to find the true lower limit of a class interval, you subtract 0.5 to the stated lower limit

6.2 Average deviation from the mean

The range is not very suitable for measuring the extent to which values differ from the mean. We need to have a measure which takes into account the variations from the mean. Deviation from the mean is a more robust measure as it gives the relationship between each value and the dispersion from the mean. For example, if 5 students got the following results in a test; n = 5

Student	Marks (%) / x
A	68
B	67
C	66
D	63
E	61
	Σx = 325

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

$$= \frac{325}{5}$$

$$= 65\%$$

The margins by which these students deviated from the mean can be as shown in the table below.

Student	Marks (x)	$d = (x - \bar{x})$	$\lvert x - \bar{x} \rvert$
A	68	+3	3
B	67	+2	2
C	66	+1	1
D	63	-2	2
E	61	-4	4
			$\Sigma \lvert x - x \rvert = 12$

The third column indicates the difference between each score and the mean. Since some of the values have a negative sign, it would make it impossible for us to find the average because when we add the deviations, the sum will be zero. To overcome this, the fourth column has been included with a heading of $\lvert x - \bar{x} \rvert$ these signs transform all the values obtained in the third column into positive value. Therefore, the general formula for average deviation is

$$AD = \frac{\Sigma \lvert x - \bar{x} \rvert}{n}$$

$$= \frac{12}{5}$$

$$= 2.4$$

The formula for average deviation for grouped data is:

$$AD = \frac{\sum f |x_i - \bar{x}|}{n}$$

Age group	*Xi*	*F*	$x_i - \bar{x}$	$f \mid x_i - \bar{x} \mid$
20 – 24	22	3	-15	45
25 – 29	27	4	-10	40
30 – 34	32	5	-5	25
35 – 39	37	6	0	0
40 – 44	42	5	5	25
45 – 49	47	4	10	40
50 – 54	52	3	15	45
		n = 30		$\Sigma f \mid x_i - \bar{x} \mid = 220$

$$AD = \frac{\sum f |x_i - \bar{x}|}{n}$$

$$= \frac{220}{30}$$

$$= 7.3$$

6.3 Variance

While discussing the average deviation, it became clear that a deviation expresses the distance of a particular score from the mean, symbolically written as $d = x - \bar{x}$. However, the average of all the deviations from the mean cannot be taken as a measure of variability. This is so because as we discussed with average deviations, the mean deviation is sometimes zero. Since $\Sigma (x - \bar{x}) = 0$. This is avoided by using the square of the deviations (instead of the absolute values as done for the average deviations). The average of the squared deviations of a distribution from the mean of a distribution is called the variance and its formula is:

$$\sigma^2 = \frac{\Sigma f (x - \bar{x})^2}{N}$$

This formula is only considered appropriate only for large population distributions. In this case, the Greek letter sigma (σ) is used. However, when dealing with samples, a more precise formula of the variance is used.

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

In this formula, the sum of the squared deviations from the mean is divided by n – 1 instead of n and the S serves as a symbol for variance. The variance S^2 is 0 only if all the scores in the distribution do not differ from the mean. Otherwise, since the variance depends on the distance of each score from the mean, the more the scores differ, the larger the spread of the variance and the more heterogeneous the group of data. The variance is expressed in squared units of measurements. If scores refer to people or years, the variance will be in square people or years which might not make sense.

6.4 Standard Deviation

Standard deviation is an improvement of the variance. In simple terms, the standard deviation of a distribution is the square root of the variance. As in the case of the variance, the Greek letter σ indicates that the formula refers to the population whereas the letter S refers to a sample. There are two ways of calculating standard deviation valid for grouped and ungrouped data. The first one is based on the actual definition of standard deviation as a square root of the variance.

$$\sigma = \sqrt{\frac{\Sigma f\,(x - \bar{x})^2}{N}}$$

The formula above is called the definitional formula for standard deviation and is used for large populations. When using samples, n as the denominator is replaced with n – 1.

$$S = \sqrt{\frac{\Sigma f\,(x - \bar{x})^2}{n-1}}$$

1. Standard Deviation of Ungrouped Data

The standard deviation for ungrouped data on the marks obtained by 5 students in a test presented on page 27 can be computed as follows;

Student	Marks (x)	Deviations $(x - \bar{x})$	Square of Deviations $(x - \bar{x})^2$
A	68	+3	9
B	67	+2	4
C	66	+1	1
D	63	-2	4
E	61	-4	16
			$\Sigma (x - \bar{x})^2 = 34$

The standard deviation formula for ungrouped data is;

$$\mathbf{S} = \sqrt{\frac{\boldsymbol{\Sigma} (\mathbf{x} - \bar{\mathbf{x}})^2}{\mathbf{n-1}}}$$

This formula requires you to first calculate the mean, find the deviation from the mean for each, square the deviations and find the sum of the squared deviations as shown in the table above. Therefore;

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

$$= \sqrt{\frac{34}{5-1}}$$

$= \sqrt{\frac{34}{4}}$

$= \sqrt{8.5}$

$= 2.9$

This means each student is expected to have deviated from the mean by 2.9 marks.

2. Standard Deviation for Grouped Data

Age group	*xi*	*F*	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
20 – 24	22	3	-15	225	675
25 – 29	27	4	-10	100	400
30 – 34	32	5	-5	25	125
35 – 39	37	6	0	0	0
40 – 44	42	5	5	25	125
45 – 49	47	4	10	100	400
50 – 54	52	3	15	22	675
		n = 30			$\Sigma f(x - \bar{x})^2 = 2400$

$S = \sqrt{\frac{\Sigma f (x - x)^2}{n-1}}$

$= \sqrt{\frac{2400}{30-1}}$

$= \sqrt{\frac{2400}{29}}$

$= \sqrt{82.75}$

$= 9.1$

This means that on average each individual is expected to have deviated from the mean by 9.1 marks.

7.0 Summary

You have been informed in this unit that measures of dispersion help us to know the extent to which values in the distribution deviate from the central value. At this point you should know that the most commonly used measures of dispersion are the range, average deviation from the mean, variance and standard deviation. Standard deviation is the most refined measure of dispersion and there are two ways of calculating standard deviation valid for grouped and ungrouped data.

8.0 Self Assessment Questions

The following frequency table of rent cost for a bed space was constructed after collecting data on rentals from boarding houses near the University of Zambia.

Table3.1: Rental Charges for Boarding Houses

Rent (in Kwacha)	Frequency
226 – 250	2
251 – 275	5
276 – 300	9
301 – 325	7
326 – 350	3
351 – 375	2
376 – 400	4
401 – 425	1
426 – 450	3
451 – 475	2

(i) Compute the mean and interpret the answer.

(ii) Compute the Standard deviation and interpret the answer.

UNIT FOUR: INTRODUCTION TO PROBABILITY

1.0 Introduction

This unit gives an introduction to probability. The unit begins by discussing the common uses of probability. We will then examine properties of probability by focusing on additional, conditional and multiplication rules of probability. Thereafter, we will demonstrate the application of these properties of probability through the use of an example, which will be the last part of this unit.

2.0 Aim of the Unit

The aim of this unit is to introduce you to probability.

3.0 Objectives of the Unit

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

- Explain the uses of probability.
- Identify the properties of probability.

4.0 Required Material


In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

5.0 Time Required


It will take you approximately one hour thirty minutes to work through this unit.

6.0 Main Contents

Whereas descriptive statistics are simply procedures for condensing information about a set of measurements or observations, inferential statistics refer to techniques for making generalizations on the basis of partial information. The study of probability introduces inferential statistics. When testing hypotheses or making generalizations, we use probability to determine the likelihood of an event occurring. For example, the level of significances, which is the probability of rejecting a null hypothesis when it is in fact true, can be easily understood if one has knowledge of probability.

4.1 Common uses of Probability

Probability statements are made by everyone when attempting to explain the present or predict the future. For example, UNZA students are demonstrating, there is a probability that they have not been paid their meal allowance. The young doctor is very dedicated, she will go far in life. Probability is used when evaluating the likelihood of an event happening such as students demonstrations at UNZA. Since there have been several instances when students have demonstrated due to late payment of meal allowance, there is a likelihood that when students demonstrate, delayed payment of meal allowances could be the issue. Probability reasoning is most used to predict the future, the young doctor satisfies pre-conditions for success in her job, so her chances for a bright future are good.

In everyday life, probability evaluations are expressed in words such as chance, likelihood, possibility, expectations or prospects. However, the probability assessments in day to day situations are usually inaccurate and often subjective or intuitive. Some common sense laws of probability are recognized such as that of the probability of an event occurring increases as more preconditions are satisfied. Some events have no probability of occurring at all, for example, the probability of a newly born baby committing suicide. On the other hand some events are certain to occur such as the a normal healthy baby learning how to walk

The basics of probability presented here aim at identifying and making more precise probability assessment based on a few probability laws and then applying them to specific fields of statistics.

4.3 Some Properties of Probability

In a questionnaire, a respondent has the probability of ticking one of the answers given to each of the multiple choice questions. Assuming that the respondent answers all the questions, the probability that any respondent ticks the answer (Yes) in a Yes or No question is $\frac{1}{2}$. This can be expressed in the following ways.

If E denotes the event under discussion (E = to tick the answer yes) the probability of this event occurring is symbolized by p (E). Therefore p (E) = $\frac{1}{2}$ = 0.5. If the response category has 5 alternatives (a 5 point scale) For instance, 'Strongly agree; Agree; No opinion; Disagree and strongly disagree' and one wants to calculate the possibility of the answer 'No opinion' being given. One probability out of 5 possibilities and it is p (E) = $\frac{1}{5}$ = 0.2. In the same way, the possibility of any answer other than 'No opinion' is 4 out of 5 possibilities and it is p (E) = $\frac{4}{5}$ = 0.8 and the possibility of a positive answer (Strongly agree or agree) is p (E) = $\frac{2}{5}$ = 0.4. In other words, the possibility of an event occurring is:

1. p (E) = $\frac{Number\ of\ favourable\ outcomes}{Number\ of\ all\ possible\ outcomes}$

If an event is impossible and can never occur, it means that the number of favourable outcomes is zero, the p (E) = 0.

If an event is certain that it will occur, it means the number of favourable outcomes is equal to the number of possible outcomes, then p (E) = 1

2. $0 \leq p\ (E) \leq 1$

The probability of any event occurring varies from 0 to 1. Therefore, it is evident that to each event E there exists there exists its complementary, non – E. If one does not pick the answer E = 'No opinion' one has to pick one of the remaining answers or possibilities which altogether constitute the complementary event of E. In simpler terms, a new born baby is either a girl (E) or a boy (non - E). Therefore, the possibility of a newly born being either a girls or a boy becomes a

certainty, since there is no alternative. The probability that an event will occur and its complementary adds up to certainty.

3. p (E) + p (non – E) = 1

The events E and non – E are called mutually exclusive events since they cannot occur simultaneously. Only complementary events are mutually exclusive. For example, you cannot toss a coin and have a head and tail up at the same time.

Addition Rule of Probability

If two events E_1 and E_2 are mutually exclusive, the probability that one or the other will occur is simply the sum of their respective probabilities. This rule is expressed by;

4. $p (E_1 \text{ U } E_2) = p (E_1) + p (E_2)$

The symbol U is borrowed from set theory and is used to denote either of the two events occurs.

If two events E_1 and E_2 are not mutually exclusive, that is they can happen at the same time, then the probability of either of them happening is the sum of the probability that each occurs without the other plus the probability that each occurs with the other. Using the symbol ∩ borrowed from set theory to denote that two events occur simultaneously; this probability can be written as p ($E_1 \cap E_2$). Therefore, the addition rule of probability for two events which are not mutually exclusive is;

5. $p (E_1 \text{ U } E_2) = p (E_1) + p (E_2) - p (E_1 \cap E_2)$

When two events are mutually exclusive, $E_1 \cap E_2$ cannot exist, in set theory it is an empty set and p ($E_1 \cap E_2$) = 0.

Conditional Probability

A frequent situation in reality is that the likelihood of an event occurring depends on whether another event has already occurred. For example, if a child is born into a wealthy family, the chance of that child enjoying a good education is higher than if he or she comes from a poor economic background. The probability of an event E_2 occurring given that some other event E_1 has already taken place is called conditional probability and is denoted by p (E_2 / E_1).

Based on the foregoing, we can distinguish between two types of events, dependent events and independent events. If the probability of one event occurring is dependent on another taking place, these two are called dependent events. Such events have been illustrated above.

On the other hand, if the probability of one event occurring does not affect the probability of the other occurring, and vice versa, these events are called independent events. In this case, the probability p (E_2) is not affected by E_1 having taken place and it is equal to the conditional probability. Symbolically;

6. p (E_2 / E_1) = p (E_2)

The above events are independent but not mutually exclusive. They can occur simultaneously.

Multiplication Rule of Probability

Given two events, mutually exclusive or not, the probability of at least one of them taking place is p (E_1 U E_2). What is the probability of both of them occurring? In other words, how is p ($E_1 \cap E_2$) found? Here a distinction has to be made between independent events and dependent events.

The probability of two independent events E_1 and E_2 both occurring is simply a product of their respective probabilities.

7. p ($E_1 \cap E_2$) = p (E_1) x p (E_2)

However, if the events are dependent, one has to take into account that once one event has taken place, the probability of the second event taking place changes accordingly. Thus one has to use the conditional probability instead of the probability of the event alone.

The probability that two dependent events E_1 and E_2 both occur is the product of the probability of one by the conditional probability of another.

8. $p(E_1 \cap E_2) = p(E_1) \times p(E_2 / E_1)$ Or

$p(E_1 \cap E_2) = p(E_2) \times p(E_1 / E_2)$

Clearly, if events are independent, using formula 6 in the above expression, formula 7 is obtained. Moreover, formula 8 allows for the conditional probability to be expressed by;

9. $p(E_2 \mid E1) = \frac{p(E_1 \cap E_2)}{p(E_1)}$ or

$p(E_1 \mid E_2) = \frac{p(E_1 \cap E_2)}{p(E_2)}$

Example

Let us illustrate the various properties of probability by an example related to probability sampling. A certain population of 150 women is divided into categories as shown below.

