COURSE GUIDE

POL 803
STATISTICAL MODELS AND COMPUTER APPLICATIONS IN POLITICAL SCIENCE

Course Team Stephen Akinyemi Lafenwa. (Course Writer) – University of Ibadan
Professor Osita Agbu (Course Editor) - Baze University, Abuja, Nigeria.
Dr. Matthew Ogwuche (Programme Leader) - NOUN

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone
Nnamdi Azikiwe Expressway
Jabi, Abuja

Lagos Office
14/16 Ahmadu Bello Way
Victoria Island, Lagos

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Printed 2022, 2024

ISBN: 978-978-786-202-5

CONTENTS

COURSE DESCRIPTION

This postgraduate course is designed to give students an in-depth understanding how statistical methods can be applied to investigate and analyze political phenomena. This is a three - unit course sets to discuss the logic and problems of measurement, relevance, scope and application of statistical models. Basic and fundamental skills required in carrying out quantitative analysis particularly, descriptive and inferential statistical analysis will be taught. Students will be exposed to various statistical tools like- measure of central tendency, measure of dispersion, regression, correlation, chi-square, T-test, Z-test, ANOVA etc. Then the course will provide useful information on how to use computer applications like Excel and the Statistical Package for the Social Sciences (SPSS) for data presentation and analysis with computer graphics. One main focus of the course is the key question of how to use statistics and related computer applications to make causal inferences, which are the main goals of most political science research.

INTRODUCTION

The importance of statistical analysis in social science research cannot be over-emphasized. Statistics is fast becoming the language of communication in the behavioural sciences. Yet classroom and extra-classroom experience has shown many social science students, particularly political science students fear and tend to avoid (where possible) statistical courses. These courses are seen by them as difficult and abstract, and has become both a source of nightmares and serious impediment to successful academic work in their university education.

Statistics itself refers to a method that is used for collecting, organizing and analyzing, interpreting numerical data for understanding social phenomenon or making a wise decision.

Although POL 803 is a fundamental course for graduate students and not supposed to be a first course in statistics for any student, this material is prepared with the assumption that students offering this course have little or no knowledge of statistics. Extra effort is taken to explore statistics from the basics. In this material, each study unit is started with the fundamental principles that would enable the student to understand concepts being explained. Therefore, students are advised to start studying this material from the first study unit through the last one. Each study unit is a prerequisite to the next. At the end of this course, you should a good understanding of statistical methods in carrying out scientific inquiry about social and political phenomena.

COURSE AIM AND OBJECTIVES

The general aim of this course is to provide an in-depth analysis of political phenomena using statistical models and computer applications. Step-by-step on how to compute each relevant parameters will be shown to learners.

The specific objectives of the course are to:

(1) Enlighten learners on the significance of statistics as a tool in Political Science research;
(2) Explain to learners some steps to follow in making statistical enquiries of political phenomena;
(3) Educate learners on how to compute basic statistical measures and use these measures to carry out analysis of political phenomena;
(4) Inspire learners to prepare and execute survey research in Political Science;
(5) Enlighten learners on how to use the SPSS computer software to analyze large volume of data;
(6) Explain to learners to understand different ways of interpreting statistical reports more competently and accurately.

WORKING THROUGH THE COURSE

To complete the course, you are required to work through, not only read the study units and other related materials. You will also need to undertake practical exercises for which you need a Scientific Calculator, mathematical set, Statistical Table (Updated version) and a Laptop or Android Phone (with SPSS 23.0 software or a higher version uploaded) and other materials that will be listed in this guide. The exercises are to aid you in understanding the concepts being presented. At the end of each unit, you will be required to submit written assignment for assessment purposes.

At the end of the course, you will be expected to write a final examination.

THE COURSE MATERIAL

In all of the courses, you will find the major components thus:
1) Course Guide
2) Study Units
3) Textbooks
4) Assignments

STUDY UNITS

There are 21 study units in this course. They are:

MODULE 1 INTRODUCTION TO STATISTICS AND SCALES OF MEASUREMENT

Unit 1 Introduction to Statistics: Definitions, Relevance and Scope
Unit 2 Branches of Statistics
Unit 3 The Nexus Between Statistics and Political Science
Unit 4 Statistical Enquiries of Political Phenomenon
Unit 5 Scales of Measurement and Application

MODULE 2 RESEARCH DESIGN AND SURVEY SAMPLING METHODS
Unit 1 Research Design
Unit 2 Meaning and Type of Sampling Method
Unit 3 Bias in Sampling Survey and Sampling Error

MODULE 3 DESCRIPTIVE STATISTICS
Unit1 Data Presentation
Unit 2 Measures of Central Tendency
Unit 3 Describing Data with Averages
Unit 4 Measures of Dispersion or Variability
Unit 5 Describing Variability: Quantitative and Qualitative/Ranked Data

MODULE 4 INFERENTIAL STATISTICS
Unit 1 Correlation Analysis
Unit 2 An Intuitive Approach
Unit 3Regression Analysis

MODULE 5 TESTING HYPOTHESES IN POLITICAL SCIENCE RESEARCH
Unit 1 Meaning and Types of Hypotheses
Unit 2Statistical Tools for Testing Hypothesis in Political Science Research: T test
Unit 3 Statistical Tools for Testing Hypothesis in Political Science Research: Chi-Square Test
Unit 4 Elementary Probability Theory
Unit 5Computer Application in Data Analysis

As you can observe, the course begins with the basics and expands into a more elaborate, complex and detailed form. All you need to do is to follow the instructions as provided in each unit. In addition, some self-assessment exercises have been provided with which you can test your progress with the text and determine if your study is fulfilling the stated objectives.

TEXTBOOKS AND REFERENCES

At the end of each study unit, you will find a list of relevant reference materials which you may yourself wish to consult as the need arises, even though I have made efforts to provide you with the most important information you need to pass this course. However, I would encourage you, as a postgraduate student to cultivate the habit of consulting as many relevant materials as you are able to within the time available to you. In particular, be sure to consult whatever material you are advised to consult before attempting any exercise.

COURSE OVERVIEW PRESENTATION SCHEME

There are 21 units in this course. You are to spend one week on most of the units. One of the advantages of Open and Distance Learning (ODL) is that you can read and work through the designed course materials at your own pace, and at your own convenience. The course material replaces the lecturer that stands before you physically in the classroom.
All the units have similar features. Each unit begins with the introduction and ends with reference/suggestions for further readings.

Units	**Title of Work**	**Week Activity**	**Assignment (End-of-Unit)**
Course Guide			
Module 1	**Introduction to Statistics and Scales of Measurement**		
Unit 1	Introduction to Statistics: Definitions, Relevance and Scope	Week 1	Assignment 1
Unit 2	Branches of Statistics	Week 2	Assignment 1
Unit 3	The Nexus Between Statistics and Political Science	Week 2	Assignment 1
Unit 4	Statistical Enquiries of Political Phenomenon	Week 3	Assignment 1
Unit 5	Scales of Measurement and Application	Week 4	Assignment 1
Module 2	**Research Design and Survey Sampling Methods**		
Unit 1	Research Design	Week 5	Assignment 1
Unit 2	Meaning and Type of Sampling Method	Week 6	Assignment 1
Unit 3	Bias in Sampling Survey and Sampling Error	Week 7	

Module 3	**Descriptive Statistics**		
Unit 1	Data Presentation	Week 8	Assignment 1
Unit 2	Measures of Central Tendency	Week 9	Assignment 1
Unit 3	Describing Data with Averages	Week 9	Assignment 1
Unit 4	Measures of Dispersion or Variability	Week 10	Assignment 1
Unit 5	Describing Variability: Quantitative and Qualitative /Ranked Data	Week 10	
Module 4	**Inferential Statistics**		
Unit 1	Correlation Analysis	Week 11	Assignment 1
Unit 2	An Intuitive Approach	Week 11	Assignment 1
Unit 3	Regression Analysis	Week 12	Assignment 1
Module 5	**Testing Hypotheses and Computer Applications in Political Science Research**		
Unit 1	Meaning and Types of Hypotheses	Week 13	Assignment 1
Unit 2	Statistical Tools for Testing Hypothesis in Political Science Research: T test	Week 14	Assignment 1
Unit 3	Statistical Tools for Testing Hypothesis in Political Science Research: Chi-Square Test	Week 15	Assignment 1
Unit 4	Elementary Probability Theory	Week 16	Assignment 1
Unit 5	Computer Application in Data Analysis	Week 16	Assignment 1
	Revision	Week 17	
	Examination	Week 18	
	Total	18 weeks	

WHAT YOU WILL NEED IN THE COURSE

There will be some recommended texts at the end of each study unit that you are expected to purchase. Some of these texts will be available to you in libraries across the country. In addition, your computer proficiency skill will be useful to you in accessing internet materials that pertain to this course. It is crucial that you create time to study these texts diligently and religiously.

TUTORS AND TUTORIALS

The course provides fifteen (15) hours of tutorials in support of the course. You will be notified of the dates and locations of these tutorials, together with the name and phone number of your tutor as soon as you are allocated a tutorial group. Your tutor will mark and comment on your assignments, and watch you as you progress in the course. Send in your tutor-marked assignments promptly, and ensure you contact your tutor on any difficulty with your self-assessment exercise, tutor-marked assignment, and the grading of an assignment. Kindly note that your attendance and contributions to discussions as well as sample questions are to be taken seriously by you as they will aid your overall performance in the course.

ASSESSMENT EXERCISES

There are two aspects to the assessment of this course. First is the Tutor-Marked Assignments; second is a written examination. In handling these assignments, you are expected to apply the information, knowledge and experience acquired during the course. The tutor-marked assignments are now being done online. Ensure that you register all your courses so that you can have easy access to the online assignments. Your score in the online assignments will account for 30 per cent of your total coursework. At the end of the course, you will need to sit for a final examination. This examination will account for the other 70 per cent of your total course mark.