Table 4.1: Marital Status of Women and Household Status

	E_5 Non Heads of Households	E_6 Heads of Households	Total
E_1 Housewives	45	0	45
E_2 Divorcees	5	10	15
E_3 Widows	5	25	30
E_4 Unmarried	35	25	60
Total	90	60	**150**

a. What is the probability of each category to be drawn from the population?

i. The probability of selecting a Housewife is $p(E_1) = \frac{45}{150} = 0.3$

ii. The probability of selecting a Divorcee is p (E_2) = $\frac{15}{150}$ = 0.1

iii. The probability of selecting a Widow is p (E_3) = $\frac{30}{150}$ = 0.2

iv. The probability of selecting an Unmarried woman is p (E_4) = $\frac{60}{150}$ = 0.4

b. What is the probability of drawing out of the population a woman who has been previously married?

The probability of selecting out of the population a woman who has been previously married is: p (E_2 U E_3) = p (E_2) + p (E_3)

= 0.1 + 0.2

= 0.3

c. What is the probability of selecting a woman who is not presently married?

The probability of selecting a woman who is not presently married is:

p (E_2 U E_3 U E_4) = p (E_2) + p (E_3) + p (E_4)

= 0.1 + 0.2 + 0.4

= 0.7

d. What is the probability of selecting a woman who is neither a housewife nor a head of household?

The probability of choosing a woman who is neither a housewife nor a head of household is: p (non - E_2 U E_6) = 1 – p (E_1 U E_6)

= 1 – 0.17

= 0.83

e. What is the probability of selecting a woman who is either unmarried or is head of household?

The probability of selecting a woman who is either unmarried or is head of household is:

p (E_4 U E_6) = p E_4 + p E_6 – p ((E_4 ∩ E_6)

= 0.4 + 0.4 – 0.17

= 0.63

f. What is the probability of selecting a woman who is either divorced or is head of household.

The probability of selecting a woman who is either divorced or is head of household is;

$p(E_2 \cup E_6) = p(E_4) + p(E_6) - p(E_4 \cap E_6)$

$= 0.1 + 0.4 - 0.07$

$= 0.43$

g. What is the probability of selecting a selecting a head of household knowing that this woman is divorced? This refers to the conditional probability $p(E_6 / E_2)$. Since there are 15 divorcees forming E_2 and out of them 10 are heads of households, this conditional probability is; $p(E_6 / E_2) = \frac{10}{15}$

$= 0.67$

h. What is the probability of selecting a divorcee knowing that this woman belongs to the group of heads of households? This refers to the conditional probability $p(E_2 / E_6)$. Since there are 60 heads of household forming E_6 and out of the 10 are heads of households, the conditional probability is; $p(E_2 / E_6) = \frac{10}{60}$

$= 0.17$

i. What is the joint probability of being a widow and being a head of household?

Here the multiplication rule of probability applies. Thus the probability is;

$p(E_3 \cap 6) = p(E_3) \times p(E_6)$

$= 0.2 \times 0.4$

$= 0.08$

8.0 Summary

You have been informed in that probability is the likelihood of an event occurring. The study of probability is important in inferential statistics as it helps in making generalisations. The probability of any event ranges between 0 and 1. The unit has also explained that dependent events are events whose occurrence is dependent on another event occurring. On the other hand, independent events are those events whose occurrence does not dependent on another event taking place.

9.0 Self Assessment Questions

The table below shows the preferred sports of 2500 students at Kalomo College of Health Sciences:

Sport	**Sex**	
	Male	**Female**
Football	380	570
Golf	440	660
Cricket	180	270

Using the data above, compute the following;

a. The probability of selecting a student who like golf
b. The probability of selecting a student who either likes football or is male
c. The probability of selecting a female student knowing that she likes cricket
d. The probability of liking football and being male

UNIT FIVE: FREQUENCY DISTRIBUTIONS

1.0 Introduction

This unit discusses frequency distributions. The unit begins with explaining the types of frequency distributions. Thereafter we shall discuss the standard normal distribution and its significance in the field of research. We will then focus on Z – score and give examples to illustrate the use of the standard normal curve.

2.0 Aim of the Unit

The aim of this unit is to discuss frequency distributions.

3.0 Objectives of the Unit

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

- Identify different types of frequency distributions.
- Distinguish between the standard normal distribution and an ordinary normal distribution.
- Explain the characteristics of the standard normal curve.
- Calculate Z - scores

4.0 Required Material


In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

 It will take you approximately one hour thirty minutes to work through this unit.

6.0 Main Contents

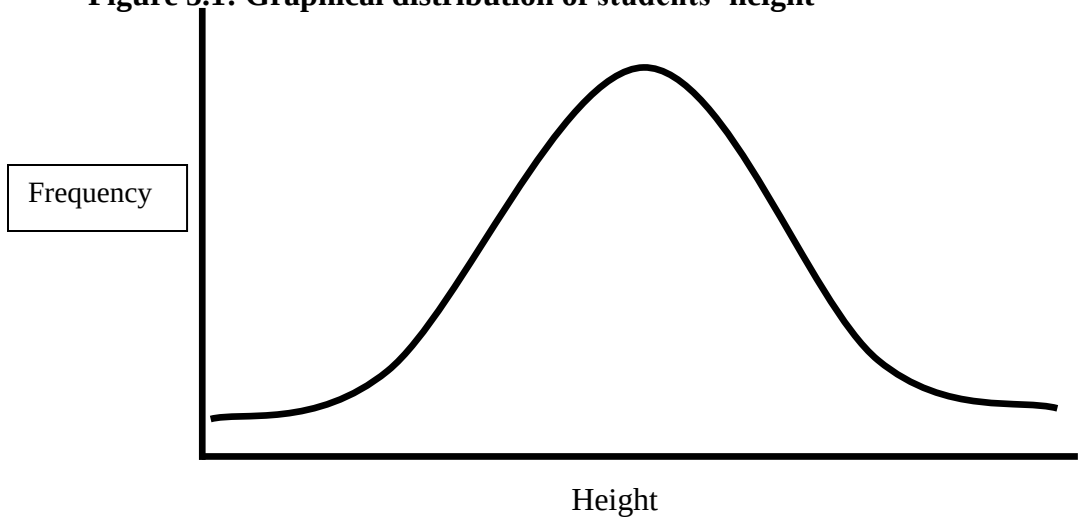
One of the first things examined in a frequency table are the peaks in the distribution. A peak in a distribution is a frequency or sets of adjacent frequencies that are larger than most of the other frequencies. For example, suppose you collected the following data on the height of students in a class n = 30

Table 5.1: Height of students in a class of 30

Height (m)	No. Of Students
1.2	4
1.5	6
1.6	10
1.7	6
1.8	4

When the above data is presented in form of a distribution, it will take the shape of a curve as shown in figure 5.1 below. The distribution of the data helps in identifying its general form.

Figure 5.1: Graphical distribution of students' height



Distributions may have distinctive forms with few low scores and many high scores; with many low scores and few high scores; with many score concentrated on the left or right side of the distribution and; with many scores concentrated on the in the middle of the distribution and few scores on either side of the distribution. The simplest way to describe a distribution is by visual representation. Examples are presented below.

Figure 5.2: Symmetrical Distribution

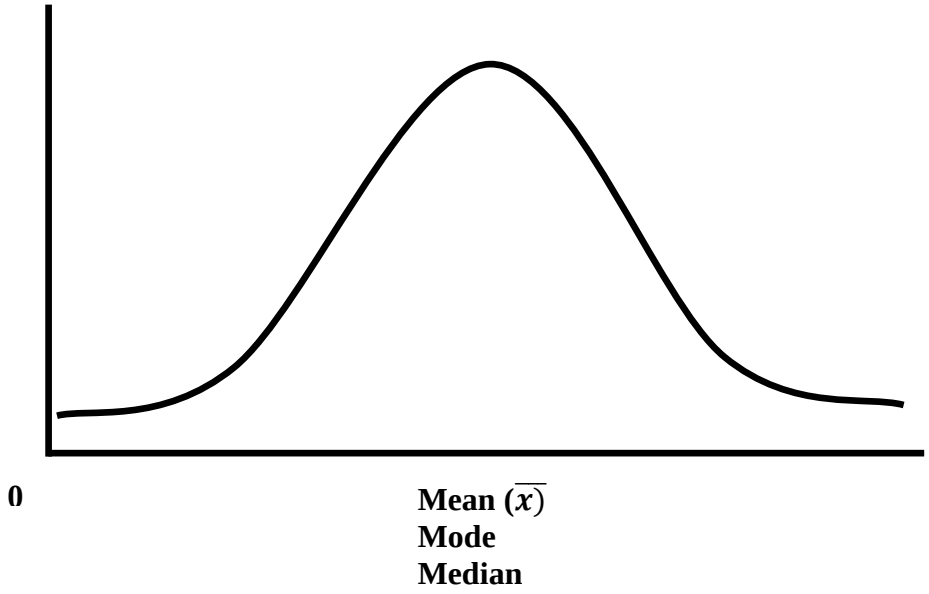


Figure 5.3: Positively Skewed Distribution

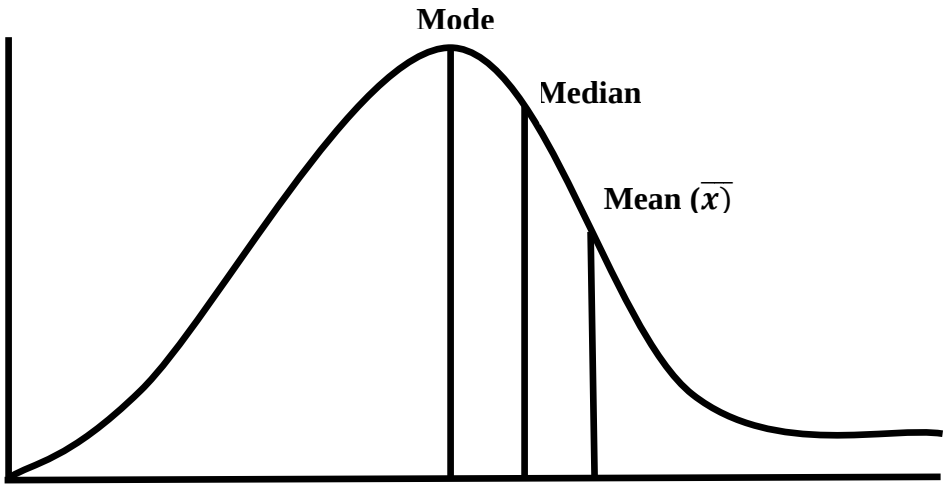
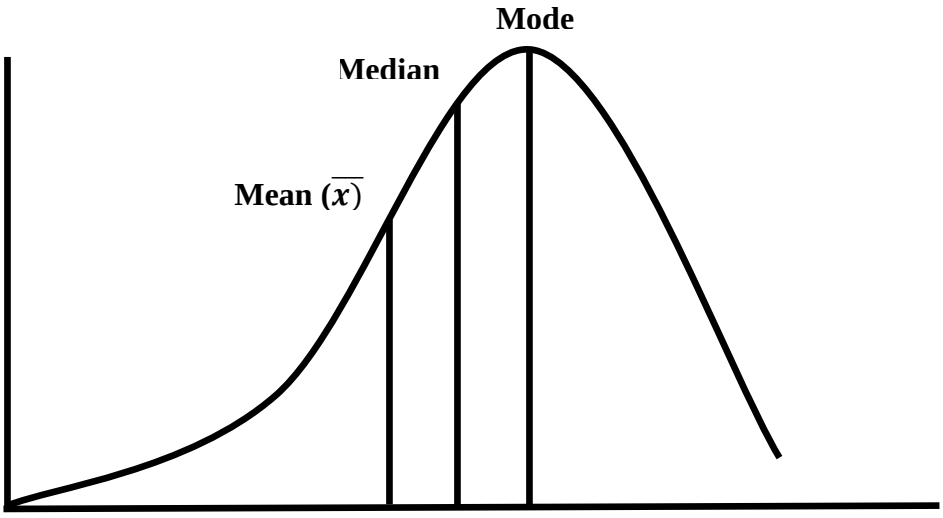


Figure 5.4: Negatively Skewed Distribution



The values of the variables are presented along the baseline and the area under the curve represents the frequencies. The distribution in figure 5.2 is symmetrical, that is the frequencies at the right and left tails of the distribution are identical, so if the distribution is divided into two halves, each will be the mirror image of the other. This means that most of the observations are

concentrated at the middle of the distribution and there are few observations with very high or very few scores. Many variables tend to be distributed symmetrically and this form of distribution plays an important role in the field of statistics.

When there are more extremely high scores, the distribution is positively skewed as shown in figure 5.3 and when there are more extremely low scores, the distribution is negatively skewed as shown in figure 5.4. Skewed distributions are also referred to as non symmetrical; they have more extreme cases in one direction of the distribution than in the other.

Skewness can also be identified according to the measures of central tendency. In a symmetrical distribution, the mean will coincide with the median and the mode. In skewed distributions, there will be discrepancies between these measures. In a negatively skewed distribution, the mean will be pulled in the direction of lower scores. In a positively skewed distribution, the mean will be pulled closer to higher scores. The property of skewed distributions makes the choice of a measure of central tendency a critical issue. Since the mean is pulled in the direction of the extreme scores, it loses its usefulness as a representative measure. In such instances, it might be useful to employ the median or the mode instead.

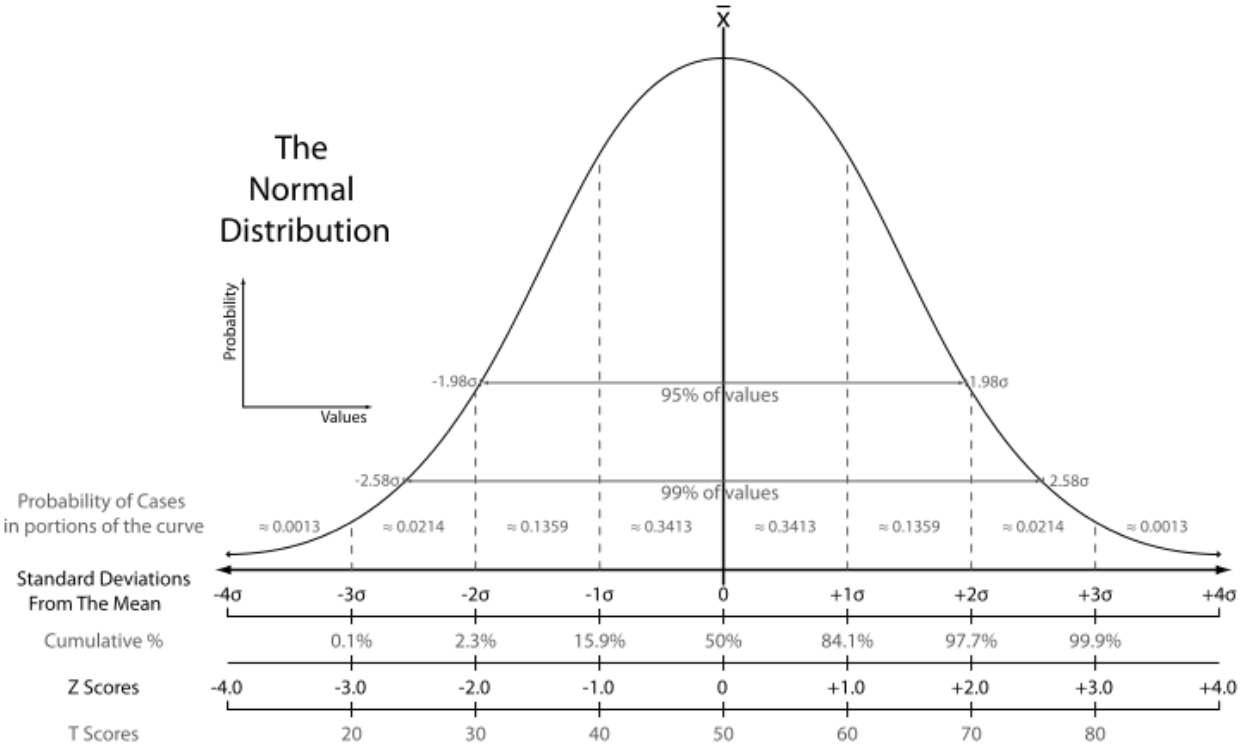
6.1 The Standard Normal Distribution

One type of symmetrical distribution called the standard normal distribution has great significance in the field of statistics. The standard normal distribution has the following properties.

- It is ***symmetrical*** and bell shaped. Each side of the curve constitutes 50% of the distribution.
- It is ***unimodal***, this means that the mean, the median and the mode coincide at the centre of the distribution.
- It is ***asymptomatic***, this means the area between the curve and baseline on both sides extends to infinity.
- The area under the normal curve represents a probability of 1. Each side of the curve represents a probability of 0.5.

- In a normal curve, a fixed proportion of observations lie between the mean and fixed standard deviation units. The proportions can be seen from the figure below.

Figure 5.5: The Normal Distribution



The mean of the distribution divides it exactly in half, 0.3413 (34.13%) of observations lie between the mean and 1 σ (one standard deviation) to the right of the mean and the same proportion falls between the mean and one standard deviation to the left. The plus sign indicates one standard deviation above the mean, the minus sign indicates standard deviation below the mean. Thus 68.28 % of all observation fall between $x \pm 1\sigma$, 95.46% of all observations fall between $x \pm 2\sigma$ and 99.73% of all the observations fall between $x \pm 3\sigma$.

In any univariate analysis that is normally distributed, the proportion of observations included within fixed distances of the mean can be determined. For example, in a distribution of intelligence tests with a mean score of 110 and a standard deviation of 10, 68.28% of subjects will have an IQ of $110 \pm 1\sigma$; that is an IQ of between 100 and 120 and 95.46% of subjects will have an IQ of $110 \pm 2\sigma$; that is an IQ of between 90 and 130.

6.2 Standard Normal Distribution and Standard Scores

On the basis of the mean and the standard deviation of the randomly distributed data it is possible to construct a standard normal distribution. The standard normal distribution resembles an ordinary normal curve. There is only one standard normal curve distribution unlike ordinary normal curve which can be several. In order to go round these problems it becomes necessary to standardize the ordinary normal curve. This standard normal curve has a mean of 0 and a standard deviation of 1 whereas ordinary normal curves will have different means and standard deviations. For example, examination results for PAM 2010 for the year 2012, 2013 and 2014 may conform to a normal distribution but these three normal distributions may have different means and standard deviations. Therefore, these normal distributions have to be standardized so that they have the same mean and standard deviation. Standardization uses a formula referred to as standard score or standard normal deviation or z-score. The formula is:

$$\mathbf{Z = \frac{x - \bar{x}}{s}}$$

Where Z = Number of standard deviation units

X = Any observation or score of the distribution;

$\bar{X}$ = Mean of the distribution and;

S = Standard deviation of the distribution.

Z also referred to as a standard score, expresses the distance between a specific observation (x) and the mean in terms of standard deviation units. Special tables have been constructed for the standard normal curve. The tables enable you to determine the proportion of observations that lie between the mean and any observation in the distribution. **A table of Areas Under the Normal Curve is attached at the end of the module**. The first two digits of Z are listed in the left hand column; the third digit is shown across the row. For example, a Z – score of 1 is .3413 (Z – scores in the table are expressed to two decimal places. In this case, the value of a

Z – score of 1 was found by first locating 1.0 in the first column and then moving across to horizontally to .00). The table only shows half of the curve's proportions because the curve is symmetrical. Thus the distance between the mean and a Z of -1.0 is identical to the distance between the mean and a Z of 1.0.

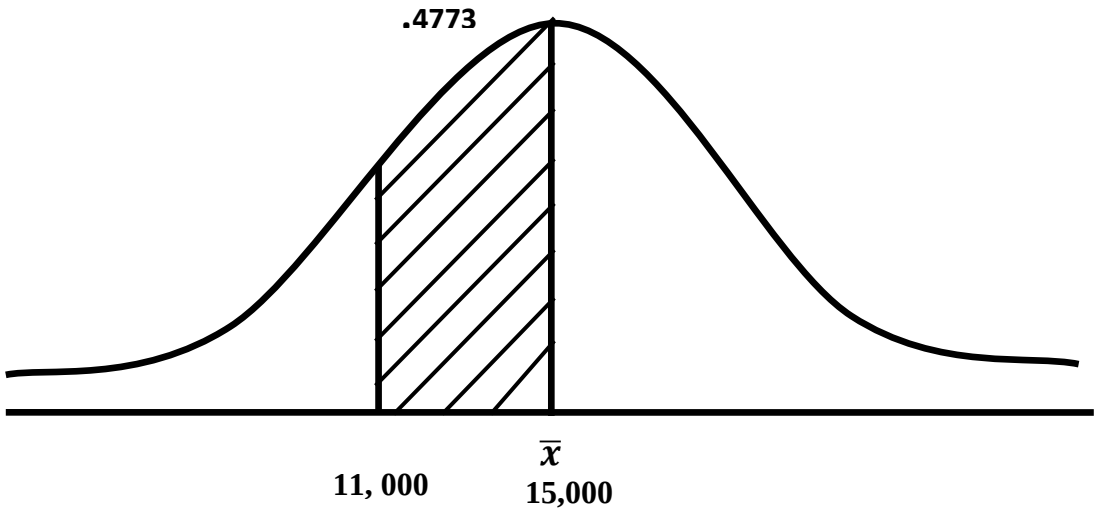
Researchers use the standard normal curve to evaluate the proportion of observations included within a desired interval and to determine the probability of various events. To do this, the raw scores must be converted to standard deviation units in order to use the tables which report the area under the standard normal curve. When raw data has been converted into standard scores, a single table can be used to evaluate distributions regardless of the scale on which the data were measured. Therefore, distributions measured on different scales can be compared. Observations are converted into standard deviational units using the formula stated on page 36.

To illustrate the use of the normal curve, suppose the distribution of income in a particular community is normal, its mean is K15, 000 and the standard deviation is K 2,000.

1. Calculating proportions

a. What is the proportion of people in the community who have an income between K11,000 and K 15,000?

Figure 5.6 Proportion of people earning between K11, 000 and K15, 000



First we convert the figure into standard deviation units.

$$Z = \frac{x - \bar{x}}{s}$$

$$= \frac{11,000 - 15,000}{2,000}$$

$$= \frac{-4,000}{2,000}$$

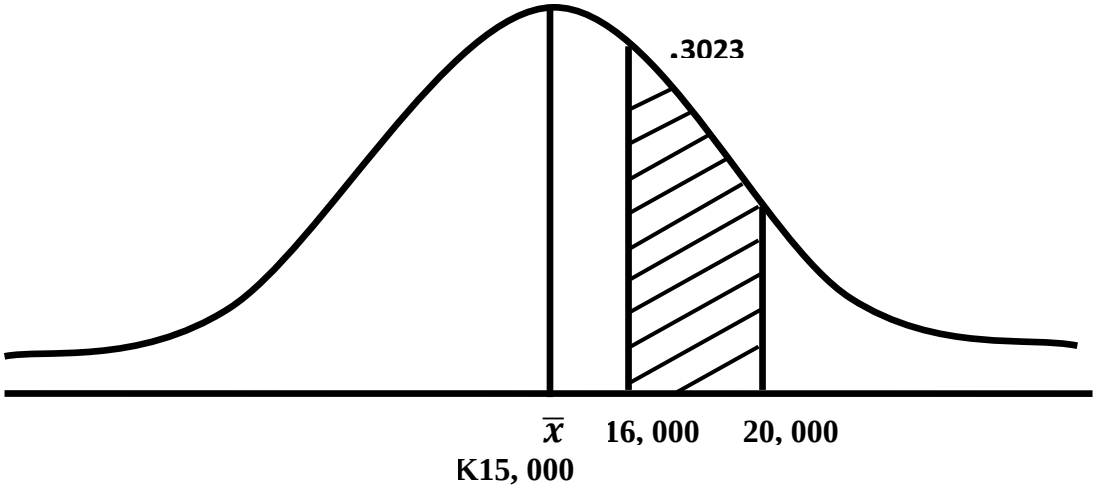
$$= -2$$

$$= .4773$$

When you go to the areas under the normal curve, the value of a Z – score of 2 is .4773, therefore, the proportion of people who have an income between K11, 000 and K15, 000 is .4773.

b. What is the proportion of people in the community who get between K16, 000 and K20, 000?

Figure 5.7: Proportion of people earning between K16, 000 and K 20,000



First we need to convert the two figures into standard deviation units.

$$Z_1 = \frac{x - \bar{x}}{s}$$

$$= \frac{16,000 - 15,000}{2,000}$$

$$= \frac{1,000}{2,000}$$

$$Z_2 = \frac{x - \bar{x}}{s}$$

$$= \frac{20,000 - 15,000}{2,000}$$

$$= \frac{5,000}{2,000}$$

= 0.5 = 2.5

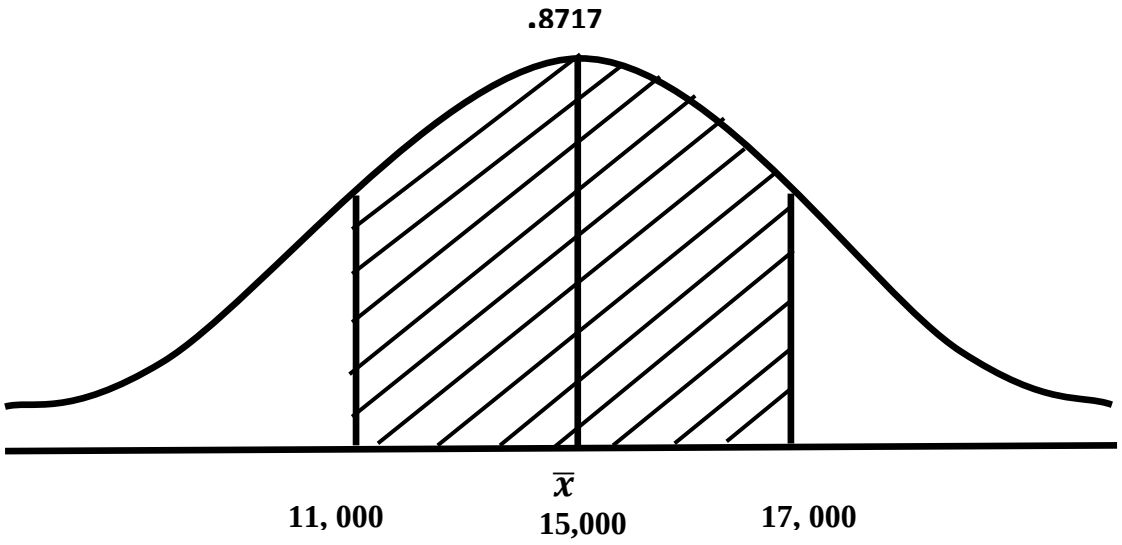
= .1915 = .4938

The Z – score value of 0.5 is .1915 and the Z- score value of 2.5 is .4938

Therefore, the proportion of people in the community who get between K16, 000 and K 20,000 is .4938 - .1915 = .3023

c. What is the proportion of people who get between K11, 000 and K 17,500?

Figure 5.8: Proportion of people earning between K 11,000 and K 17, 500



$Z_1 = \frac{x - \bar{x}}{s}$

$= \frac{11,000 - 15,000}{2,000}$

$= \frac{-4,000}{2,000}$

= -2

= .4773

$Z_2 = \frac{x - \bar{x}}{s}$

$= \frac{17,500 - 15,000}{2,000}$

$= \frac{2,500}{2,000}$

= 1.25

= .3944

Therefore, the proportion of people who get between 11,000 and 17,500 is .4773 + .3944 = .8717

2. Calculating percentages

a. What is the percentage of people whose income is between K11, 000 and K15, 000?

$Z = \frac{x - \bar{x}}{s}$

$= \frac{11,000 - 15,000}{2,000}$

$= \frac{-4,000}{2,000}$

= -2

= .4773

= .4773 X 100 (Note that to find percentages, you calculate the Z – score value by 100)

= 47.73%

This means that 47.73% of people in this community have an income of between K11, 000 and K 15, 000.

b. What is the percentage of people in the community who get between K16, 000 and K20, 000?

First we need to convert the two figures into standard deviation units.

$Z_1 = \frac{x - \bar{x}}{S}$

$= \frac{16,000 - 15,000}{2,000}$

$= \frac{1,000}{2,000}$

= 0.5

= .1915

$Z_2 = \frac{x - \bar{x}}{s}$

$= \frac{20,000 - 15,000}{2,000}$

$= \frac{5,000}{2,000}$

= 2.5

= .4938

The Z – score value of 0.5 is .1915 and the Z- score value of 2.5 is .4938

The proportion of people in the community who get between K16, 000 and K 20,000 is .4938 - .1915 = .3023, hence the percentage of people who get between K16, 000 and K20, 00 is .3023 x 100 = 30.23%

c. What is the proportion of people who get between K11, 000 and K 17,500?