TUTOR-MARKED ASSIGNMENTS (TMAs)

Usually, there are four online tutor-marked assignments in this course. Each assignment will be marked over ten percent. The best three (that is the highest three of the 10 marks) will be counted. This implies that the total mark for the best three assignments will constitute 30% of your total course work. You will be able to complete your online assignments successfully from the information and materials contained in your references, reading and study units.

FINAL EXAMINATION AND GRADING

The final examination for POL 803: Statistical Models and Computer Applications in Political Science will be of three hours duration and have a value of 70% of the total course grade. The examination will consist of multiple choice and fill-in-the-gaps questions which will reflect the practice exercises and tutor-marked assignments you have previously encountered. All areas of the course will be assessed. It is important that you use adequate time to revise the entire course. You may find it useful to review your tutor-marked assignments before the examination. The final examination covers information from all aspects of the course.

HOW TO GET THE MOST FROM THIS COURSE

1. There are 21 study units in this course. You are to spend one week in most of the units. In distance learning, the study units replace the university lectures. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace, and at a time and place that suites you best. Think of it as reading the lecture instead of listening to the lecturer. In the same way a lecturer might give you some reading to do. The study units tell you when to read and which are your text materials or recommended books. You are provided exercises to do at appropriate points, just as a lecturer might give you in a class exercise.
2. Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit, and how a particular unit is integrated with other units and the course as a whole. Next to this is a set of learning objectives. These objectives let you know what you should be able to do, by the time you have completed the unit. These learning objectives are meant to guide your study. The moment a unit is finished, you must go back and check whether you have achieved the objectives. If this is made a habit, then you will significantly improve your chance of passing the course.
3. The main body of the unit guides you through the required reading from other sources. This will usually be either from your reference or from a reading section.
4. The following is a practical strategy for working through the course. If you run into any trouble, telephone your tutor or visit the study centre nearest to you. Remember that your tutor’s job is to help you. When you need assistance, do not hesitate to call and ask your tutor to provide it.
5. Read this course guide thoroughly. It is your first assignment.

6. Organise a study schedule - Design a 'Course Overview' to guide you through the course. Note the time you are expected to spend on each unit and how the assignments relate to the units.
7. Important information; e.g. details of your tutorials and the date of the first day of the semester is available at the study centre.
8. You need to gather all the information into one place, such as your diary or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates and schedule of work for each unit.
9. Once you have created your own study schedule, do everything to stay faithful to it.
10. The major reason that students fail is that they get behind in their coursework. If you get into difficulties with your schedule, please let your tutor or course coordinator know before it is too late for help.
11. Turn to Unit 1, and read the introduction and the objectives for the unit.
12. Assemble the study materials. You will need your references for the unit you are studying at any point in time.
13. As you work through the unit, you will know what sources to consult for further information.
14. Visit your study centre whenever you need up-to-date information.
15. Well before the relevant online TMA due dates, visit your study centre for relevant information and updates. Keep in mind that you will learn a lot by doing the assignment carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the examination.
16. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study materials or consult your tutor. When you are confident that you have achieved a unit's objectives, you can start on the next unit. Proceed unit by unit through the course and try to space your study so that you can keep yourself on schedule.
17. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in the course guide).

CONCLUSION

This is a practical cum theoretical as well as empirical course and so, you will get the best out of it if you can read wide, study the procedures and formulas where necessary and always do exercises relevant to each study unit. You can watch the videos on YouTube or other media platforms on step -by -step on how to use computer applications in analyzing data relevant for political analysis.

SUMMARY

This Course Guide has been designed to furnish you with the information you need for a fruitful experience in the course. In the final analysis, how much you get from it depends on how much you put into it in terms of learning time, effort and planning.

I wish you all the best in POL 803 and in the entire programme!

REFERENCES FOR FURTHER READINGS

Adeniyi Gbadegesin, Razaq Oloopoenia and Afeikhena Jerome, (2005), *Statistics for Social Sciences*, Ibadan:IUP

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers, Pvt Limited.

Akintunde Elijah, (2017) Introduction to Statistics and SPSS, a manual prepared for Politica Science students.

Champney, L. (1995). *Introduction to Quantitative Political Science*. New York: Harper Collins College Publishers.

Clegg, F. (1990) *Simple Statistics: A Course Book for the Social Sciences* (CUP).

Freedman, David, Robert Pisani, and Roger purves. (2007) *Statistics* 4th eds. Norton

Gupta, C. B. (1982*), An Introduction to Statistical Methods*, Delhi: Vikas Publishing House Limited

Harry Frank, and S.C. Althoen (1994) *STATISTICS: Concepts and Applications.* (Cambridge) Cambridge University Press.

Johnson J. B. and Joslyn, R. A. , 1991, "Political Science Research Methods", Washington D.C. *Congressional Quarterly* Inc. Chapters 1 -3

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

Kitchens, L.J. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference* (2nd Ed.). USA: Duxbury Press.

Nwolise O.B.C, (1987) "Factors explaining high Defence expenditures in Africa 1967 – 1977,"Ph.D dissertation, Department of political science, University of Ibadan

Nwolise O.B.C., 2005, "Measurement" in Adeniyi Gbadegesin, Razaq Oloopoenia and Afeikhena Jerome, (2005), *Statistics for Social Sciences*, Ibadan:IUP

Obasi, I.N. (1999). *Research Methodology in Political Science*. Enugu: Academic Publishing Company

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Olu Ojo (2005), *Fundamentals of Research Methods*, Ibadan: Standard Publications

Ott, Lyman et al (1983). *Statistics: A Tool for the Social Sciences* (3rd Ed.). Boston, Massachusetts: Duxbury Press.

Robert S. Witte and John S. Witte, 2017, *Statistics* (11th Edition), NJ: John Wiley and Son Inc.

Spiegel, M.R. and L.J. Stephens. (2000) *Introduction to Probability and Statistics Schaum's Outline Series* (3rd Ed.). New York: McGraw-Hill.

Witte Roberts S. and John S. Witte, 2017, Statistics, (Eleventh ed.) Hoboken, NJ: John Wiley & Sons, Inc.,

Wright, D.B. (1997) *Understanding Statistics: An Introduction for the Social Sciences* (Sage).

Any Statistical Textbooks for Political/Social Scientists.
All Goggle Materials on Statistics

WEBSITE LINKS FOR FURTHER READING
http://www.le.ac.uk/bl/gat/virtualfc/Stats/descrip.html

https://med.und.edu/daccota/_files/pdfs/berdc_resource_pdfs/data_analysis_using_spss.pdf

Arithmetic Notations and their Meanings

You should study the following signs carefully and understand their usage. You will come across them in the later part of this course.

1. Equality sign is denoted by =
   It shows that a quantity written on its left side is same as the quantity on its right-hand side. e.g. 10 x 2 = 20
2. The decimal point is denoted by ·
   It is used to indicate the fraction part of a figure. e.g. 15·33 means 15 is a whole number but .33 is the fractional part of that quantity. In fact, the same quantity could be written as 15⅓ without a decimal point.
3. The addition sign is denoted by +
   When placed between two numbers, the two numbers are to be added
   e.g. 11 + 3 = 14
   Another word for addition is sum.
4. Subtraction sign is denoted by -
   When placed between two numbers, the second number is to be subtracted or removed from the first.
   e.g. 20 – 8 = 12
   Another word for subtraction is difference.
5. Multiplication is denoted by x
   When placed between two numbers, one of the numbers must be multiplied by the other.
   e.g. 3 x 5 = 15
   Another word for multiplication is product.
6. Division sign is denoted by ÷
   When placed between two numbers, use the last number to divide the first or the number below to divide the one on top of the slash (/)
   e.g. 14 ÷ 2 = 7

   or $\frac{12}{4} = 3$
7. Greater than sign denoted by >
   When placed between two numbers, it means the first number has a bigger magnitude than the second number.
   e.g. 25 > 12
8. Less than sign is denoted by <
   When placed between two numbers, it means the first number is of smaller magnitude than the second number.
   e.g. 16 < 20
9. Greater than or equal to denoted by ≥ implies the first number may be larger than the second or may have same magnitude as the second number.
   e.g. $20 \geq 2x$
10. Less than or equal to sign denoted by ≤

Means the first number may be smaller or have same magnitude as the second number.
e.g. $5 \leq 2x$

11. Not equal to sign denoted by $\neq$
Means the figure on the left and the figure on the right are not of the same magnitude. E.g. $10 \neq 3 \times 4$

12. Bracket sign denoted by ( )
It may be used to separate single arithmetic operations to be performed.
e.g. $(2 \times 3) + (10 \div 5)$
Or when a figure is placed behind it and there is another figure inside it, then it means multiplication e.g. $2 (4) = 2 \times 4 = 8$

13. Therefore sign denoted by
Used to show logical continuation of a step from the previous one.
e.g. If $2x = 10$

then $\therefore$

$$2x = \frac{10}{2}$$

$$\therefore x = 5$$

14. For example sign is denoted by e.g.
It means for instance or as an illustration.

15. Square root sign is denoted by $\sqrt{}$
When placed over a number, it means we should look for a smaller number which when multiplied by itself will give exactly the number under the square root sign.
e.g. $\sqrt{25} = \sqrt{5x5} = 5$

16. Summation sign is denoted by $\sum$
It means add all figures together from the first to the last.

17. Square sign is denoted by X^2
Means multiply the figure by itself twice.
e.g. $5^2 = 25$

MAIN COURSE

CONTENT **PAGE**

MODULE 1 INTRODUCTION TO STATISTICS AND SCALES OF MEASUREMENT

Most of students in the field of political science do not understand what statistics entails. This lecture will examine the definitions of Statistics, its scope and limitations, and its importance to political enquiry. We will also explore the branches of Statistics and their inter-relationship, and then go ahead to define and explain some basic concepts in statistics. Also, the link between statistics and political science will be explained in terms of relevance of statistics to the political science as a discipline. More importantly, the issue of measurement is also dealt with in this module. Therefore, this module is thematically structured into five units that comprehensively present vital details that will clear your doubt and help you.