$Z_1 = \frac{x - \bar{x}}{s}$

$= \frac{11,000 - 15,000}{2,000}$

$= \frac{-4,000}{2,000}$

= -2

= .4773

$Z_2 = \frac{x - \bar{x}}{s}$

$= \frac{17,500 - 15,000}{2,000}$

$= \frac{2,500}{2,000}$

= 1.25

= .3944

The proportion of people who get between 11,000 and 17,500 is .4773 + .3944 = .8717 , hence the percentage of people who get between 11,000 and 17,500 is .8717 x 100 = 87.17%

3. Calculating probabilities

a. What is the probability of a member of this community getting between K11,000 and K15,000?

$Z = \frac{x - \bar{x}}{s}$

$= \frac{11,000 - 15,000}{2,000}$

$= \frac{-4,000}{2,000}$

= -2

= .4773

The probability of a member of the community getting between K11,000 and 15, 000 is 0.4773. Note that the proportion of the area under the normal curve is also indicate of the probability of a given event.

4. Computing the total number of people in a given interval

a. If the community has a population of 1200, how many people get between K11, 000 and K17, 500?

$Z_1 = \frac{x - \bar{x}}{s}$

$= \frac{11,000 - 15,000}{2,000}$

$= \frac{-4,000}{2,000}$

$Z_2 = \frac{x - \bar{x}}{s}$

$= \frac{17,500 - 15,000}{2,000}$

$= \frac{2,500}{2,000}$

= -2 = 1.25

= .4773 = .3944

The proportion of people who get between 11,000 and 17,500 is .4773 + .3944 = .8717, hence the percentage of people who get between 11,000 and 17,500 is .8717 x 100 = 87.17%. To find the actual number of people who get between K11, 000 and K17, 000, we simply find 87.17% of 1200.

$= \frac{87.17}{100} X1200$

= 87.17 X 12

= 1046.04

= 1046 people

The number of people cannot be expressed in decimals, so when dealing with people, the answer should be rounded off to a whole number.

The basic rule in using z –scores is that if the z – scores are located on the opposite sides of the centre, you add the corresponding proportions of the area under the curve. On the other hand, if z-scores are on the same side of the centre, you subtract the corresponding proportions.

7.0 Summary

You have been informed in this unit that one type of symmetrical distribution called the standard normal distribution has great significance in the field of statistics. It is symmetrical, unimodal, asymptomatic, the areas under the curve represents a probability of 1 and it has a fixed proportion of observations which lie between the mean and fixed standard deviation units. The unit has also shown that z – scores or standard scores express the distance between a specific observation and the mean in standard deviation units.

8.0 Self Assessment Questions

After reading through unit five, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit five again.

Using the table of the areas under the normal curve attached in the appendices, find the following;

a. The areas associated with a Z- score of 2.33
b. The area associated with a Z- score of -1.08
c. The area between a Z – score of -0.52 and 2.01
d. The area between a Z –score of – 3.06 and -0.08

In a normally distributed sample of 1200 children, the mean age is 14.4 years and the standard deviation is 2.5 years.

a. How many children are between 12 and 16 years old?
b. How many children are older than 18 years?
c. What is the percentage of children who are younger than 8 years?
d. What is the probability of a child in this sample being above 15 years?

UNIT SIX: INTRODUCTION TO ESTIMATES WITH SAMPLES

1.0 Introduction

This unit gives an introduction to estimates with samples. The unit begins with discussing the sampling distribution of means. Thereafter, we shall discuss the standard error of the mean and the standard error of a proportion. We will then focus on confidence levels, confidence limits and confidence intervals. The last part of this unit will dwell on the estimation of parameters based on samples.

2.0 Aim of the Unit

The aim of this unit is to introduce you to estimation of parameters based on sample statistic.

3.0 Objectives of the Unit

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

- Estimate parameters based on samples.
- Distinguish between point estimates and confidence interval estimates

4.0 Required Material

In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

 It will take you approximately two hours to work through this unit.

6.0 Main Contents

The purpose of sampling is to enable researchers estimate the population parameters based on sample statistics. For example, when you want to estimate the outcome of a presidential election, you may sample a few people and use the results obtained from the sample to estimate what the results of the actual election will be. Similarly when you want to know the average income of people in Zambia you would select a sample and use the value obtained from the sample to estimate the population mean, although the sample mean is unlikely to be exactly the same like the population mean. There will inevitably be some degree of error in the sample.

If you took a very large number of samples of the population, you would expect all the sample means to be close to the true population means, but spread around the mean with some sample means higher and some sample means lower than the true population mean. These means will not all be the same and they can be plotted as a frequency distribution. ***This distribution is called a sampling distribution of the means.*** For example, if you draw 100 samples of the same size to find the mean income of households in Zambia, you will end up with 100 different sample means. The sampling distribution of means will be normal and it will bear all the characteristics of the normal curve.

The sampling distribution of sample means is important in statistical inferences because it is used in significance tests and hypothesis testing. Although the sampling distribution of means is similar to the ordinary normal curve, the differences are that whereas the ordinary normal curve deals with ordinary observations, the sampling distribution of means deals with the distribution of sample means.

6.1 The concept of Standard Error of the Mean

The standard deviation of a sampling distribution is known as the standard error of the mean. It is given by the formula;

$S_{\bar{x}} = \frac{S}{\sqrt{n}}$

$S_{\bar{x}}$ = Standard error

S = Standard deviation of the sample

$\sqrt{n}$ = Square root of the sample size

The importance of the standard error is that it is used to estimate how accurately the sample mean estimates the population mean. The smaller the standard error, the more accurately the sample mean can estimate the population mean. When you increase the sample size, the Standard Error (SE) reduces.

6.2 Standard Error of a Proportion

Many surveys and samples attempt to estimate a proportion, rather than an arithmetic mean. This is true of surveys of attitudes or opinions about an issue or the percentage of times an issue occurs. Suppose you want to know what proportion of people would vote for a name political party in elections, several sample may be obtained and the proportion of votes in each sample could be arranged into a sampling distribution which would be; normally distributed; It would have a mean equal to the proportion of voter in the population who would vote for a named political party and; it would have a standard error equal to the standard deviation of a proportion.

The standard error of a proportion is given by the formula;

$\boldsymbol{S_{\bar{X}} = \sqrt{\frac{p.q}{n}}}$ where; p = proportion; q = 1- p and; n = sample size

6.3 Confidence Levels, Confidence Limits and Confidence Intervals

The properties of a normal distribution, together with the rule that sample means are normally distributed around the true population mean with a standard deviation of equal to the standard error, we can predict the following (using the normal distribution table);

- 68% of all the sample means will lie within 1 standard error of the mean. In other words, there is a 68% probability that the population mean lies within the range: sample mean ± 1standard deviation.
- 95% of all sample means will be within 1.96 standard errors of the population mean. In other words, there is a 95% probability that the population mean lies within the range: sample mean ± 1.96 standard deviations.
- 99% of sample means will be within 2.58standard errors of the population mean. In other words, there is a 99% probability that the population mean lies within the range: sample mean ± 2.58 standard deviations.

These degrees of certainty are known as confidence levels and the ends of the ranges around the sample mean are called confidence limits. The ranges are called confidence intervals. The word confidence is used because you are trying to find out how confident you can be that your sample is representative of the population as a whole.

6.4 Estimation of Parameters

There are two types of parameter estimates that can be used; these are point estimates and confidence interval estimates.

a. ***Point estimates*** are the simplest way of estimating population parameter; you merely take the sample statistic to represent the population parameter.

   Sample statistic = Population parameter

   When using point estimates, an error in the estimation process is likely to occur as the sample statistic might be below or above the population parameter. ***This error is referred to as sampling error***, which is the difference between the population parameter and the sample statistic.

b. ***Confidence Interval Estimates***. A confidence interval estimate is a particular kind of estimate of a population parameter. It denotes the range within which a population parameter might be found it also denotes the level of confidence or probability that the population parameter lies within the stated confidence limits.

C.I.E = Sample statistics ± Critical value x Standard error

A critical value is the Z – score value which shows the ranges in which a given confidence level falls under the normal curve. It is like a confidence limit for a given level of confidence. For example, a confidence level of 99% will be represented by a Z – score of ± 2.58.

Example 1: Suppose you are trying to estimate the mean age of UNZA students, you sample 100 students; the mean age for this sample is 24.7 years and the standard deviation is 1.02 years. Use 95% confidence level to establish the mean age of UNZA students.

C.I.E = Sample statistic ± Critical x standard error

= 24.7 ± 1.96 x 0.1

= 24.7 ± 0.196

= 24.5 – 24.9

Interpretation: You can be 95% certain that the mean age of all UNZA students lies between 24.5 years and 24.9 years.

Note that the standard error was computed using the formula;

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

$$= \frac{1.02}{\sqrt{100}}$$

$$= \frac{1.02}{10}$$

$$= 0.1$$

The critical value for was found by following the following steps;

95% confidence level means we should identify the z – scores in which 95% of the area under the normal curve falls. You know that the whole area under the normal curve adds up to 100% with 50% on either side of the mean. Similarly, a confidence interval of 95% will have $\frac{95\%}{2}$ = 47.5% on either side of the mean. 47.5% represents a proportion of $\frac{47.5}{100}$ = 0.4750 or .4750 of the area under the normal curve. Having done this, you know you go to the table of the areas under the normal curve in the appendices of this module and look for the z – score which will represent an area of .4750. In this case you will note that a z – score of 1.96 represent .4750 of the area under the normal curve.

Example 2: A researcher randomly selects 100 students at UNZA to determine their utilization of the University library. Out of the sample, 55% said they prefer studying from their rooms. Estimate the proportion of students who prefer studying from the library at 99% level of confidence.

Standard error of a proportion is = $\sqrt{pq/n}$

= $\sqrt{(0.55x0.45)/100}$

= 0.04975

In this case p = $\frac{55}{100}$ = 0.55

q = (1 –p)

1 – 0.55

0.45

C.I.E = Estimated proportion ± critical value x standard error

= 0.55 ± 2.58 x 0.4975

= 0.421645 – 0.678355

= 42% - 68%

Interpretation: You can be 99% confident that 42% - 68% of UNZA students prefer studying from their rooms.

7.0Summary


You have been informed in this unit that the purpose of sampling is to enable researchers estimate population parameters based on sample statistics. You have further been informed that the importance of the standard error is that it is used to estimate how accurately the sample mean estimates the population mean. The smaller the standard error, the more accurately the sample mean can estimate the population mean. When you increase the sample size, the standard error reduces.

8.0 Self Assessment Questions


After reading through unit six, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit six again.

Given a random sample of 100 malnourished adults with a mean weight of 26kg and a standard deviation of 5.2kg;

a. Find the point estimate of the malnourished adults mean weight.
b. Compute the standard error and interpret the answer.
c. Establish the 90% confidence interval estimate for the population mean.
d. Establish the 98% confidence interval for the population mean.

UNIT SEVEN: HYPOTHESIS TESTING

1.0 Introduction

This unit discusses hypothesis testing. The unit begins with highlighting the steps in hypothesis testing and defining the major concepts in hypothesis testing. Thereafter, the unit will discuss hypothesis testing between means of large samples. We will then look at hypothesis testing between proportions of large samples as the last part of this unit.

2.0 Aim of the Unit

The aim of this unit is to discuss the procedure for hypothesis testing.

3.0 Objectives of the Unit

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

- Outline the steps in hypothesis testing.
- Define the major concepts in hypothesis testing.
- Conduct hypothesis testing for difference between means of large samples.
- Conduct hypothesis testing for differences between proportions of large samples.

4.0 Required Material


In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

McNabb D. (2009) *Research Methods for Political Science: Quantitative and Qualitative Methods*, New Delhi: PHI Learning Private Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

 It will take you approximately one hour thirty minutes to work through this unit.

6.0 Main Contents

We use samples to make estimates of the population mean or proportion. In the previous section, it was explained how we can estimate a confidence interval for a population mean or proportion, at a given level of confidence. We can also use sample statistics to test a theory or assumption that we have made about the population. This is referred to as hypothesis testing. The purpose of hypothesis testing is to test an assumption that we can make about the population.

6.1 Procedure in Hypothesis Testing

The following are the steps which are followed in hypothesis testing:

1. Formulation of the research and null hypothesis

a. Research hypothesis. A research hypothesis is a hypothesis which is derived from theory, observations and literature review. A research hypothesis is a hypothesis which a researcher is interested in testing in order to prove or disapprove it. Hypotheses are expressed in a directional way. Research hypothesis is denoted by H_1 . For example, a research hypothesis may read as; Male students participate more in demonstrations than female students. In this case a sign μ_1 will be used to denote male students and μ_2 will be used to denote female students. Hence the research hypothesis in question would be presented as:

$$H_1 : \mu_1 > \mu_2$$

b. Null hypothesis. A null hypothesis is an antithesis or opposite of the research hypothesis. It is the hypothesis that is directly tested and contradicts the true hypothesis. The null hypothesis is denoted by $H_{O.}$ The null hypothesis of the research hypothesis above would read as; Male students are less involved in demonstrations than female students. Hence the null hypothesis in question would be presented as:

$$H_O : \mu_1 < \mu_2$$

In hypothesis testing rejection of the null hypothesis increases the probability that the research hypothesis would be correct. Whereas acceptance of the null hypothesis increases the likelihood that the research hypothesis could be wrong.

2. Choosing appropriate statistical tests and making assumptions

Once the hypothesis has been formulated, the next thing is to decide on the appropriate statistical test that must be done. This choice has certain assumption or conditions which must be met. These assumptions are used to justify the choice of the appropriate statistical test. The most important assumptions are that; (i) assumptions relating to the distribution of parameters; (ii) assumptions relating to the scale of measurements of variables and; (iii) assumptions relating to the sampling designs, this is always assumed to be random sampling. Based on these assumptions, there are two major types of statistical tests used in hypothesis testing, these are ***parametric tests*** and ***non parametric tests***. A parametric test is a statistical test based on several assumptions about the parameters of the population from which it is drawn. Among the most important ones are that the observations must be drawn from a normally distributed population and that the variables are measured on either the interval or ration scale. On the other hand, a non parametric test does not require the normality of the distribution or the variables to be interval or ratio.

3. Obtaining the appropriate sampling distribution

There are two types of sampling distribution which are used in hypothesis testing. These are determined by the sample size. The standard normal distribution or Z – distribution will be used if the sample size is more than 30 ($n > 30$). If the sample size is less than 30 ($n < 30$), then the t – distribution may be used.

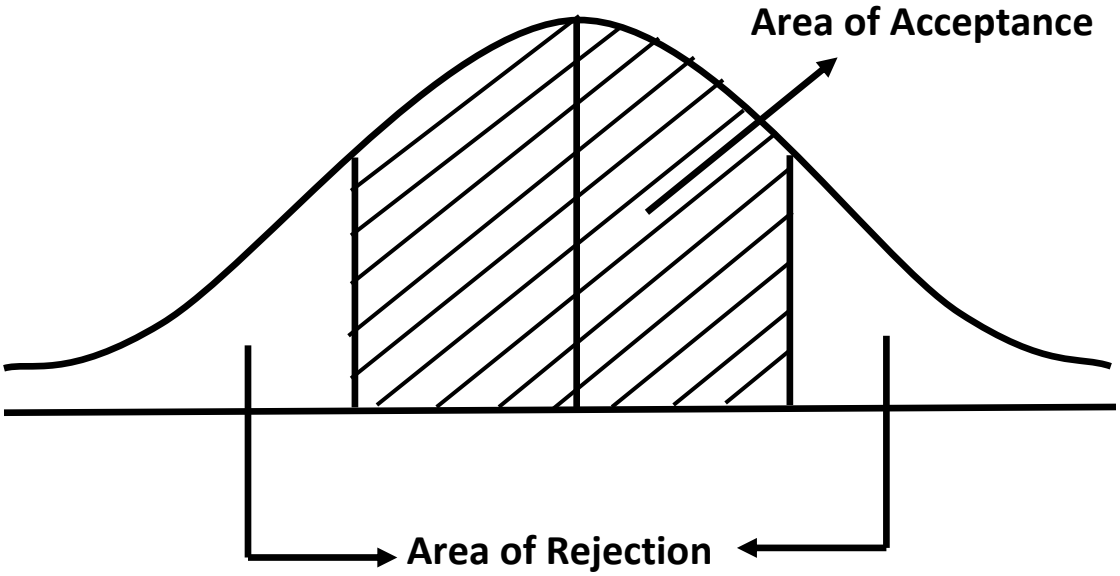
Once the choice of an appropriate sampling distribution has been made, the sampling distribution selected will be used to choose the critical values.

4. Choosing the significance level and critical regions

a. Critical Regions

Once the sampling distribution has been obtained, determination of the critical values follows. The critical value in the case of a Z – distribution is the Z – score obtained from the standard normal curve. The critical value in hypothesis testing determines the boundary or basis on which you accept or reject the null hypothesis. In hypothesis testing, a test statistic must be computed. The test statistic is also known as the observed value.

Figure 7.1: Areas of rejection and acceptance in hypotheses testing



b. Level of significance

The level of significance is used to compute the critical values to use in demarcating the areas of rejection or levels of acceptance. The level of significance is the probability of rejecting the null hypothesis (H_0) when it is in fact true. The probability of rejecting a null hypothesis when it is in fact true is referred to as type I error. The level of significance is denoted by α (Alpha) symbol. The opposite of type one error is type II error. Type two error is the probability of accepting the null hypothesis when it is false. Symbolically, it is denoted by beta (β).

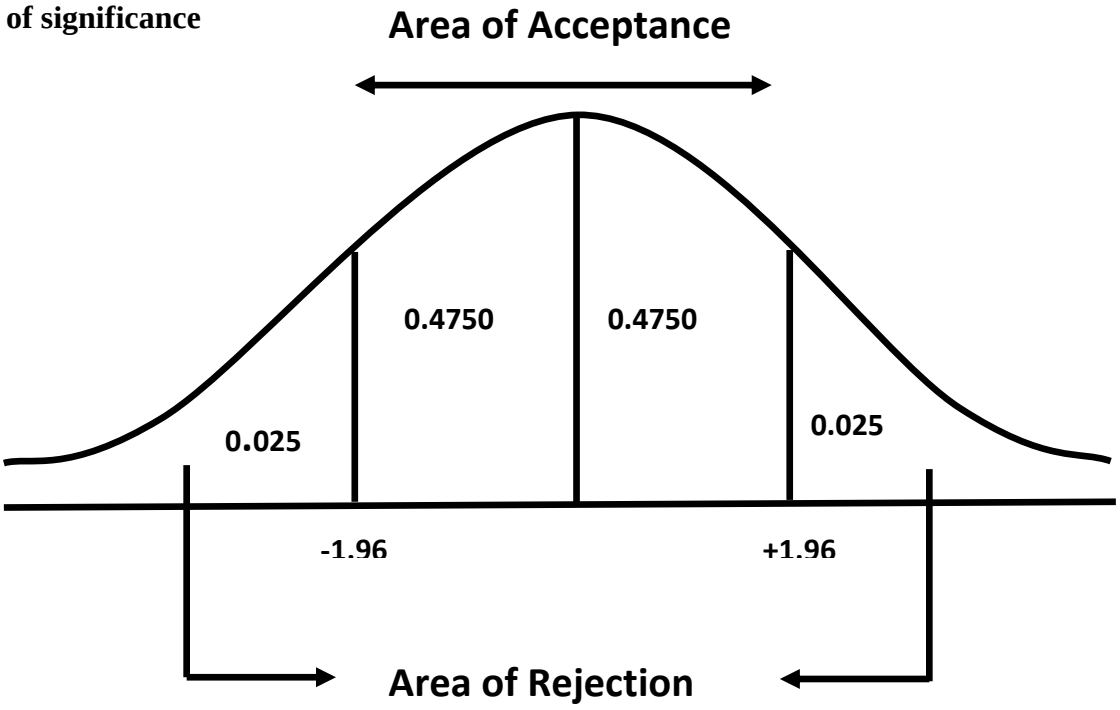
c. Setting the critical regions using the level of significance.

i. Critical regions for a two tailed hypothesis test. A two tailed hypothesis test refers to a situation where the research hypothesis is non – directional. For example, the average income of civil servants is not equal to K2,500.00

$$H_1 : \mu \neq \text{K2, 500.00}$$

At 5% level of significance, the total proportion of the area of rejection is .05. Since this proportion is evenly split between the two sides of the standard normal distribution it is going to be divided by two in order to find the area of rejection on both sides of the curve. Therefore, the critical region will be $\frac{0.5}{2}$ =0.025. The Z-score which demarcates the region with a proportion of 0.025 on the left of the curve is -1.96 and on the right of the curve is +1.96 as shown in the diagram below.

Figure 7.2: Areas of acceptance and rejection for a two tailed hypothesis at 5% level of significance



At 1% significance level, the proportion of the area of rejection is .01. Since this proportion is evenly split between the two sides of the standard normal distribution, it is going to be divided by two in order to find the region of rejection on either side of the curve. Therefore, the critical region will be $\frac{0.1}{2}$ = 0.05. The Z –score which demarcates the region of 0.05 on the left of the curve

e is – 2.58 and on the right of the curve is + 2.58.

ii. Critical regions for a one tailed hypothesis test. A one tailed research hypothesis refers to situation where the research hypothesis is directional. For example, the average income of civil servants is more than K2, 500.00 ($H_1 : \mu >$ K 2,500.00) or there are more men involved more in decision women ($H_1 : \mu_1 > \mu_2$). In this case, you place the critical value in one tail only, in the right side of the curve. The area of acceptance is the area to the left of the critical value; the extreme right is the area of rejection.

At 5% level of significance, the proportion of the area of rejection under the standard normal curve is 0.05. Since this area is only on one side of the curve, it will not be divided by two as was the case in a non – directional hypothesis. Therefore, the z – score which represents the critical region of 0.05 is 1.65. At 1% level of significance the proportion of the area of rejection under the normal curve is 0.01, the Z-score which represents this critical region is 2.58.

The z – score values which correspond to the level of significance determine where the critical regions are located, that is the cut off points or values for the rejection or acceptance of the null hypothesis.

5. Computing the test statistic

At this stage you run the data in your computer and come up with a test statistic also known as the observed value. This will come in form of a z – score or t- score depending on the sample size.

6. Decision making

Depending on the area of acceptance or rejection of the null hypothesis, make a decision whether to accept or reject it based on the test statistic or observed value.

7. Conclusion

Make a conclusion based on your decision, go beyond the mathematics and express yourself in your own words.

Example of hypothesis testing

In a dispute between Chimuka Mining Company Limited and the workers' union over employee remuneration, management claims that the workers earn more than K 7,500 but the union dispute the claim. As a labour consultant, you are hired to ascertain this claim which the union disputes. You select a random sample of workers at the mine, n = 100. The mean income of workers based on the sample is K7, 900 and a standard deviation of K 1,500. You decide to establish if the claim by management is true by using a 5% level of significance.

Step 1. Formulate the research and null hypothesis

H_0 : μ = K 7,500

H_1 : μ > K 7,500

Step 2. Assumptions

i. The subjects are randomly and independently selected
ii. The population distribution is normal
iii. The scale of measurement is interval

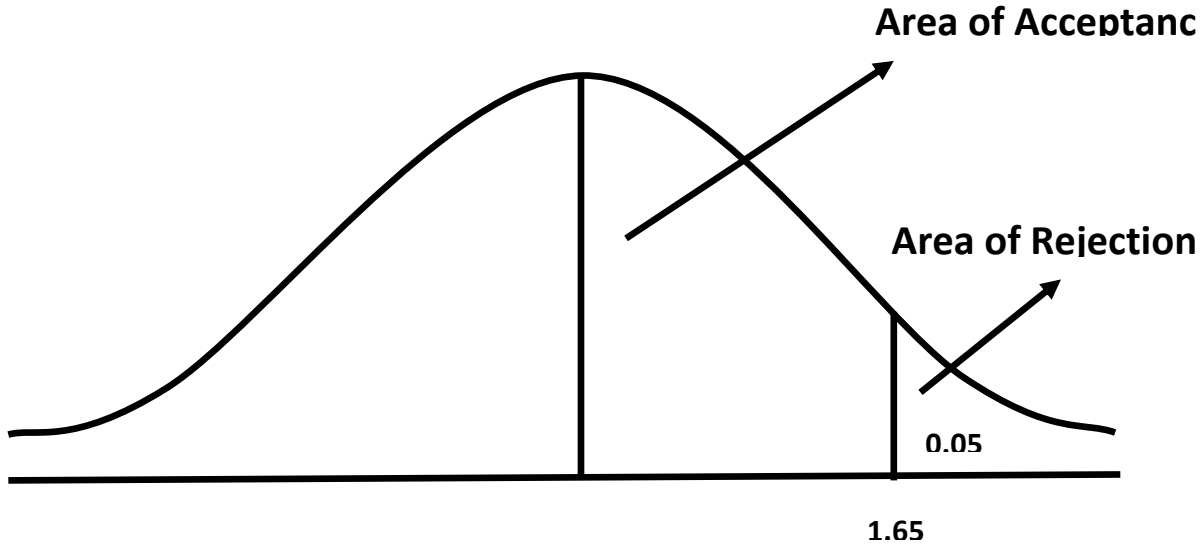
Step 3. Decision Rules

Given 5% level of significance and using a standard normal distribution, the critical value for a one tailed directional test is 1.65.

If Z_{obs} < +1.65 Accept H_0

If $Z_{obs} \geq$ +1.65 Reject H_O

Figure 7.3: Areas of Acceptance and rejection for a one tailed hypothesis at



Step 4. Computation

$Z_{obs} = \frac{x - \mu}{S_{\bar{x}}}$

$S_{\bar{x}}$ = Standard error

$S_{\bar{x}} = \frac{s}{\sqrt{n}}$

$= \frac{1,500}{\sqrt{100}}$

$= 150$

$Z_{obs} = \frac{x - \mu}{S_{\bar{x}}}$

$= \frac{7,900 - 7,500}{150}$

$= \frac{400}{150}$

$= 2.67$

Step 5. Decision

Compare the observed value with a critical value of 1.65. Because Z_{obs} is greater than 1.65, we reject the null hypothesis (H_0).

Step 6. Conclusion

It is highly likely that Chimuka Mining Limited is right in its claim that it pays its workers more than K 7,500.

6.2 Testing Hypothesis about a Proportion

A nutritionist at the Ministry of Health in Lusaka claims that more than 50% of children in Lusaka are malnourished. A researcher selects a sample of 400 children and finds that 54% of children have a protein deficiency. Test this claim by the researcher at 5% level of significance.

Step 1. Formulation of research and null hypothesis

H_0 : P = .50

H_1 : P $>$ .50

Step 2. Assumptions

i. The children are randomly and independently selected
ii. The population distribution is normal
iii. The level of measurement is interval

Step 3. Decision rules

Given α .05, for a directional hypothesis we require a one tailed test, with the critical value of 1.65.

If $Z_{obs} <$ +1.65 Accept H_0

If $Z_{obs} \geq$ +1.65 Reject H_O

Step 4. Computation

$$Z_{obs} = \frac{P' - P}{SE}$$

P = Hypothesized proportion

SE = Standard error of a proportion

$$SE = \sqrt{\frac{p.q}{n}}$$

q = 1 – p

= 1 - .50

= .50

$$SE = \sqrt{\frac{p.q}{n}}$$

$$= \sqrt{\frac{.50(.50)}{400}}$$

= 0.025

$$Z_{obs} = \frac{P' - P}{SE}$$

$$= \frac{0.54 - 0.50}{0.025}$$

= 1.60

Step 5. Decision

Compare the observed value with a critical value of 1.65. Because Z_{obs} is less than 1.65, we accept the null hypothesis (H_0).

Step 6. Conclusion

The claim by the nutritionist at the Ministry of Health that more than 50% of children in Lusaka are malnourished is not true.

6.3 Testing a Hypothesis concerning differences between Means for Large Sample

Test of difference between two means is used when you want to compare two means reflecting two samples or you may compare two means of two populations and then establish if there is a significant difference between in the mean of female student and male student.

For example, there is a debate going on at the University of Zambia concerning the performance of concerning self sponsored students and those on bursary. Management argues that the performance of self sponsored students and those on bursary is the same while others are saying it is not. As an independent researcher, you are asked to establish management's claim that there is no significant difference in the performance of the two categories of students at 1% level of significance. You sample both sets of students and come up with the following information.