Unit 1 Introduction to Statistics: Definitions, Relevance and Scope
Unit 2 Branches of Statistics
Unit 3 The Nexus between Statistics and Political Science
Unit 3 Statistical Enquiries of Political Phenomenon
Unit 4 Scales of Measurement and Application

You are advice to study each of the unit carefully as you are expected to answer some questions to evaluate your understanding on the various issues as discussed. Possible answers to the questions are provided under each of the unit accordingly.

UNIT 1 INTRODUCTION TO STATISTICS: DEFINITIONS, RELEVANCE AND SCOPE

Unit Structure

1.1 Introduction

This unit examines the various definitions of Statistics as it relates to political science as a discipline and other disciplines in the social sciences. The unit exposes you to the various reasons why statistical analysis is significant to political scientists who are preoccupied with scientific and systematic study of politics. More importantly, the scope of statistics in political analysis will also be highlighted and discussed. This introductory aspect of this course in this unit, is critical to your understanding of statistical tools and their application to analysis of political phenomena.

1.2 Learning Outcomes

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

- give at least two definitions of statistics by identifying its main features.
- explain the relevance of statistical analysis in the study of politics
- discuss the scope of statistics in political science

1.3 What is Statistics?

Statistics, in a simple sense means a collection of information shown in numbers. It is defined as science of collecting and analyzing data. More broadly, it is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

Statistics can also be defined as the scientific study of handling quantitative information. It embodies a methodology of collection, classification, description and interpretation of data obtained through the conduct of surveys and experiments. You can also define statistics as a quantitative method of analyzing data obtained through the conduct of surveys and experiments.

It is one of the scientific methods - others include experimental method, comparative method and case study method) for conducting political inquiry and analysis. It is an analytical tool used in solving problems and in decision making in areas such as Natural and Biological Sciences,

Engineering, Economics, Management, Marketing, Political analysis and other problem-oriented disciplines.

It is important for you to note that the main aim of statistics in political science is to allow for scientific explanation and prediction. It helps in making decisions vital for public policies and programs. Basically speaking, I want you to know that you can use statistics to do the following:

(i) Collection of data:
Statistics involves collection or gathering of data from the field, particularly through survey research. Data collected may be from primary sources or secondary sources.

(ii) Presentation of data:
Data collected using statistical methods can be presented using statistical tools like pictogram, graphs, frequency tables etc.

(iii) Analysis of data:
Data presented can also be analyzed using descriptive or inferential analytical methods depending on your major interest or focus in the application of statistics to analyze political phenomenon.

(iv) Interpretation of data
There are procedures in statistics that will allow you to interpret your data precisely so as to reach logical conclusion; from which you can make intelligent decisions.

If I may ask you, do you know what statistical data are? Statistical data refer to numerical descriptions of quantitative aspects of things. This description may take the form of counts or measurements. For instance, statistical data on election may include:

- Number of polling units in a particular local government.
- Number of male or female legislators in the National Assembly.
- Number of single or married victims of political assassinations
- Number of presidential aspirants in a general election.
- Other variables like the experience, education, income, age or political affiliation of candidates etc. I want you to think and list in your exercise book more examples of statistical data for political analysis.

We can classify Statistical Data into different groups namely:

1. Discrete and Continuous Data
2. Primary and Secondary Data
3. Time Series and Cross-Sectional Data

Discrete Data: This is described as data collected on variables that can be measured precisely e.g. the number of voters in an election, the number of bills passed by a state legislative house in a given year, the number of unemployed youth in a country.

Continuous data: these are data collected on variables whose values cannot be measured precisely. Their values can only be approximated or pinned to an interval. e.g. The distance between Lagos and London could be said to be 5,0007 km ± 1 km. This is not a precise figure but an interval.

Primary Data: This is the data used for the specific purpose for which it was collected. Its source includes census and samples.

Secondary Data: These are data that are being used for some purposes other than that which they were originally collected. Examples are data collected from Federal Office of Statistics (F.O.S.), World Bank, UNESCO, INEC, Local Government, Military establishments among others and used by a researcher either in the University of Ibadan or elsewhere.

Time Series Data: These are sets of observations taken sequentially in time, usually at equal intervals e.g. Government bank deposits are often arranged in accordance with the time the deposits were made.

Cross Sectional Data**:** These are data recorded or collected at a specific time, usually for a day, a week or a year. The examples include among others the annual spending of political parties in Nigeria and the weekly contributions of a campaign team.

1.4 The Significance of Statistics in Political Science

In the field of social sciences, statistical methods refer to a body of methods that are used for collecting, organizing and analyzing numerical data for understanding social phenomenon or making a wise decision. It is imperative to point out that the importance of statistical analysis in social science research cannot be over-emphasized. Statistics as a tool is fast becoming the language of communication in the behavioral sciences – psychology, sociology, political science and the like. For effective communication in research, statistical thinking is as necessary as engine in a motor car.

Specifically, the following are major reasons why statistics is important to you as a political science researcher:

1. Statistics enables you to know how to evaluate published political and other data, when to believe them, when to be skeptical or cynical and when to reject them. By asking questions on the source of statistical data, how and when they are collected, you should be able to determine the validity of the data and what to use them for.
2. It enables you to meet up with the requirement of interpreting the results of sampling (surveys or experimentation), employing statistical methods of analysis to make inferences, describing the characteristics of a group, interpreting a computer printout containing statistical information, or writing up a report based on statistical analysis. As you we see later, statistics help political scientists to describe political phenomenon and draw inferences from it using statistical tools.
3. It enables you to gather facts, test hypothesis and develop theories. By relying on numerical data gathered through survey or experimentation, we can test our hypotheses or assumptions for example using T test, F test, Z test, Chi -Square test and so on. From hypothesis testing and interpretation of results, we can draw our conclusion.
4. Statistics further enable you to draw generalizations and make predictions as necessary. In other words, we can use statistics to explain and predict political phenomena.

According to Robert W. Bugess cited in Atoyebi (2003):
"The fundamental gospel of Statistics is to push back the domain of ignorance, prejudice, rule of thumb, arbitrary or premature decisions, traditions and dogmatism, and to increase the domain in which decisions are made and principles are formulated on the basis of analyzed quantitative facts".
From the above quotation, you should take note of the significance of statistics in decision making and in establishing the truth

Self -Assessment Exercises (SAEs) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 3 minutes.

1. Statistics is concerned with:
(a) calculation of mean, median and mode.
(b) computation of regression and standard deviation.
(c) collection, organization, presentation and analysis of data.
(d) none of the above.
2. Some of the uses of Statistics include the following except:
(a) computation and interpretation of difficult mathematical questions.
(b) presenting facts in definite form.
(c) Drawing of generalization and prediction.
(d) Evaluate and simplifies unwieldy and complex mass of political data.
3. …….. is the data used for the specific purpose for which it was collected.
4. Data collected on variables that can be measured precisely are known as …….

1.5 The Scope of Statistics

1.5.1 Scope

In terms of the scope, statistical methods can be meaningfully applied to understand and interpret any phenomenon, particularly in the field of social sciences, which satisfies these two major conditions:

1. *Any political phenomenon that is capable of being quantified* (i.e. expressed in the form of figures or numbers) either through a process of counting or measurement. For examples how many women registered to vote in the general elections, voter-turnout in Oyo state during the 2015 presidential elections, what is the size of the GDP compare to number of people living in abject poverty etc.
2. *Any political phenomenon that is affected by a multiplicity of causes*. That is to say, the changes that are brought about in the characteristic (of the phenomenon) under study are caused by a number of forces acting simultaneously. In the field of social sciences in general and specifically in political science discipline, most if not all variables are multi-causal and not mono-causal; that is there are more than one factor or variable for explanation. For instance, Why do people participate in politics? How to win elections? What makes a good leader?

Generally speaking, Statistics is applicable in all fields where empirical data (evidence) can be measured and collected.
There are key concepts that are central to the scope of statistical analysis in political science research. They include:

1.5.2 Population and Sample

A Population is the entire set of existing units being considered within a specific study (usually people, objects, transactions or events). Examples of population includes:

- all legislators in Nigeria.
- all registered voters in North west geo-political zone in Nigeria.
- all APC delegates in 2022 presidential party primaries.
- all political appointees in President Muhammadu Buhari's administration.

The list is exhaustive. You should think of more examples of population and add them to the four above.

A population can be finite or infinite. A finite population is that whose figure is ascertainable (i.e. a population that is numerically countable). While an infinite population is that whose figure cannot be ascertained (not countable). Examples of a finite population are the four listed above, while example of an infinite population is the population of all fishes in the Atlantic Ocean and the population of all corrupt politicians in Nigeria.

A ***Concept*** is an abstraction based on characteristics of perceived reality. It is a word or general notion that expresses generalizations from particulars. An example of a concept would be 'weight' that expresses numerous observations to the extent to which things are more or less heavy, just as the concept of security expresses observations about the extent of safety and freedom from danger and anxiety.

Variables

In studying a population, we focus on one or more characteristics or properties of the units in the population. For example, we may be interested in the income, gender, age, and occupation of a population of registered voters in the 2019 general elections in Nigeria. We call such characteristics variables. The opposite of a variable is the term ***constant*** which signifies that member of the unit or group have same magnitude of the characteristic of interest e.g., the number of hours in a day is constant, it is always 24 hours. Also, the number of days in a week is constant; it is 7.

Sample

A sample is a representative subset of the population. It is a small part of the whole population selected for the purpose of a study. Whatever deductions made out of the sample can be used to generalize for the entire population provided the sampling technique used in not biased.

For instance, in order to study the voting pattern of market women in Jabi Market in Abuja (consisting of about 3,000 women), I may carefully select a sample of 500 women, study their voting pattern and then use my findings to generalize for the entire population of Jabi market women.

One important question may cross your mind: why do we need to select only 500 market women, why not study the entire 3,000 market women? Well, it is a good question but go to the next section to get the answer.

1.5.3 Importance of Drawing a Sample

1. It is economical because it is cheaper to work with a sample when the population is large. Assuming the population of market women in Nigeria is 2,500,000 or more, can we interview all of them?
2. It saves time and energy.
3. Where a population is partly inaccessible, drawing a sample will be more appropriate.
4. Where members of the population can easily be destroyed while studying them, samples are more appropriate.

A **Subject** is a single member of a sample. For example, the 2003 general election is a subject in the sample of general elections conducted so far in the Nigeria's fourth republic drawn from the population of all elections conducted in Post -1999 transition to democracy.

Self -Assessment Exercises (SAEs) 2

Attempt these exercises to assess what you have learnt so far. This should not take you more than 12 minutes

1. The relationship between population and sample is that

(a) Population is a census while sample is also a census.

(b) Sample is a subset of population and can be used to generalize about a population.

(c) Both concepts deal with human beings only.

(d) Both concepts are unrelated.

2. ……. is the opposite of variable

3. Which of the following is not a variable?

(a) Income

(b) Gender

(c) Furniture

(d) Speedometer of Buhari's presidential car

(e) Campaign expenditure

4. The Scope of Statistics in political science research covers any political phenomenon that is capable of being quantified only. TRUE OR FALSE

5. Statistics helps political scientists to make good decisions for proper planning and execution of public policies TRUE OR FALSE

1.6 Summary

Statistics as a method is different from experimental, comparative and case study methods in political science. It is a tool used in gathering, presenting, describing and analyzing data as well as drawing inferences from analyzed data. Generally, statistics is useful for scientific explanation and prediction of political phenomenon. Statistical data may be classified as discreet or continuous, primary or secondary, time series or cross sectional. Samples are drawn from population in order to minimize cost and to save time.

Statistics have many advantages for the study and practice of politics. We can summarize these advantages simply by noting that statistics offer insight into issues and problems in a field that would otherwise go unnoticed and unheeded. In short, applying statistics in the field of political science enables us to make good political decisions based on quantifiable, measurable and accurate fact or information that we can use to predict and plan for the future.

The scope of statistics in investigating political phenomena covers those phenomena that can be quantified or measured on the one hand, and on the other hand, those phenomena that are affected by multiplicity of causes. The use of Statistics makes Social Science research more appealing and research outcomes are more clearly presented.

1.7 References/Further Reading

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Adeniyi Gbadegesin, Razaq Oloopoenia and Afeikhena Jerome, (2005), *Statistics for Social Sciences*, Ibadan:IUP

1.8 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. c
2. a
3. Primary data
4. Discrete data

Answers to SAEs 2

1. b
2. Constant
3. c
4. False
5. True

UNIT 2 BRANCHES OF STATISTICS

Unit Structure

2.1 Introduction

After defining what statistics entails and its scope, it is important to highlight and explain the major branches of statistics that we can explore in our analysis of political phenomena. When analyzing data, for example, the total number of votes obtained by 50 candidates in an election, it is possible to use both descriptive and inferential statistics in your analysis of their votes. Typically, in most research conducted on groups of people, you will use both descriptive and inferential statistics to analyze your results and draw conclusions. I want to inform you that there are two main branches of Statistics; namely Descriptive Statistics and Inferential Statistics.

2.2 Learning Outcomes

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

(a) Identify the two major branches of statistics.
(b) Describe and give examples of Descriptive statistics.
(c) Discuss with examples what Inferential statistics entails.

2.3 Descriptive Statistics

Descriptive statistics is the analysis of data that helps describe, display or summarize data in a meaningful way such that, for example, patterns might develop from the data. It utilizes numerical and graphical methods (such as Bar chart, Pie chart, Histogram etc.) to look for patterns, summarize and present the information in a set of data.

Descriptive or Deductive Statistics can be described as the phase of Statistics which seeks only to describe and analyze a given group without drawing any conclusion or inferences about a larger group. Meanwhile, descriptive statistics do not allow us to make conclusions beyond the data we have analyzed or draw conclusions regarding any hypotheses we might have made. In a nut shell, with descriptive statistics you are simply describing what is, what the data shows.

In essence, therefore, descriptive statistics allows you to present the data in a more meaningful way, which allows simpler interpretation of the data. For instance, if we had the election results of 100 independent candidates, you may be interested in the overall performance of those candidates. You might also be interested in the distribution or spread of the votes. This type of statistics allows you to do this. You should take note that there are two general types of statistic that you can use to describe your data.

The first one is called measures of central tendency (details later). These are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of votes obtained by the 50 candidates from the lowest to the highest. You can describe this central position using a number of statistics, including the mean, median, and mode.

The second category of descriptive statistics is what we known as measures of dispersion or measures of spread. These measures involve ways of summarizing a group of data by describing how disperse or spread out the scores is. Take for example; the mean vote of the 100 candidates may be 72,000. However, you should take note that not all candidates will have polled 72,000 votes. Relatively, their votes will be spread out. Some will be lower and others higher. Measures of dispersion help us to summarize how spread out these scores is.

It is important to inform you that to describe this spread, a number of statistics are available to us, including the range, quartiles, absolute or mean deviation, variance and standard deviation. The computational procedures of each of these measures will be explained in details later in this course.

2.4 Inferential Statistics

Inferential Statistics on the other hand, is concerned with studying a small portion of a group called a sample and making valid conclusion about the larger population based on the sample characteristics. The phase of Statistics which deals with conditions under which inferences is valid is called Inductive or Inferential Statistics.

We have seen that descriptive statistics provide information about our immediate group of data. Descriptive statistics are applied to populations, and the properties of populations, like the mean or standard deviation, are called 'parameter**s'** as they represent the whole population (i.e., everybody you are interested in).

Often, however, you do not have access to the whole population you are interested in investigating, but only a limited number of data instead. For example, you might be interested in the exam marks of all students in Nigeria. It is not feasible to measure all examination marks of all students in the whole of Nigeria so you have to measure a smaller 'sample' of students (e.g., 100 students), which are used to represent the larger population of all Nigerian students. Properties of samples, such as the mean or standard deviation, are not called parameters, but 'statistic'.

Inferential statistics are techniques that allow you to use these samples to make generalizations about the populations from which the samples were drawn. Conclusions from inferential statistics extend beyond the immediate data alone. For example, we use inferential statistics to try to infer from the sample data what the population thinks. Or, we use inferential statistics to make judgements of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in the study.

It is, therefore, important that the sample accurately represents the population. The process of achieving this is called sampling (this will be explained later in the course of the lecture). Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population. Let me inform you that he methods of inferential statistics include

(i) the estimation of parameter(s) and
(ii) Hypothesis testing.

Self -Assessment Exercises (SAEs) 1

1. The two main branches of statistics are:
(a) Correlation and Regression analysis.
(b) Description and analysis of Variance.
(c) Inferential and Inductive statistics.
(d) Descriptive and inferential statistics
2. Consider the statements below and indicate which one is Descriptive or Inferential?
(i) I read political magazines two times every week.
(ii) There are more male learners offering POL 803 this session than female learners.

(iii)	The histogram shows that PDP candidates won the local government election.
(iv)	More market women will register to vote after COVID 19 pandemic.
(v)	There is positive relationship between age and level of political satisfaction
(vi)	Older people are more politically satisfied than younger people.

2.5 Summary

In the foregoing we stated and highlighted the two major branches of statistics. They are; descriptive and inferential statistics. Descriptive or Deductive Statistics can be described as the phase of Statistics which seeks only to describe and analyze a given group without drawing any conclusion or inferences about a larger group. Simply put, it assists political scientists to describe and analyze a given political phenomenon under investigation. For instance, you may want to use information gathered through a survey to describe the gender or age of political parties' candidates in the forthcoming general elections in your country. You can do this by using descriptive statistical methods – measures of central tendency (e.g. mode, median and mean) and measures of dispersion (e.g. range, standard deviation and variance). This will be discussed in details later.

Inferential statistics goes beyond mere description to drawing inferences or conclusions from the phenomena analyzed. This is used to infer from the sample data what the population thinks. In other words, you may use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in the study. Estimation of parameter(s) and hypothesis testing are major ways that political scientists can use to draw inferences or conclusions.

2.6 References/Further Reading

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers Pvt Limited.

Ott, Lyman et al (1983). *Statistics: A Tool for the Social Sciences* (3rd Ed.). Boston, Massachusetts: Duxbury Press.

2.7 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. d
2.
(i) Descriptive
(ii) Inferential
(iii) Descriptive
(iv) Descriptive
(v) Inferential
(vi) Inferential

UNIT 3 THE NEXUS BETWEEN STATISTICS AND POLITICAL SCIENCE

Unit Structure

3.1 Introduction

In the last lecture, we identify and explain the major branches of statistics that we can explore in our analysis of political phenomena. It was pointed out that basically, we have two types of statistics: Descriptive and Inferential Statistics. In this lecture, the link between statistics and political science will be explained. This will be done by discussing the general advantages of statistics in political science. Also, the uses of statistics in political science will be outlined and discussed with examples of political careers that need statistics.

3.2 Learning Outcomes

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

- identify the general advantages of statistical approach in political science.
- explain the uses of statistics to political scientists.