Self Sponsored Students	Sponsored Students
n_1 = 50	n_2 = 50
x_1 = 88.7	x_2 = 83.1
s_1 = 7.8	s_2 = 11.4

Step 1: Formulation of the research and null hypothesis

$H_0 : \mu_1 \neq \mu_2$

$H_1 : \mu_1 = \mu_2$

Step 2: Assumptions

i. Subjects are randomly and independently selected
ii. The groups are independent from one another
iii. The population distribution is normal
iv. The scale of measurement is interval

Step 3: Decision Rules

Given α .01, for a non – directional hypothesis, if -2.58 $< Z_{obs} <$ +2.58, accept H_0

If Z_{obs} < -2.58 or $Z_{obs} \geq$ +2.58, Reject H_0

Step 4: Computation

$$Z_{obs} = \frac{x_1 - x_2}{\sqrt{\frac{s^2 1}{n_1} + \frac{s^2 2}{n_1}}}$$

$$= \frac{88.7 - 83.1}{\sqrt{\frac{7.8^2}{50} + \frac{11.4^2}{50}}}$$

= 2.87

Step 5: Decision

Compare the observed value with the critical value formulated under decision rules. Because Z_{obs} is greater than 2.58, you reject the null hypothesis (H_0).

Step 6: Conclusion

There is no significant difference between the performance of self sponsored and students on bursary. The claim by management is correct.

6.4 Testing Hypothesis concerning differences between Proportions

Test of difference between two proportions is used when you want to compare two two populations and then establish if there is a significant difference between the proportions of the two populations.

For example; A research firm is investing citizen's satisfaction with service delivery by government. The comparison is being made between a rural and urban area.

Rural	Urban
n_1 = 1000	n_2 = 1, 091
$P_1 = \frac{200}{1000}$	$P_2 = \frac{240}{1{,}091}$
= 0.2	= 0.22
q_1 = 1 - p_1	q_2 = 1 - p_2
= 1 – 0.2	= 1 – 0.22
= 0.8	= 0.78

Given this information would you agree that people in urban areas are more satisfied with service delivery by government than those in rural areas? Test the hypothesis using 5% level of significance.

Step 1: Statement of hypothesis

$H_0 : P_1 = P_2$

$H_1 : P_1 < P_2$

Step 2: Assumptions

i. are randomly and independently selected
ii. The groups are independent from one another
iii. The population distribution is normal
iv. The scale of measurement is interval

Step 3: Decision Rules

Give α .05 for a directional hypothesis, if Z_{obs} > - 1.65 Accept H_0,

if Z_{0bs} < -1.65 Reject H_0

Step 4: Computation

$$Z_{obs} = \frac{P_1 - P_2}{\sqrt{\frac{P_1 q_1}{n_1} + \frac{P_1 q_1}{n_1}}}$$

$$= \frac{0.20 - 0.22}{\sqrt{\frac{0.20\ (0.8)}{1000} + \frac{0.22\ (0.78)}{1090}}}$$

$$= -\ 1.12$$

Step 5: Decision

Compare the observed value with the critical value formulated under decision rules. Because Z_{obs} is greater than - 1.65, we accept the null hypothesis (H_0).

Step 6: Conclusion

There is no difference in satisfaction with service delivery by the government between people in rural and urban areas.

7.0Summary

You have been informed in this unit that the purpose of hypothesis testing is to test an assumption that we can make about the population. The procure for hypothesis testing involves a series of steps, theses are; formulation of the research and null hypothesis; choosing the appropriate statistical test and making assumptions; obtaining the appropriate sampling distribution; choosing the significance level and critical regions, computing the test statistic, decision making and; conclusion.

8.0 Self Assessment Questions

After reading through unit seven, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit seven again.

a. The management of Spar Supermarket in Choma is suspicious that the 500 grams of beans

supplied to the store are ***below*** the required weight. To test this suspicion, 100 packages of beans are selected at random, and the average weight of the packages in the sample is 497 grams with a standard deviation of 5 grams. Is the suspicion of management justified? Test at 5% level of significance.

b. A sample of 400 voters in Lubasenshi constituency and 100 voters in Katombora constituency showed that only 33% and 10% of were interested in casting their ballots in the next Presidential, Parliamentary and Local Government elections. At 10% level significance, test the hypothesis that there is ***no*** significant difference in interest in voting between the two constituencies.

UNIT EIGHT: ANALYSIS OF VARIANCE (ANOVA)

1.0 Introduction

This unit discusses the analysis of variance (ANOVA). The unit begins with explaining how the ANOVA is used in hypothesis testing. Thereafter, will describes the steps used when undertaking an analysis of variance. A practical example will be used to explain how the analysis of variance is computed.

2.0 Aim of the Unit

The aim of this unit is to explain the use of ANOVA in hypothesis testing.

3.0 Objectives of the Unit

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

- Distinguish between ANOVA and other methods of hypothesis testing.
- Outline the steps in the computation of the F test statistic
- Test hypotheses using ANOVA.

4.0 Required Material


In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required


It will take you approximately one hour thirty minutes to work through this unit.

6.0 Main Contents

Analysis of variance tests are tests of hypothesis that compare the variation or dispersion between different samples or groups to establish whether there is a significant difference between the samples. This hypothesis test makes use of the F test which is based on the F distributions. To carry out a test, we compute an F test statistic from the data in the samples and compare this with a critical value in an F distribution table. There is a different F distribution table for each level of significance. If the F test exceeds the critical value, we reject the null hypothesis and assume that there is a significant difference between the samples. **This F distribution table is attached as an appendix at the end of this module**.

6.1 The F test Statistic

The F test statistic compares two or more different samples and the variances of the values in those samples.

$$F = \frac{VBG}{VWG}$$

VBG is the abbreviation for variance between groups and VWG is an abbreviation for variance within groups.

For example, suppose you are studying three tutorial groups to establish whether there is a significant difference between in the performance of students in three different tutorial groups. Four (4) random samples are collected from each of the three tutorial groups to establish if there is a significant difference in the performance of students among the three groups using 5% level of significance.

	Tutorial Group 1	Tutorial Group 2	Tutorial Group 3	Total
	80	70	63	213
	92	81	76	249
	87	78	70	235
	83	74	58	215
Total	242	303	267	912

Step 1: Statement of Hypothesis

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

Step 2: Assumptions

i. Subjects are randomly and independently selected
ii. Groups are independent from one another (mutually exlusive)
iii. Population variances are equal
iv. Population distribution is normal with respect to the variable of interest which is performance
v. The scale of measurement is interval

Step 3: Decision rules

ANOVA uses a distribution related to the normal distribution called the F – distribution. In testing hypothesis concerning three means, the computed F – ratio has to be compared to the critical value given in the table. You need the degree of freedom (df) for the two variance estimates of between means and variance within means and also the level of significance.

Degree of freedom between groups (dfb) = j – 1

j stands for the number of groups; therefore dfb = 3 – 1 = 2 df

Degree of freedom within groups (dfw) = n – j

n = total sample size for all the groups; therefore dfw = 12 – 3 = 9 df

To get the critical value of F at 5% level of significance, look for the point of intersection the degree of freedom between means (Degree in freedom in the numerator) and the degree of freedom within means (Degree of freedom in the denominator). Using the F distribution, the F critical value for the degrees of freedom calculated for the data given is 4.26

If F_{obs} < 4.26, Accept H_0

If $F_{obs} \geq 4.26$, Reject H_0

Step 4: Computation

i. Computation of the grand mean ($\bar{\bar{x}}$)

This involves summing all the observations in all the columns and in all the roles and then dividing by the total number of observations.

$\bar{\bar{X}} = \frac{\sum xij}{n}$

$= \frac{912}{12}$

= 76

ii. Computation of the Total Sum of Square (TSS)

Computation of the TSS involves finding the sum of squared deviations of each observation from the grand mean.

SST = $\sum$ (xij – $\bar{\bar{x}}$)²

$\bar{\bar{x}}$ = 76

T.G 1	$(x_i - \bar{\bar{x}})^2$	T.G 2	$(x_i - \bar{\bar{x}})^2$	T.G 3	$(x_i - \bar{\bar{x}})^2$
80	16	70	36	63	169
92	256	81	25	76	0
87	121	78	4	70	36
83	49	74	4	58	324
	442		69		529

SST = 442 + 69 + 529

= 1040

iii. Computation of the sum of squares between groups

This involves first computing means for each group or category and thereafter finding the squared deviation of each category to the grand mean.

	T.G 1	T.G 2	T.G 3
	80	70	63
	92	81	76
	87	78	70
	83	74	58
Total	342	303	267
Mean($\bar{x}_i$) = $\frac{\Sigma x}{n}$	85.5	75.75	66.75

The means as calculated above are T.G 1 = 85.5, T.G 2 = 75.75 and T.G 3 = 66.75

Squared deviations of the category mean to the grand mean ($\bar{x}_1$ - $\bar{\bar{x}}$) for each tutorial will be calculated as follows:

	N	($\bar{x}_i$ - $\bar{\bar{x}}$)	($\bar{x}_i$ - $\bar{\bar{x}}$)²	n($\sum \bar{x}_i$ - $\bar{\bar{x}}$)²
T.G 1	4	9.5	90.25	361
T.G 2	4	-0.25	0.0625	0.25
T.G 3	4	-9.25	85.5625	342.25
				703.5

The calculations above show that 703 is the sum of squares between groups (SSB).

iv. Computation of the variance between groups (VB)

This involves dividing the sum of squares between groups by the degree of freedom between groups,

VB = $\frac{SSB}{dfb}$

Note that the degree of freedom between groups (dfb) was calculated earlier in this section.

VB = $\frac{SSB}{dfb}$

= $\frac{703.5}{2}$

= 351.75

v. Computation of the sum of squares with groups

This involves subtracting the total sum of squares between groups (SSB) from the total sum of squares (TSS)

SSW = TSS – SSB

= 1040 – 703.5

= 336.5

vi. Computation of the variance within groups (VW)

This involves dividing the sum of squares within groups (SSW) by the degree of freedom within groups.

VW = $\frac{SSW}{dfw}$

Note that the degree of freedom within groups was calculated earlier in this section..

VW = $\frac{SSW}{dfw}$

= $\frac{336.5}{9}$

= 37.39

vii. Compute the F statistic

$F_{obs} = \frac{VB}{VW}$

$= \frac{351.75}{37.39}$

$= 9.41$

Step 5: Decision

We will reject the null hypothesis because the F statistic (ratio) is greater than the critical value of 4.26. If $F_{obs} \geq 4.26$, Reject H_0

Step 6: Conclusion

There is a significant difference in the avaerage performance of students is the three tutorial groups.

7.0 Summary

You have been informed in this unit that Analysis of variance tests are tests of hypothesis that compare the variation or dispersion between different samples or groups to establish whether there is a significant difference between the samples. This hypothesis test makes use of the F test which is based on the F distributions. The F test statistic compares two or more different samples and the variances of the values in those samples. F = $\frac{VBG}{VWG}$, where VBG is the abbreviation for variance between groups and VWG is an abbreviation for variance within groups.

8.0 Self Assessment Question

After reading through unit eight, you should answer the question below. It is important for you to attempt the question in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit eight again.

Mr. Kakoma planted four different varieties of beans. Each variety was planted on three separate plots of land and when he harvested, he obtained the following 25 kg bags of beans.

Variety A	60	61	56
Variety B	59	52	51
Variety C	55	55	52
Variety D	58	58	55

Perform an analysis of variance to test at 5% level of significance to establish if there is a significant difference between the samples. (The F distribution table is provided in the appendices).

UNIT NINE: CHI – SQUARE TEST

1.0 Introduction

This unit discusses the chi-square as a test for significance. The unit begins by explaining the applications of *chi-square*. Thereafter, we will describe the steps to be followed when undertaking a *chi–square* test. A practical example will be used to explain how to test for significance using chi – square. The last part of this unit will look at the *chi-square* test of goodness fit.

2.0 Aim of the Unit

The aim of this unit is to explain chi-square in significance testing.

3.0 Objectives of the Unit

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

- Distinguish between *chi-square* and other methods of hypothesis testing.
- Test hypotheses using *chi-square*.

4.0 Required Material

In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

McNabb D. (2009) *Research Methods for Political Science: Quantitative and Qualitative Methods*, New Delhi: PHI Learning Private Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

 It will take you approximately one hour thirty minutes to work through this unit.

6.0 Main Contents

The chi – squared distribution is another distribution used in significance testing. It is often referred to by means of the Greek letter χ^2**.** The chi – squared distribution can be used to test a sample of items in order to decide whether items in a sample are distributed according to a preconceived or expected distribution pattern. Because chi – square tests whether something conforms to a given pattern or not, chi – square tests are often referred to as goodness of fit tests.

The main application of χ^2 test is testing for association or correspondence between attributes. Significance tests on attributes usually take the form of a contingence table. A number of research problems in the social sciences often involve the use of more than one variable. If you have observations taken on more than two or more variables we have what is referred to as a bivariate table and where you have more than two variables, it is referred to as a multivariate table. Bivariate data is often drawn in a two way table with one of the variables along the column and another variable down the rows. Two way tables are called contingency tables. This is because the research hypothesis normally states that two variables are related and this is the same as saying the two tables are contingent.

Example:

A research wants to determine whether there is a relationship between religious affiliation and attitudes towards abortion.

Attitudes towards abortion	Religious Affiliation		
	Protestant	Catholic	Total
For	126 (64%)	99 (38%)	225
Against	71 (36%)	162 (62%)	233
Total	197	261	458

Step 1: Statement of Hypothesis

H_0 : There is no relationship between religious affiliation and attitude towards abortion

H_1 : There is a relationship between religious affiliation and attitude towards abortion

Step 2: Assumptions

i. The subjects for each group are randomly and independently selected
ii. The groups are independent
iii. Each observation qualifies for only one category
iv. The scale of measurement for chi – square can either be nominal or ordinal. In this case, the data is nominal.

Step 3: Decision rules

Given $\alpha = 0.05$ and df = (r – 1) (c – 1)

r = number of row

c = number of columns

df = (2 -1) (2-1) = 1

To find the critical value, go to the ***Critical Values of Chi–Square, the*** **table attached at the module**, find the intersection between the df of 1 (in the first column) and the probability of 0.05 or .05 across the row. The intersection between these two gives you the critical value of 3.84.