3.3 General Advantages of a Statistical Approach in Political Science

It has been mentioned in Module one of this course that statistics have many advantages for the study and practice of politics and related fields. First and foremost, statistics have great power to describe systematically a body of information or data in policy and political analysis, for instance. No other approach matches the precision and quantification that statistics bring to this task. Statistics can explicate in a precise manner the main tendencies, as well as the spread of the data about them using descriptive statistics.

Statistics are also advantageous for subjecting our intuitive ideas about how a political process or phenomenon operates to empirical test. Empirical, in this context means observable or based on data. This confrontation of informed conjecture and speculation with actual data and observation is called hypothesis testing. A hypothesis is an informed guess or conjecture or assumption about an issue or problem of interest. For example, belonging to the ruling political party or a nomination to contest general elections by the godfathers will enhance a candidate's chances of winning the next general elections in Nigeria. Statistics are helpful not only for determining the extent to which the data available support or refute our hypotheses but also for generating the kind of hypotheses that can be tested.

Moreover, statistics are the foremost method for drawing an accurate inference from a subset or sample of data to its parent, the full population. Rarely do political scientists have the luxury of working with the complete population; instead, the data available are almost always a sample of observations. For example, a political analyst may have a sample of all political appointees, civil servants and records – whatever the units of analysis might be – and may want to generalize to the entire population. Political or policy analysts need to know what the sample suggests about the population. Statistics provide an excellent methodology for drawing this linkage. They allow the analyst to evaluate the risk of error when making an inference from sample to population. Since we do not have the data from the entire population, we can still make an error in inferring from the sample. Yet, statistics are valuable, for they enable the analyst to estimate the probability or extent of this error. That is the essence of statistical inference in policy and political studies.

Another major benefit of statistics is that they can help the political scientist keep track of an almost innumerable collection of measured characteristics or attributes, called variables, at the same time. The ability to examine a large number of variables simultaneously – and to sort out and make sense of the complicated relationships among them – is a great advantage of statistical methods for dealing with highly complex situations. An appreciation of statistics can help the policy and political analysts become much more discerning consumer of quantitative information. Like it or not, analysts from various sub-fields of political science are bombarded with "facts" or assertions based on statistical analysis. They appear regularly in myriad sources, including reports, evaluations, memoranda, briefings, hearings, press releases, newspaper accounts, books, academic journals, etc. Policy and political analysts need the skills to evaluate the conflicting claims and

representations often made and to avoid being misled. Statistics offer major benefits in this area.

3.4 Uses of Statistics in Political Science Research

Whether we realize it or not, the relevance of statistics is manifesting almost all the sub-fields of political science. From public policy to public administration, to international relations, statistics in everyday life are used to identify, analyze and affect ideas and behavior. Some of the most interesting statistics are applied in the political realm. It is significant to note that behind the scenes in every arena of politics, statisticians are generating information that fuels political theory, campaign strategy, and policy development. Let's consider some of the uses of statistics related to political systems.

3.4.1 Collection and Dissemination of Public Information

It is imperative for citizens of a given community, state or nation to know and understand how their government functions. Beyond mere knowledge, they want to interpret how political structures, policies, and practices impact their lives. In democratic societies, these citizens exercise their voice by voting. Election polls and public opinion polls are key tools in collecting and disseminating public information for public understanding. A critical element of communicating public information is the media. Writers, reporters and other media personnel rely on statistical reports to inform and educate their audiences. The media is also a forum for calls to action – challenging individuals and groups to act as agents of change.

Another aspect of information-sharing is in the sphere of formal education. Teachers in public and private schools—from primary to university levels—depend on applied statistics in teaching political science and public policy. Consider, just as an example, how many textbooks and educational websites include graphs and charts generated by statisticians.

3.4.2 Election Forecasts

During any election season, media channels clamor for the most current and accurate forecasts of the expected results. Statisticians in discipline of political science develop complex models that consider numerous dynamic factors to deliver the most likely predictions. As data scientists discover new ways to collect and interpret data, election forecasting continues to evolve. Forecasts are important to the general public, the news media, and the candidates.

3.4.3 Political Campaign Strategy

Political candidates spend huge sums of money on election campaigns. While the public primarily sees funding spent on advertising, there are other critical – and costly – financial aspects, too. One large chunk of spending is statistical research that leads to strategy. Research in political elections has ramifications for every aspect of a candidate's campaign. Examples of interesting statistics that lead to actionable information include the following:

- Public opinion that affects a candidate's position on issues
- Voter attitudes that influence campaign messaging
- Demographics that determine targeting
- Media habits and preferences that drive advertising placement

Applied statistics can have a dramatic impact on the outcome of a political campaign, and this creates high value for the role of statisticians.

3.4.4 Micro-targeting in Elections

One particular aspect of campaign strategy has emerged in contemporary context. Micro-targeting is a technique that relies on statistical methods to draw conclusions about individuals from big data. By linking variables in the raw data, data scientists now have the ability to identify consequential patterns that can be applied to predict response on particular issues. Using big data, modern political campaigns have unprecedented access to huge volumes of information about voters. Unlike politics of the past, today's campaigns can target individuals with tailored messages based on their preferences and interests.

3.4.5 Formulation and Execution of Public Policy

Statistical information drives planning and decision-making in public policy, and major research institutes have been established to facilitate these processes. National Bureau of Statistics (NBS), Nigeria Institute for Social and Economic Research (NISER), Nigerian Institute of International Affairs (NIIA) among others are charged with collecting and analyzing data across major government entities such as the Energy Ministry, the Education Ministry and the Environment Labor Ministry.

3.4.6 For Legislation

Public opinion and congressional action are closely connected. It is important for legislators to listen to the people they represent — the people who elected them to office. Members of National Assembly and State Houses of Assembly receive large volumes of communication

from constituents. With the ever-increasing use of technology, the number of contacts from citizens is steeply rising, too. Office staff members are tasked with receiving all this input and translating it to information that reflects the whole. Further, lobbyists represent special interests of the people. As lobbyists seek to persuade legislators, one key resource they use is quantitative information. The receiving lawmaker applies this information in the broader context to more clearly understand issues and potential outcomes of proposed legislation.

3.4.7 Diplomacy and International Initiatives

Data analysis in foreign affairs strengthens diplomacy by providing useful information for programming and policy decisions. Evaluation of international aid or lending based on a comparison of detailed levels of data, global health initiatives based on tracking of mortality rates, fighting human rights violations based on data collected by non-profit groups are some of the areas that political scientists need statistics.

The Social Sciences provide a foundational understanding for improving social systems and building communities. For every public personality leading change in society, there are players in the background studying and catalyzing movements. Likewise, political science as a sub-system of social system provides vital information and data for solving problems associated with the political system. Thus, Statistical information is a core resource for political scientists.

Even, organizations that provide statistics in everyday life are critical to social science research in politics. Political categorization gives insights about the underlying perspectives and values among the citizens of a political state. This work depends on statisticians to develop, execute and communicate research.

Interesting statistics such as social characteristics like gender, education and occupation, public opinion on matters of policy like government health insurance and military spending, evaluation of political candidates, involvement in politics and government accountability are useful to political scientists. In democratic societies, accountability to the people is a core value. Governments, therefore, rely on factual, systematic information from political scientists to guide decision making.

Whatever your particular interests related to politics, there is a wide range of political careers that involve statistics in everyday life. For many of these options, having a higher degree will increase our job opportunities and salary potential. The field of political science includes specialties such as policy analyst, political analyst, political consultant,

political researcher, and political research scientist. Jobs for those who major in political science; right out of college are often entry-level research positions in politics, government, and non-profit work.

One valuable way to specialize in the field of political science is through advanced study in statistics. Statisticians are problem-solvers in many arenas, including political science. Through quantitative analysis, political scientists study theories, systems, trends, and policies. Working as a statistician in the field of political science is a way to apply analytical skills in an area of personal interests. Political officers also, focus on analysis and reporting related to international issues, providing information for policymakers. Public diplomacy officers promote international understanding of State policies. In addition, by discovering and presenting useful information such as polls and graphs, political lobbyist may deliver proprietary information that benefits the lobbyist's interests. To effectively communicate the interest they represent, lobbyists must be skilled in interpreting data. The lobbyist will need to clearly present the views of targeted voting segments, with an understanding of how this relates to potential legislation.

Self -Assessment Exercises (SAEs) 1

1. Identify at least three major benefits of statistics in political science:
2. List at least five ways by which a political scientist can use statistics.

3.5 Summary

In this study unit we discussed the link between statistics and political science by identifying the advantages of statistics to making inquiry in political science. We also identified and explained specific ways by which a political scientist can make use of statistics. Statistics have many advantages for the study and practice of politics and related fields. In the field of political science, statistics help to discuss systematically a body of information or data as well as assist us in subjecting our hypotheses, assumptions and propositions to verification tests. Moreover, it would have been difficult for us to use inferences drawn from a sample to generalize for a particular population without statistics.

More importantly, statistics is linked to political science in terms what political scientists use statistics for. In this study unit, we stated that statistics can be used to collect and disseminate public information, forecast election, formulate and execute of public policy, analyze political campaign strategy, in studying legislation among others. Policy and political analysts, Diplomats, lobbyists are among political experts that make use of statistics. This is why the only valuable way to

specialize in the field of political science is through advanced study in statistics.

3.6 References/Further Reading

Michigan Tech (2022) What Role Do Statistics Play in Politics? https://onlinedegrees.mtu.edu › news › role-statistics-pla.. Accessed at 23 April, 2020

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

2.7 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. Statistics help to describe a body of political information or data. Statistics are used for hypothesis testing. Statistics help in drawing inferences from samples drawn from a population since political scientist cannot study the entire population.
2. - collection and dissemination of public information
   - forecast elections
   - Analyze political campaign strategy
   - formulation and execution of public policy
   - in studying legislation

UNIT 4 STATISTICAL ENQUIRIES OF POLITICAL PHENOMENON

Unit Structure

4.1 Introduction

There are major steps to be followed in carrying out political science research using statistical method. Therefore, in this unit we shall highlight, identify and explain these general steps.