If $\chi^2_{0bs} < 3.84$ accept H_0

If $\chi^2_{obs} \geq 3.84$ reject H_0

Step 4: Computation

$\chi^2_{obs} = \sum_{ij} \frac{(O_{ij} - E_{ij})}{E_{ij}}$

O = Observed value

E = Expected value

$E_{ij} = \frac{row\ total\ X\ column\ total}{N}$

O_{IJ}	E_{IJ}	$O_{ij} - E_{ij}$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{IJ}}$
126	96.78	29.22	853.84	8.82
99	128.22	-29.22	853.84	6.66
71	100.22	-29.22	853.84	8.52
162	132.78	29.22	853.84	6.43
			$\chi^2_{obs} =$	30.43

Step 5: Decision

We will reject the null hypothesis because $\chi^2_{obs} > 3.84$.

Step 6: Conclusion

There is a relationship between religious affiliation and attitude towards abortion.

6.1 Chi–Square Goodness of Fit Test

The *chi–square* goodness of fit test is used to determine the extent to which our expectations match with reality.

For example

The Managing Director of Michelo Stores believes that the number of customers who come to the shop from Monday to Saturday is 1,632. However, the cashier disputes this claim. You are hired as a consultant to verify this claim.

Step 1: Statement of Hypothesis

H_0 : Number of customers is evenly spread

H_1 : Number of customers is not evenly spread

Step 2: Assumptions

i. Nominal scale
ii. Sample is large

Step 3: Decision Rules

Given α + 0.05 the df = K – 1= 6-1 = 5df

The critical value is 11.07

If $\chi^2{}_{0obs} < 3.84$ accept H_0

If $\chi^2{}_{obs} \geq 3.84$ reject H_0

Step 4: Computation

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

	O	E	O – E	$(O – E)^2$	$\frac{(O - E)^2}{E}$
Monday	1525	1632	-107	11499	7.02
Tuesday	1711	1632	79	6241	3.82
Wednesday	1655	1632	23	529	0.32
Thursday	1497	1632	-135	18225	11.17
Friday	1603	1632	-29	841	0.52
Saturday	1801	1632	169	28561	17.5
				χ^2_{obs} =	40.35

Step 5: Decision

We will reject the null hypothesis because $\chi^2_{obs} > 3.84$.

Step 6: Conclusion

The number of customers who visit the shop is not evenly spread.

7.0 Summary

You have been informed in this unit that the chi – squared distribution can be used to test a sample of items in order to ascertain whether items in a sample are distributed according to a preconceived or expected distribution pattern. Because chi – square tests whether something conforms to a given pattern or not, chi – square tests are often referred to as goodness of fit tests. The main application of χ^2 test is testing for association or correspondence between attributes.

8.0 Self Assessment Questions

After reading through unit nine, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit nine again.

Suppose a bottling company expects to sale 480 pallets of soft drinks over the next four months, with an equal volume of sales in each month. In the past four months, sales by the company were 150, 110, 100 and 120 pallets. Does the historical data suggest that the demand each month is uniform? Test at 5% level of significance.

UNIT TEN: REGRESSION

1.0 Introduction

This unit discusses regression and correlation. The unit begins with giving an overview of regression and discussing linear regression. Thereafter, we will discuss the formulation of the regression equation. We will then discuss correlation and will later focus on the coefficient of determinant as the last part of this unit.

2.0 Aim of the Unit

The aim of this unit is to discuss regression.

3.0 Objectives of the Unit

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

- Distinguish between regression and correlation.
- Compute the regression equation.
- Calculate the correlation coefficient.
- Explain the coefficient of determinant.

4.0 Required Material

In this unit, you will require the following readings:

Bless C. and Kathuria R. (1993) *Fundamentals of Social Statistics: An African Perspective*, Cape Town: Juta and Co. Ltd

Kenny D. (1987) *Statistics for the Social and Behavioural Sciences*, Canada: Little Brown and Co. Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

 It will take you approximately three hours to work through this unit.

6.0 Main Contents

Regression is a type of analysis which focuses on understanding the relationship or association between two or more variables. It deals with variables that are measured on an interval scale. The relationship or association is shown by two different samples of numbers that are linked together. For example, questions of association might be the relationship between population growth and crime, the association between the time that a car travels and the relationship between years of work experience and income.

If two variables are associated, then as the numbers in one sample vary, their partner numbers in the second sample vary in a related fashion. In statistical analysis we say that the numbers co-vary. The simplest way in which numbers can co-vary is in a linear fashion. A relationship is said to be linear if a change in one variable results in a change in the second variable. The term linear is used when you plot a line on the graph which depicts that the change in X results in a change in Y. Therefore, regression helps to predict the future situation based on the established relationship or association.

6.1 Linear Regression

Linear regression is a way of expressing a relationship between two interval variables using the linear function. Scientists use regression to find some algebraic expression by which to represent formal relationships between variables. The equation $Y = a + bx$ is a linear regression equation meaning that the function describing the relationship between x and y is a straight line. The letter a in the equation is a constant, it shows the value of y when x is 0. The letter b in the equation is the regression coefficient, it shows the changes in y which will be occasioned by a change in x. Ordinarily, researchers display the observations of x and y and the regression line connecting them in form of a graph. The variables x and y are represented by two intersecting axes. X which is the independent variable is also referred to as the prediction or predicator. Y

which is the dependent variable is also referred to as the criterion. The formula for finding the value of a in the regression equation is a = $\bar{Y}$ - b$\bar{x}$ while the formula for b = $\frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$

An example of the regression of x and y is depicted using the data set on time and distance which a car travels as shown in figure 10.1.

Figure 10.1: Time and Distance Travelled by a Car

The regression line does not always pass through the intersection of x and y axes. When a straight line intersects the y axis, a constant needs to be introduced to the regression equation. The constant is represented by the letter a is called the y intercept. The intercept reflects the value of y when x is 0. For example, three different regression line which are considered in figure 10.2, 10.3 and 10.4 have different values for a and b. The three different values of a (6, 1, 2) are reflected in the three different intersections of the graphs being considered (figure 10.2, 10.3 and 10.4). The different values of b (-3, 0.5, 3) reflect the steepness of the slope. The higher the value of b, the steeper the slope. The sign of b expresses the relationship between x and y. When b is positive, an increase in x is accompanied by an increase in y. On the other hand, when b is negative, y decreases as x increase

Figure 10.2: Negative Regression

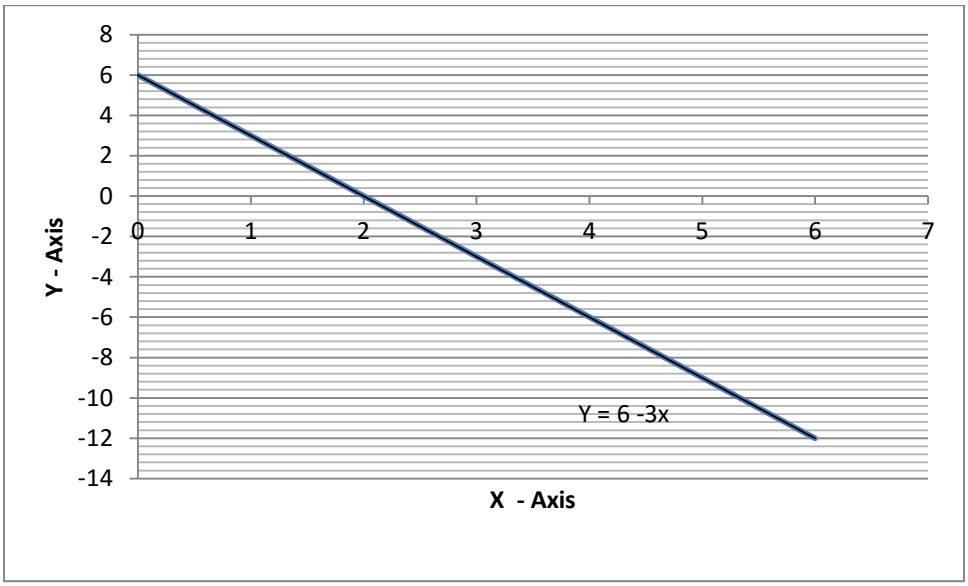


Figure 10.3: Positive Regression

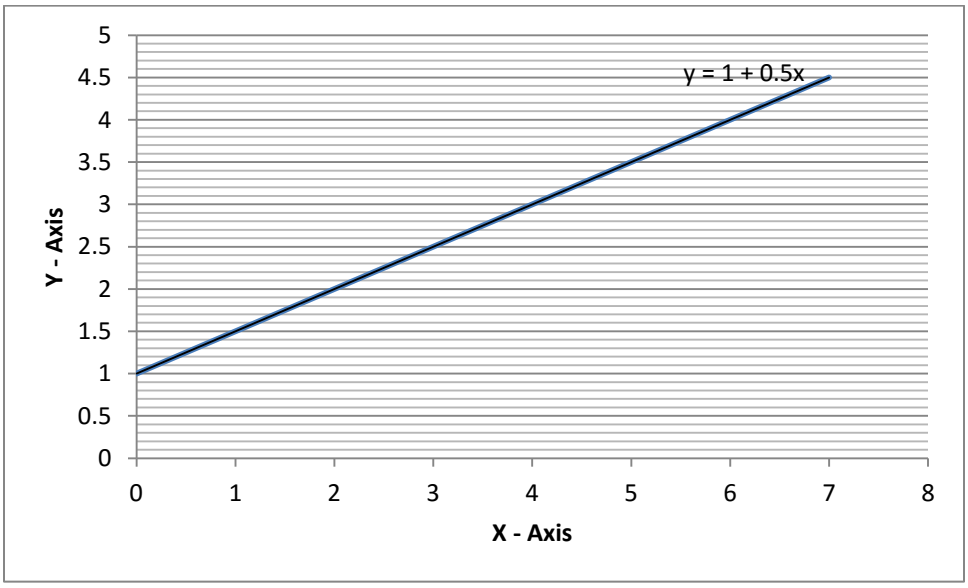
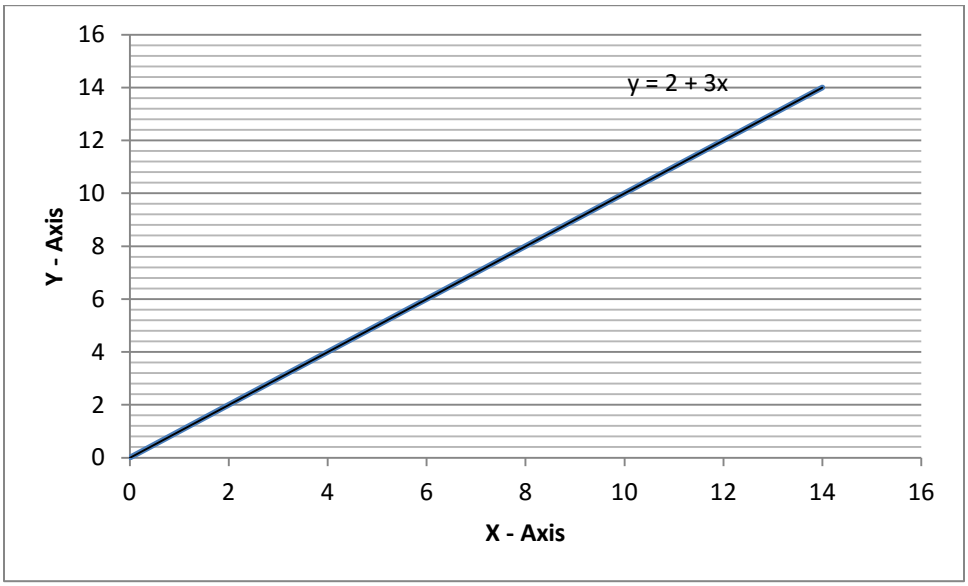


Figure 10.3: Steep Positive Regression



Given the data below on percentage of urban population and robbery rates in the USA, compute the following:

Table 10.1: Percentage of urban population and robbery rates in 10 states in the USA. (Adopted from Nachmias and Nachimias, p. 378)

State	Percentage of urban population (x)	Robberies per 100,000 population (y)	XY	X^2	Y^2
Massachusetts	84.3	150	12,645	7,106.49	22,500
Wisconsin	65.7	105	6,898.5	4, 316.49	11,025
South Dakota	50	26	1,300	2,500	676
Virginia	69.4	132	9,160.8	4, 816.36	17,424
Sacramental	54.6	176	9, 609.6	2, 981.16	30,976
Texas	80.3	180	14, 454	6, 448.09	32,400
Arizona	87.5	174	15, 225	7,656.25	30,276
California	92.6	331	30, 650.6	8,574.76	109,561
Arkansas	53.5	126	6, 741	2,862.25	15,876
Illinois	89	131	11,659	7,921	17,161
Totals	726.9	1,531	118,343.5	55,182.85	287, 875

a. The regression equation

The regression equation is given by$Y = a + bx$. To find the regression equation, we first have to calculate the value of b and a.

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

$$= \frac{10(11\ 345.5) - (726.9)(1531)}{10(551\ 82.85) - (726.9)^2}$$

$= \frac{1\ 183\ 455 - 1\ 112\ 883.9}{551\ 828.5 - 528\ 383.61}$

$= \frac{70\ 551.1}{23\ 444.89}$

= 3.01

a = $\bar{Y}$ – b$\bar{x}$

Therefore, to find a, we should first compute the mean of the y variable. $\bar{Y} = \frac{\sum y}{n} = \frac{1{,}531}{10}$ = 153.1, similary, we should also calculate mean of x. $\bar{X} = \frac{\sum x}{n} = \frac{726.9}{10}$ = 72.69.

a = 153.1 – (3.01) (72.69)

= 153.1 – 218.76

= -65.7

Therefore, the regression equation is:

$\hat{Y}$ = -67.5 + 3.01 (x)

In the above equation, the intercept is -67.5 and the regression coefficient is 3.01.

6.2 Correlation

Correlational analysis aims at measuring the degree of the relationship between two variables and expresses it through a correlational coefficient. The correlational coefficient is given by the formula:

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

Correlation is an extension of regression analysis insofar as regression expresses mathematically the law underlying the relationship between variables whereas correlation aims at measuring the type and strength of the relationship which exists between variables. The other difference between correlation and regression is that the regression coefficient as a measure of association is asymmetrical in the sense that its values depends on which variable is considered

as the criterion (dependent variable) and which one is the predicator (independent variable). It measures the amount of change in the criterion as a function of one unit change in the predicator. However, it is sometimes not possible to specify which variable is the predicator and which variable is the criterion. For example, when measuring the association between a students' performance in chemistry and in biology one variable is not clearly the predicator and the other the criterion. It is therefore desirable to come up with a measure of association that is asymmetrical and expresses the degree of relationship between variables. The correlation coefficient meets both these requirements.