4.2 Learning Outcomes

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

- identify the five major steps in statistical inquiry in political science research
- explain statement of inquiry and sources of research questions in political science research.
- highlight and explain relevant data collection methods in political science research
- describe what organization analysis and interpretation of data entails in statistical inquiry in political science.

4.3 Statement of Inquiry and Data Collection in Political Science Research

4.3.1 Statement of Inquiry

The first and the most important thing in a social inquiry is the preparation of a statement of purpose, clearly and carefully stated. The purpose of the inquiry may be:

(a) To examine the validity of an existing theory or hypothesis in the field of political science.
(b) To discover a new theory or hypothesis.
(c) To investigate the existing state of affairs.
(d) To solve a problem involving the inter-relations of several groups of facts.

A statistical inquiry may be concerned with "Women, Social stratification and Political Participation". This is a very broad subject; the aspect the investigator is interested in must be clearly stated. Your focus may be market women their social stratification and participation in voting during election. You need to identify research questions and formulate them into objectives of the study.

Thus, this first step involves formulation of the research idea/problem which depends on your interest/observation/experience as a researcher or your on-going work as well as matters arising from the work of others. Second, you will conduct a literature review of already done work in the area under investigation. Third, you will need to identify and define your key concepts. This will be followed by formulation of research questions, objectives and hypotheses as appropriate.

4.3.2 Type of Questions

There are three basic types of questions that political inquiries can address:

Descriptive: When a study is designed primarily to describe what is going on or what exists. Public opinion polls (survey of public's view) that seek to describe the proportion of people who hold various opinions are primarily descriptive in nature. For example, if our interest is to know the percentage of the population that would vote for the ruling party or the major opposition party in the next presidential elections in Nigeria, we are simply interested in describing something.

Relational: When a study is designed to look at the relationships between two or more variables. A public opinion poll that compares what proportion of males and females say they would vote for a ruling party or a major opposition party candidate in the next presidential election in Nigeria is essentially studying the relationship between gender and voting preferences.

Causal: When an inquiry is designed to determine whether one or more variables (e.g., a treatment variable or program) causes or affects one or more outcome variables. For instance, if we did a public opinion poll to try to determine whether a recent political advertising campaign changed the voter preferences, we would essentially be investigating whether the campaign strategy (cause) changed the proportion of voters who would vote the ruling party or a major opposition party candidate (effect).
It is significant to note that the three question types can be viewed as cumulative.

4.3.3 Collection of Data

Here, your attention should be paid to:
- Scope of inquiry.
- Determination of statistical units.
- Techniques of data collection.
- Degree of accuracy.

(a) ***Scope of Inquiry***

(i) Scope: This is in reference to space, time and number of items to be covered. With regard to space, one may be interested in a particular city, a local government, a state or the entire country.
(ii) Time: It must be noted that the work of collection of the data must be finished within a reasonable time frame. If more than reasonable time is taken, conditions might have changed and the data collected may be rendered useless.
(iii) Number of Items: Here, we consider whether to work with a sample or the entire population. The choice is that of the researcher but you may go back to lecture one if you are in doubt of using a sample.

(b) ***Determination of Statistical Units***

The researcher must operationalize the necessary concepts properly. That is, the researcher must find the empirical/observable measure that adequately captures enough of the complex reality labelled by the concept.
Characteristics of Statistical Units

- The unit of measurement must be definite and specific.
- It must be of such a nature that may be correctly ascertained.
- Ensure homogeneity and uniformity.
- It should be stable.
- It should be appropriate for the purpose.

Classification of Units

- Natural units e.g. a person, a participant, a cow, a tree, etc.
- Produced units – objects e.g. a house, a car, a table, a ship, etc.
- Measurational units e.g. ton, Kilogram, meter, the year, etc.
- Pecuniary value units e.g. Naira, Dollar, Pound, Cedi, etc.

4.3.4 Techniques of Data Collection

(a) Primary Source: making a special survey i.e. conducting a field inquiry.

We can do this by - Direct interview
- Questionnaire
- Telephone interview
- Observation

(b) Secondary Source: going to the records of some institutions; whether public or private, that collects and publishes data as a routine. e.g. Federal Office of Statistics (FOS), World Bank Publications.

Data can be collected using the following devices:

1. *Personal Interview*: Here, the interviewer and the respondent have a direct contact during the process of interview.

Its advantages include:

- Timeliness, Completeness and accuracy of information, The data collected are original and direct from the source, The questions can be varied to suit the prevailing condition.

Its disadvantages include:

- It is expensive and time-consuming and there is the possibility of personal bias.

2. *Questionnaire***:** This is another mode of data collection. Here, printed list of questions is distributed among intended respondents and statistical facts are generated from their responses. For this technique to be effective and reliable, the following facts must be considered:

- The questions must not be personal, offensive or misleading.
- The questions must be clear and precise.
- The questions must not be difficult or ambiguous.

- The questions must be properly arranged.

Note that a questionnaire may be self - administered or administered by post.

3. *Direct Observation*: In this method of data collection, the researcher observes the characteristic of interest directly from the population or sample. The population may consist of a set of inanimate objects, or a laboratory experiment or when the desired variable can be measured without soliciting response from the respondent.

(d) ***Degree of Accuracy***

The researcher should set a reasonable and achievable level of accuracy.

4.4 Organization, Analysis and Interpretation of Data

4.4.1 Organization of Data

Classification, quantification or operationalization: After collecting your data through any of the above methods considered as appropriate and adequate for the analysis, the next thing is for you to classify or categorize your data, you quantify or operationalize relevant concepts and relate them together. You can use diagrams to present any data that have been adequately organized.

4.4.2 Analysis and Discussion of Data

Using relevant statistical tools to test for association or relationship among the variables under consideration is the next stage in making statistical investigation in Political Science. In short, testing statistical hypothesis/hypotheses constitute the main task of a political researcher at this stage. This involves formulation of a hypothesis, setting up a suitable level of significance, selecting appropriate statistical technique, performance of necessary computations and making of decision from the interpretation of result. Before making decisions on any statistical technique, consider the following:

- What questions do you want to address e.g. is there a relationship between gender and level of political satisfaction? Are women more politically satisfied than men? These are two different questions that require different statistical techniques, the choice depending on data you have collected.
- Find the questionnaire items and scale that you will use to address the questions.

- Identify the nature of each of your variables – dependent, intervening, independent.
- Identify the level of measurements (nominal, ordinal, interval or ratio) for each of your variables. Different statistics are required for variables that are categorical and continuous.
- Decide whether your data meets the basic assumption for the statistical technique you want to use
  Details on steps involved in hypothesis testing will be discussed in another module.

4.4.3 Interpretation of Results and Drawing of Conclusion

This is the last stage of our statistical investigation of social phenomena. It involves drawing of valid inferences from conclusion to explain and predict, identifying areas of further research, suggestions/recommendations (where necessary or possible). Valid statements from the analysis should be made and further testable hypothesis and generalization may also be made. It is important to inform you that the outcome(s) of any statistical inquiry may lead to generation of a generalization, or a set of interrelated generalizations known as theory or an empirical law capable of explaining and predicting political phenomena regardless of time and places. Conclusion –drawing valid inferences from our findings to explain and predict, identifying areas of further research, suggestions/recommendations.

Self- Assessment Exercises (SAEs) 1

1.	State the major steps involved in statistical inquiry of political phenomena.
2.	Highlight four purposes of carrying out political inquiry
3.	Mention four primary sources of data in political science research

4.5 Summary

In the unit we have considered five major steps that are necessary when carrying out statistical investigation or inquiry of political phenomenon. The steps include statement of inquiry, collection or gathering of relevant data, organization of data, analysis of data and interpretation of results from which conclusions are drawn. It is important for you to note that, the same steps can be used when study man in the social context.

3.6 References/Further Reading

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers Pvt Limited.

Ott, Lyman et al (1983). *Statistics: A Tool for the Social Sciences* (3rd Ed.). Boston, Massachusetts: Duxbury Press.

4.7 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

- Statement of inquiry
- collection of data
- organization of data
- analysis of data
- interpretation of results and conclusion

2.

(a) To examine the validity of an existing theory or hypothesis in the field of political science.
(b) To discover a new theory or hypothesis.
(c) To investigate the existing state of affairs.
(d) To solve a problem involving the inter-relations of several groups of facts.

3.

- Direct interview
- Questionnaire
- Telephone interview
- Observation

UNIT 5 SCALES OF MEASUREMENT AND APPLICATION

Unit Structure

5.1 Introduction

In this unit, we shall examine how to determine appropriate scale of measurement and how best to measure our data and for easy presentation and interpretation. The first part of this study unit deals with the meaning, significance and characteristics of measurement scales while the second section identify with examples the main types of measuring scale. I will also discuss with you how to construct measuring scales for social variables.

5.2 Learning Outcomes

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

- define scales of measurement and rules for measurement.
- identify and discuss techniques of measurement
- mention and explain the four levels of measurement.
- give examples of nominal, ordinal, interval and ratio measurement scales.
- understand how to design measurement scales.

5.3 Meaning and Significance of Measurement

Let me start this discussion by telling you that measurement is significant in social research. Appropriateness of measuring scale constructed helps a lot in realizing the aim of every social research. The theory of measurement assumes that a concept representing some phenomenon that the analyst or researcher is concerned about cannot be directly measured. Police officer effectiveness, organizational achievement, legislative performance, level of corruption and electoral violence are all concepts that cannot be measured directly. Such concepts are measured indirectly through indicators specified by operational definitions. Specifically, an operational definition is a statement that tells the researcher or analyst how a concept will be measured. An indicator is a variable or set of observations that results from applying the operational definition. Then what is measurement?