The correlation coefficient is represented by '**r**' and it ranges between -1 and 1. That it -1 is equal to or less than r and r is less than or equal to 1 ($-1 \leq r \leq 1$). On this basis, a researcher can conclude whether the observed relationship is real (significant). If the correlation coefficient (**r**) is close to -1, then the relationship between variables is a strong negative correlation. If the correlation coefficient **r** is close to 1, then the observed relationship is a strong positive correlation. If r is equal to 1, then it is a perfect positive correlation. If r is equal to -1, then it is a perfect negative correlation. If r is equal to 0, then there is absolutely no correlation.

The correlation coefficient for the data on population and crime in the USA considered under regression would be calculated as follows;

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

$$r = \frac{10(118{,}343.5) - (726.9)(1531)}{\sqrt{[10\ (55\ 182.85) - (528\ 383.61)][10\ (287\ 875) - (2\ 343\ 961)]}}$$

$$r = \frac{1\ 183\ 435 - 1\ 112\ 883.9}{\sqrt{[551\ 828.5 - 528\ 383.61][2\ 878\ 750 - 2\ 343\ 961]}}$$

$$r = \frac{70\ 551.1}{\sqrt{[23\ 444.89][534\ 789]}}$$

$$r = \frac{70\ 551.1}{\sqrt{1253806928^{x10}}}$$

$$r = \frac{70\ 551.1}{111973.52}$$

r = 0.63

The correlation coefficient we have computed show that there is a strong positive correlation between the percentage of urban population and crime per 100 000 population in the USA.

6.3 Coefficient of Determinant

Having computed the correlation coefficient, we can also calculate the coefficient of determinant. The coefficient of determinant allows one to determine the proportion of variability of y that can be attributed to x. The coefficient of determinant is given by the formula:

r^2 x 100 and is expressed as a percentage. In the case of the example we have considered, the coefficient of determinant will be:

r^2 x 100

$(0.63)^2$ x100

0.3969 x 100

39.69%

This means that the percentage of urban population accounts for 39.69% of the factors which cause crime in the USA. Other factors, beside percentage of urban population account for 60.31%

7.0 Summary

You have been informed in this unit that regression is a type of analysis which focuses on understanding the relationship between two or more variables which are measured on the ratio or interval scale. Linear regression is a way of expressing the relation between variables using the linear function. You have further been informed that correlation analysis aims at measuring the degree of the relationship between two variables and expresses it through a correlation coefficient. The coefficient of determinant allows one to determine the proportion of variability of Y that can be attributed to X.

8.0 Self Assessment Questions

After reading through unit ten, you should answer the questions below. It is important for you to attempt the questions in order to gauge how much you have comprehended the material presented in this unit. If you are having challenges in answering the question, read unit ten again.

The table below shows the relationship between copper prices and Government revenue in Zambia during a seven year period.

Government Revenue and Copper Price, 1967 - 1973

Year	Copper Price (K/ton) X	Government revenue (K million) Y
1967	810	276
1968	887	306
1969	1, 048	401
1970	1, 011	432
1971	767	309
1972	746	315
1973	1, 156	385

a. Compute the regression equation to show the dependence of government revenue on copper price.
b. Interpret the meaning of the observed regression coefficient within the context of the equation.
c. If government revenue in 1979 was 210 (K million), what was the price of copper?
d. Compute the correlation coefficient and interpret the answer.
e. Compute the coefficient of determinant and interpret the answer

UNIT ELEVEN: THE RESEARCH REPORT

1.0 Introduction

This unit explains the format for PAS 2014 research report. The unit begins by giving an outline of the research report. We will then discuss in detail the each component of the research report.

2.0 Aim of the Unit

The aim of this unit is to introduce you to research report writing.

3.0 Objectives of the Unit

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

- Outline the components of a research report.
- Write your research report for PAS 2014.

4.0 Required Material

 In this unit, you will require the following readings:

McNabb D. (2009) *Research Methods for Political Science: Quantitative and Qualitative Methods*, New Delhi: PHI Learning Private Ltd

Nachmias, C. and Nachmias, D. (1996) *Research Methods in the Social Sciences. 5th ed*, London: St. Martins Press

5.0 Time Required

 It will take you approximately three to work through this unit.

6.0 Main Content

Writing a research report is the last step in the research process. The report informs your readers what you have done, what you have discovered and what conclusions you have drawn from your findings. The research report will have all the sections which were in your proposal; this is because the research report is an extension of the research proposals. You cannot write a valid research report if you did not write the research proposal. Besides having all the sections contained in the proposal, the report will have findings, discussions and conclusions.

6.1 Format of PAS 2014 Research Report

The PAS 2014 report should have the following:

1. **Cover Page**

The cover page should have the title of the research and all the other relevant details pertaining to the institution, school, course, personal details and other relevant information.

2. **Preliminary pages**

The preliminary pages will contain acknowledgements, table of contents, list of abbreviations, list of tables and list of figures.

3. **CHAPTER ONE: INTRODUCTION**

Note that you already have chapter one in your research proposals. All the contents for this chapter will be as they are in your approved (marked) proposals. The only thing you have to do is to change the tense. When writing the contents for this chapter, you should use past tense. The following will be part of the introduction.

- Introduction / Background
- Statement of the Problem
- Study objectives
- Overall objective of the study
- Specific of the study
- Research questions
- Hypothesis to be tested (if applicable)

- Definition of concepts / conceptual framework or theoretical framework
- Literature review

4. CHAPTER TWO: METHODOLOGY

Just like chapter one, you have much of what constitutes chapter two albeit some part should not be included in the report. The report should have the sections listed below but should also be written in past tense. Ensure that all the corrections which you were advised to make in your proposal are fully addressed as you work on your report. The following will be part of the methodology.

- Study design
- Study site and study population
- Sampling size and sampling design
- Data collection instruments
- Data processing procedure
- Problems and Limitations
- Ethical issues involved (if any) and how you propose to deal with the

5. CHAPTER THREE: FINDINGS AND DISCUSSIONS

Findings, discussions and conclusions are what distinguish a proposal from a report. The chapter on findings and discussions should have the following. You write your findings after having analyzed your data. Data analysis entails turning raw data you have collected at the preceding stage in to meaningful information. If your study is purely qualitative, you can write your report based on your field notes from interviews, observations or reviewing of documents by manually analyzing the contents of your notes (contents analysis) or by using computer software NVIVO or Ethnograph. If you are using quantitative data analysis, it is necessary to decide upon the type of analysis required such as frequency tables, cross tabulation and other statistical analysis.

When presenting findings, you can firstly present finding on the characteristics of your respondents such as age, sex and other relevant characteristics. Thereafter, start presenting

findings in line with your research objectives or research questions. This means that you will have sub – headings under this chapter which will conform to your objectives. Having presented findings on your research objectives, you should then moves on to present findings on your research hypotheses.

If you study is quantitative, you should use frequency tables, bar charts, pie charts and cross tabulations to present findings on your objectives. However, this does not mean that you should present a table for every question you asked, you be selective as to what tables to include in the report. The other information can be presented in narrative form. When testing you hypotheses, the scale of measurement for the variables will determine the kind hypothesis test. In a majority of cases, the variables will be nominal and ordinal, this mean chi – square should be used. If you are using SPSS, the p – value which you get after running a chi – square tests determines whether you accepted or reject the null hypothesis. A p – values of ≤ 0.05 means rejection of the null hypothesis and accept the research hypothesis.

For every set of findings which you present, you should come up with a discussion. Discussion of findings entails you as researcher being able to contextualize your findings. Explaining the meaning of your findings in light of the prevailing situation and where possible you can also compare your findings to those of other people who have conducted similar studies. You can also link your findings to the literature which you reviewed and the try to see how your findings fit into your conceptual or theoretical framework.

6. CHAPTER FOUR: CONCLUSIONS

When writing the conclusions, highlight the salient findings. This chapter does not require you to write a lot of things. All you should do is you again follow this sequence of your objectives and at each stage explain what you have concluded. You may cite the page in your report where evidence for arriving at that conclusion can be found. This will mostly be linked to the chapter on findings and discussions.

- Bibliography
- Appendices
- Appendix A: Proposed time frame for the project.
- Appendix B: Questionnaire (s)

- Appendix C: Interview guide (s)

7.0 Summary

In this unit you have been informed that a writing of a research report is the last step in the research process. The report informs your readers what you have done, what you have discovered and what conclusions you have drawn from your findings. You cannot write a valid research report if you did not write the research proposal. Besides having all the sections contained in the proposal, the report will have findings, discussions and conclusions. The research report for PAS 2014 should have four chapters. Chapter one is the introduction, chapter two is methodology, chapter three is findings and discussion while chapter four is the last chapter and it consists of conclusions and recommendations

8.0 Self Assessment Question

After reading through unit eleven, you should write your PAS 2014 research report in line with the guidance provided in this unit.

REFERENCES

Bless C. and Kathuria R. (1993) **Fundamentals of Social Statistics: An African Perspective,** Cape Town: Juta and Co. Ltd

Bowen, B. and Herbert, F. (1980) **An Introduction to Data analysis**, San Francisco: W. H. Freeman and Co. Ltd

Kenny D. (1987) **Statistics for the Social and Behavioural Sciences**, Canada: Little Brown and Co. Ltd

Kotari C. (2004) **Research Methodology: Methods and Techniques,** 2nd Ed. New Delhi: New Age International Publishers

Kumar, R. (2005), **Research Methodology: A Step – by – Step for Beginners**, 2nd Ed. London: Sage Publications

McNabb D. (2009) **Research Methods for Political Science: Quantitative and Qualitative Methods**, New Delhi: PHI Learning Private Limited

Nachmias, C. and Nachmias, D. (1996) **Research Methods in the Social Sciences,** 5th Ed. London: St. Martins Press

O'Sullivan, E. and Rassel, G. (1995) **Research Methods for Public Administration,** 2nd Ed. New York: Longman

Upgrade, V. and Shende, A. (2000) **Research Methodology**, New Delhi: S. Chand and Co.

APPENDIX E

Areas Under the Normal Curve

Fractional parts of the total area (10,000) under the normal curve, corresponding to distances between the mean and ordinates that are Z standard deviation units from the mean.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3718	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4762	4767
2.0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2.9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986
3.0	4986.5	4987	4987	4988	4988	4988	4989	4989	4989	4990
3.1	4990.0	4991	4991	4991	4992	4992	4992	4992	4993	4994
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									

TABLE IV Critical Values of Chi Square

	Level of significance for a non-directional test					
df	.20	.10	.05	.02	.01	.001
1	1.64	2.71	3.84	5.41	6.64	10.83
2	3.22	4.60	5.99	7.82	9.21	13.82
3	4.64	6.25	7.82	9.84	11.34	16.27
4	5.99	7.78	9.49	11.67	13.28	18.46
5	7.29	9.24	11.07	13.39	15.09	20.52
6	8.56	10.64	12.59	15.03	16.81	22.46
7	9.80	12.02	14.07	16.62	18.48	24.32
8	11.03	13.36	15.51	18.17	20.09	26.12
9	12.24	14.68	16.92	19.68	21.67	27.88
10	13.44	15.99	18.31	21.16	23.21	29.59
11	14.63	17.28	19.68	22.62	24.72	31.26
12	15.81	18.55	21.03	24.05	26.22	32.91
13	16.98	19.81	22.36	25.47	27.69	34.53
14	18.15	21.06	23.68	26.87	29.14	36.12
15	19.31	22.31	25.00	28.26	30.58	37.70
16	20.46	23.54	26.30	29.63	32.00	39.29
17	21.62	24.77	27.59	31.00	33.41	40.75
18	22.76	25.99	28.87	32.35	34.80	42.31
19	23.90	27.20	30.14	33.69	36.19	43.82
20	25.04	28.41	31.41	35.02	37.57	45.32
21	26.17	29.62	32.67	36.34	38.93	46.80
22	27.30	30.81	33.92	37.66	40.29	48.27
23	28.43	32.01	35.17	38.97	41.64	49.73
24	29.55	33.20	36.42	40.27	42.98	51.18
25	30.68	34.38	37.65	41.57	44.31	52.62
26	31.80	35.56	38.88	42.86	45.64	54.05
27	32.91	36.74	40.11	44.14	46.96	55.48
28	34.03	37.92	41.34	45.42	48.28	56.89
29	35.14	39.09	42.69	46.69	49.59	58.30
30	36.25	40.26	43.77	47.96	50.89	59.70
32	38.47	42.59	46.19	50.49	53.49	62.49
34	40.68	44.90	48.60	53.00	56.06	65.25
36	42.88	47.21	51.00	55.49	58.62	67.99
38	45.08	49.51	53.38	57.97	61.16	70.70
40	47.27	51.81	55.76	60.44	63.69	73.40
44	51.64	56.37	60.48	65.34	68.71	78.75
48	55.99	60.91	65.17	70.20	73.68	84.04
52	60.33	65.42	69.83	75.02	78.62	89.27
56	64.66	69.92	74.47	79.82	83.51	94.46
60	68.97	74.40	79.08	84.58	88.38	99.61

Find the row corresponding to the indicated degrees of freedom, find the column corresponding to the chosen level of significance, the critical value of χ^2_{crit} is at the intersection of that row and that column. If $\chi^2_{obs} \geq \chi^2_{crit}$ then H_0 is rejected.

TABLE III Critical Values of F (.05 level in roman type, .01 level in **bold face**)

Degrees of freedom for the numerator (columns); Degrees of freedom for the denominator (rows)

	1	2	3	4	5	6	7	8	9	10	11	12
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.36**	19.37 **99.37**	19.38 **99.39**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**
3	10.13 **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**

	14	16	20	24	30	40	50	75	100	200	500	∞	
1	245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,258**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**	1
2	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**	2
3	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**	3
4	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**	4
5	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**	5
6	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**	6
7	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**	7
8	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**	8
9	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**	9
10	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**	10

F_{obs} is found at the intersection of the column corresponding to df_{num} and the row corresponding to df_{den}. The critical values printed in bold are the ones corresponding to the .01 level of significance. If $F_{obs} \geq F_{crit}$ then H_0 is rejected.

TABLE III *(continued)*

Degrees of freedom for the numerator

Degrees of freedom for the denominator