In its daily usage, measurement means a system of measuring, or the act or process of measuring something. In other words, measurement is used to ascertain the extent, quantity, capacity and dimensions of something: for instance; measuring yam flour, cassava, land and height of candidates.

Statistically speaking, measurement is the assignment of numerals or other symbols to empirical properties or events or objects according to certain specified rules. In the social sciences, it is referred to as 'operationalization' of concepts or terms, that is; the pinning of numbers on concepts or variables involved in social research according to some pre – established rules.

Measurement in statistical analysis is significant in social research because:

1. It enhances the exactness or accuracy of calculations;
2. It promotes the scientificness of studies;
3. It enables researchers to quantify or operationalize or pin number on concepts;
4. It helps to avoid erroneous conclusion due to imprecise measurement of data and
5. It promotes the reliability and validity of research findings and conclusions

5.4 Scales of Measurement

5.4.1 Definition of Scales of Measurement

Measurement scales refer to different levels of measurement using a scale of numbers. To get accurate scale, you must follow the rules of measurement. A rule specifies the procedure a researcher uses to assign numerals or a number to objects or events. The function of each rule is to tie the measurement procedure to reality. For instance, assign numerals 1 through 5 to Nigerian leaders since independence according to how radical they are. The rule might further say, if a leader in Nigeria is very radical assign 5 to it, if a Nigerian leader is viewed as not radical at, you may decide to assign the number 1.

5.4.2 Characteristics of Measurement Scales

1. Origin: It marks the beginning of a scale. This may be zero, one or any other number. The best type of scales starts with origin which is zero (natural or absolute) e. g. ruler, speedometer. Fahrenheit thermometer as a measuring scale begins with 32^0 F.
2. Order: it is an ordering measuring property enabling the researcher to put the objects or events or variables in the proper positions I the scale in a pre-arranged manner. Thus making it possible to arrive at logical conclusions.
3. Distance: this refers to the relative position of objects or variables on any given scale.

On the basis of these characteristics, there are four popularly used measurement scales. They are discussed below.

Self-Assessment Exercises (SAEs) 1

Attempt these exercises to assess what you have learnt so far. This should not take you more than 2 minutes.

1. …….. is a system of measuring, or the act or process of measuring something.
2. The theory of measurement assumes that a concept representing some phenomenon that the analyst or researcher is concerned about cannot be directly measured. TRUE OR FALSE
3. Measurement promotes erroneous conclusion due to imprecise measurement of data. TRUE OR FALSE
4. Which of the following is NOT a characteristic of scales of measurement?

(a) Order (b) Quantity (c) origin (d) distance

5.5 Techniques of Measurement

There are four main techniques in measurement: classification, quantification, use of statistical coefficients, and indexing.

(a) Classification

Classification is the simplest form of measurement. It means the assigning of individuals to classes or categories on the basis of a trait or characteristic or attribute, e.g., place of residence, political party, income group, age, marital status, religion etc. Each class is homogeneous, and the classes are mutually exclusive, as members of one category. This is a form of nominal scale does not reveal the magnitude of the differences.

Nominal scales measure discrete phenomena, e.g., sex, race and religion, but the underlying traits are not measured quantitatively. The attitudes of people, for example, over issues and situations do not exist in a simple pro-con dichotomy. It is better to measure them in gradations: strongly favorable, unfavorable, and strongly unfavorable.

(b) Quantification/operationalization

Quantification is the pinning of numbers/weights on concepts, variable, or observation, such as attitudes, the good life, war intensity, etc. Gurr observes that to "operationalize is to decide how to measure a variable". However, the problem is not pinning numbers to things but "What procedures can be used in the measurement so as to get valid and reliable data for each variable?"

(c) Use of statistical coefficients

By measuring relationships or associations between variables through statistical computations, one arrives at coefficients. For example, the correlation coefficient (r) tells us whether two variables are correlated or not; it also tells us the direction or strength of the relationship between the variables; the chi-square index (x^2) is a measure of association for discrete variables such as sex, state, religion, age, education, etc. Also, means or standard deviations are single figures used as measures of typicality for a set of data. A mean can be used to measure the intellectual ability or performance of a group.

(d) Indexing

Indexing is the use of an index (one figure) to represent a set of data. It enables us to measure a one-dimensional concept through one single figure. For example, the cost-of-living index shows the change in the

prices of goods and services; the human development index (HDI) tells us how much development has taken place in a country, taking various factors, especially social factors, into consideration (not just GNP). The infant mortality rate gives a good picture of the state of health of a nation; while the IQ of a student says a lot about his level of intelligence. We need an index when we have to measure or rank a complex phenomenon or need to decide how to rank performance in a contest or class, E.G a beauty contest or an examination performance contest.

In the science one may have to deal with visible/observable or unobservable phenomena to which numbers have to be assigned in the process of measurement.

When one is dealing with attitudes, for example, one needs to note that attributes are not directly observable but must be inferred from behaviour. The same statement is valid for variables like intelligence, personality traits, motives etc. in measuring such variables, the measurement has three sub-processes according to Nwolise (2005):

(a) Identification of behavioural specimens which are considered acceptable as the bases for drawing conclusions about the underlying term. (if measuring attitude, this depends on what is meant by attitude.)
(b) Collection of the behavioural specimens.
(c) Converting the behavioural specimens into a quantitative variable.

TABLE 1

VARIABLE	**DIMENSIONS**	**WEIGHT PER OCCURRENCE**	**WEIGHT PER WEEK/DAY DURATION**	**INTENSITY DEATHS PER 100,000 POPULATION**
Systemic threat	(i) External war (ii) civil war (iii) External raid	10 8 5	8 6 4	+ + +
Elite instability	(i) Coup (ii) Attempted coup)Plot	4 3 2	- - -	+ + +
Repressio	(i) Riot	2	2	+

n	(ii) Demonstrati on)Strike	1 1	1 1	+ +

Source O.B.C Nwolise, (1987)

NOTE Wars were measured on weekly basis, and raids, riots, demonstrations, and strikes on daily basis. The indicator for systemic threat from the table is obtained as the sum of the values of the indices of external wars, and external raids. The indicator for civil war was calculated as the sum of frequency * 8, duration * 6, and intensity + which is measured by the number of deaths per 100,000 population.

5.6 Levels of Measurement

(a) Nominal level
(b) Ordinal level
(c) Interval
(d) Ratio level

5.6.1 The Nominal Level (scale): This is the simplest and most basic level of measurement. It was derived from the Latin word for 'name'. This involves the classification of observations into a set of categories that have no direction to them. It has no order, no distance and no true origin or absolute zero. It is just a label for identification. In terms of its operation, it determines equality or a difference. The numbers assigned to the categories and used as the measure of each case in that category serve simply as names or labels for those categories and cases e.g. Political Party Affiliation in Nigeria's Second Republic can be categorized as follows:

1 = Unity Party of Nigeria (UPN) 2 = National Party of Nigeria (NPN) 3= Nigerian Peoples Party (NPP) 4= Great Nigerian Peoples Party (GNPP) 5 = Peoples' Redemption Party (PRP) 6= National Advance Party (NAP)
Marital Status

1= Married 2= Single 3= Others

Even when numeric values are attached to nominal categories, this is just a way of using numbers as symbols for categorizations that can be easily read by the computer or easily coded and analyzed manually. In essence therefore, only statistical tools that do not assume ordering or meaningful distances should be used to analyze data measured and collected at this level.

5.6.2 The Ordinal Level (scale): This measurement scale only has order. This is a higher level of measurement than the Nominal. The ordinal level of measurement involves classification of data into a set of categories that have direction to them. At the ordinal level, the categories may be placed in order (ascending or descending) and the numbers assigned to the categories reflect that order. It is useful for ordering and to show some kind of relativity – higher, greater, lower etc. It has no statistical connotation other than ranking. In terms of operation, it determines greater or lesser values. For instance, in classifying the income status of Academic staff of the National Open University of Nigeria (NOUN) we may use:

1 = Low income status 2 = Middle income status 3 = High income status

We know that Medium is better than low but we do not know by how much. Likewise, the distance between High and Medium participation is not ascertainable.

5.6.3 Interval Level (scale):

This is the highest level of measurement in social sciences field since it involves the process of assigning real numbers to observations and its intervals are equal. In other words, measuring on the interval scale permits the investigator to measure precise distances between categories, and between the cases in those categories, by using an actual unit of measurement. Thus, it has order and distance but no true origin unless one is assumed for it. It is useful for precision because it helps to present information with little or no ambiguity in its interpretation. At this level, we may discuss how much more or less or how much bigger or smaller a category is than another and identifies the exact distance between any two categories and any two cases within the categories. The scale has magnitude and equal intervals but lacks the real or absolute zero point. In terms of operation, it determines equality of intervals or differences. We can measure age in years, distance in meters and temperature in Fahrenheit on the interval scale. By moving from a higher to a lower level, some information may be lost.

5.6.4 Ratio Scale:

In terms of precision, this is the highest level of measurement. This is similar to the interval scale in all respect but in addition it has an absolute zero point not arbitrary zero value e.g. height, weight, number of books on a shelf, number of arms and ammunitions, population, etc. Although it is the most useful for statistical analysis, it is rare in social research. Also, it facilitates conversion from one unit of measurements to the other using relevant conversion factors. With regard to its

operation, it determines equality or ratio. It is important for you to take note that only continuous variables can be measured at this level of precision.

See the table below for the summary:

Characteristics/ Scales	*ORIGIN*	*ORDER*	*DISTANCE*	**Appropriate Statistical tools**
Nominal	NO	NO	NO	Mode, Percentage Chi-Square
Ordinal	NO	YES	NO	Mode, Median, Percentile Percentage, Chi-Square, Rank Correlation
Interval	NO	YES	YES	Mode, Median, Standard Deviation, T-test, F-test and Product moment correlation
Ratio	YES	YES	YES	Geometric mean, Percent Variance, Linear Regression, Multiple Regression, ANOVA* and ANCOVA**

Table 1: Characteristics and appropriate statistics for each type of Scale of Measurement
*** Analysis of Variance ** Analysis of Co-variance**

It has to be noted also that while some variables can be measured with a single item on our instrument, some variables are difficult to measure with a single item. For instance, Age (in years), gender (female, male and other), religion (Islam, Christianity, traditionalist and other), marital status (single, married, separated, divorced and other) and height in inches, distance in miles etc. are variables that can be measured with a single item. However, such variables with multiple dimensions or aspects as liberty, democracy, performance, stability, power and

tolerance require multi-item measures. For these variables, direct indicators or single questions/entries on our measuring instruments (e.g. the questionnaire) will not be adequate. This is where scaling (discussed below) comes in.

Self-Assessment Exercises (SAEs) 2

Attempt these exercises to assess what you have learnt so far. This should not take you more than 8 minutes.

1. What level of measurement is suitable for the following variables?

(a) Hours of campaign per week _____

(b) Amount in Naira donated to political parties____

(c) Years of experience as a Military Strategist _____

(d) State of residence _____

(e) Speedometer of Barrack Obama's Presidential car _____

(f) Perception about Presidential Speech (Excellent, Good, Fair, Worse)___

2. Indicate the appropriate scale of measurement in this passage: Balewa is a male student while Sophie is a female student ……., Balewa is younger than Sophie ……, Balewa is 30 years old and Sophie is 60 years old …….., and that Sophie is two times as old as Balewa……..

3. Mention at least two major techniques of measurement 1……. and 2 ……..

5.7 Scaling

This is a method of measuring the amount of property possessed by a class of objects or events. Scaling is a more complex process of measurement that involves assigning series of ordered items by using a multiplicity of operational indicators. It helps to

1. Provide a means of ascertaining whether and /or how different aspects of a phenomenon hang together;
2. measure in empirically justifiable, objective and readily interpretable manner;
3. overcome the problem of simple measures which may be difficult to interpret;
4. reduce data to a more manageable size;
5. and ensure universality of the meaning of complex concepts and in the use of scales to measure concepts,
6. as well as yield more accurate and adequate data, in general terms.

In social sciences, scaling is mostly associated with the measurement of attitudes. For measuring psycho-social variables, three major scales are often designed. These include:

(a) **Likert's Summated Rating Scale:** This scale was devised by Rensis Likert and its construction involves generation of statements about the variable being measured and providing a set of graduated response options. In table 2 below, you can see example of this type of rating scale.

A	**B**	**C**	**D**	**E**	**F**
Strongly Agree	Very Satisfactory	Very Good	Very often	Very supportive	Highly appropriate
Agree	Satisfactory	Good	Often	Supportive	Appropriate
Undecided	Neutral	No Opinion	Not certain	Neutral	Neutral Disagree
Disagree	Unsatisfactory	Poor	Rarely	Unsupportive	Inappropriate
Strongly Disagree	Very Unsatisfactory	Very Poor	Never	Very unsupportive	Very Inappropriate

Table 2

Let me show you how to design your own scale using any of the above examples. Using Example A; the instruction: Please read each item and tick (√) the most appropriate column that describes your view. Note that

Strongly Agree (SA): 5 points
Agree:(A) 4 points
Undecided: (UD) 3 points
Disagree: (D) 2 points
Strongly Disagree (SD): 1 point

S/No	Statements	SA	A	UD	D	SD
1	Most Nigerians registered to vote during the 2019 general elections					
2	Most Registered Nigerian voters participated in voting during the 2019 general elections					
3	Most Nigerians that voted were first accredited with Smart Card Readers					
4	Most Nigerians rejected the outcome of the elections					

Table 3

The respondent is expected to indicate his/her position on each statement by ticking the appropriate column in the graduated response options. Numerical values or weights are assigned to options. These are summed up for each individual so as to obtain a total score which represents the respondent's stand on the variable or attribute that is being measured.

(b) **Thurstone Equal Appearing Interval Scale**: If you want to design this type of scale you need to construct a number of statements (items) which are related to attitude being measured. After, you will present these statements to a panel of 25 (or more) judges, who are requested to sort the items into 7 or more categories, ranging from low to high intensity. The same set of judges is asked to rate each of the items in terms of the degree of intensity on a 7 –point scale or more. The average rating from each item is computed, that is the average of the categories into which each item was sorted. See the tables below for the code for rating.

Very high	High	Slightly above average	Average	Slightly below Average	Low	Very low
7	6	5	4	3	2	1

Table 4 (a) Rating code

Category (A)	7	6	5	4	3	2	1	Total
No of Judges (B)	2	2	3	4	3	5	6	25
(A) X (B)	14	12	15	16	9	10	6	82

Table 4 (b) Hypothetical rating of an item by 25 Judges

The mean value is = 82/25 = 3.28 or approximately 3.3
This represents the average point assigned to the item by the 25 judges. You have to subject each of the items constructed in the questionnaire to the same procedure, with a target of obtaining a numerical value or weight for each item. Then you will select about 20 items on the basis of their weights from the pool of items for inclusion in the final instrument.

On the basis of the scale you have designed, respondents are requested to tick 3 to 5 of the twenty statements or items which best represent their opinions on the issue. The individual's score is calculated by taking the mean or median of the scale values of these statements or items that were picked by the respondent. Let us assume that an individual picks statements 2, 8, 12, 14, and 19, with scale values of 2.5, 2.7, 3.4, 3.7 and

4.2 respectively, the mean score is 3.3 and the median is 3.4 (the mean is obtained by taking the average of the 5 scores while median is the middle number). Either of the mean or median values can be used to represent the individual's score. The interpretation is that the higher the score, the more intense is the person's attitude toward the issue under consideration and vice – versa.

(c) **Guttmann's Cumulative Scale**: This is different from Likert and Thurston scales in the sense that it considers the uni-dimensionality of items, that is, the extent to which all the items measure one aspect of a particular variable. For example, if the variable to be measured is 'attitude towards family socialization, there could be many dimensions, e. g. attitude toward authority pattern in the family, parents' socio-economic status, parents civic knowledge or parents involvement in political activities. Respondents may be favourably disposed to all or some or one of these. As you are aware that using Likert and Thurston scales, two respondents who express different patterns of attitude and skill may obtain the same score for their attitude toward family socialization, but on the Guttmann scale, it is assumed that they would have expressed similar patterns of attitude toward family socialization before they can obtain the same score.

Let me point out that the Guttmann scaling technique is more of a procedure for determining the uni-dimensionality of a set of statements or items making up a given scale than a procedure for measuring the attitude of a respondent.

Self-Assessment Exercises (SAEs) 3

Attempt these exercises to assess what you have learnt so far. This should not take you more than 2 minutes.

1. In social sciences, scaling is mostly associated with the measurement of ……..
2. To determine psycho-social variables of politicians, highlight three relevant scales to measure

5.8 Summary

Every social science student, scholar, or researchers owes the profession or discipline the duty of acquiring quantitative ability in order to be able to measure concepts or variables. Measurement is the assignment of numbers to some phenomenon. Some variables cannot be measured so precisely but only in terms of categories. Measurement enhances exactness, truth and predictability, which are the hallmarks of science.

Measurement and measuring scales are very important issues in social statistical analysis in general and in political science research in particular. The nature of the research will determine the measuring scale that is appropriate for the design of research instrument(s).

In this unit, you have learnt that there are different techniques of measurement and our variables can be measured at four levels namely, nominal, ordinal, interval and ratio. Measurement scales refer to different levels of measurement using a scale of numbers. When the variable has no order, distance and true zero, it is called nominal, when it has only order, it is ordinal, if it has distance, and order but no true zero; it is called interval. If a variable has all the three characteristics of a measurement scale – true origin/zero, order and distance, it is referred to as ratio. You can design Likert, Thurston and Guttmann scales to measure attitude in social science research.

5.9 References/Further Reading

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers Pvt Limited.

Johnso J. B. and Joslyn, R. A. , 1991, "Political Science Research Methods", Washington D.C. *Congressional Quarterly* Inc. Chapters 1 -3

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

O.B.C Nwolise, (1987) "Factors explaining high Defence expenditures in Africa 1967 – 1977,"Ph.D dissertation, Department of political science, University of Ibadan

O.B.C. Nwolise, 2005, "Measurement" in Adeniyi Gbadegesin, Razaq Oloopoenia and Afeikhena Jerome, (2005), *Statistics for Social Sciences*, Ibadan:IUP

5.10 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. Measurement
2. True
3. False
4. b

Answers to SAEs 2

1. (a) Interval
   (b) Interval
   (c) Interval
   (d) Nominal
   (e) Ratio
   (f) Ordinal

2. Nominal, Ordinal, Interval, Ratio
3. Classification, Quantification/operationalization, Statistical Coefficients and Indexing

Answers to SAEs 3

1. Attitude
2. Likert rating scale, Thurstone interval scale and Guttsman's cumulative scale

MODULE 2 RESEARCH DESIGN AND SURVEY SAMPLING METHODS

Since a research design is the total plan of a given study, it is imperative to understand what it stands for and what types of research design are relevant to statistical inquiry in political science. Also, owing to cost and time constraints in most cases, it is difficult to use the entire population in survey design. Therefore, it is important for you to understand sampling process. To achieve these objectives, this module is thematically structured into three units under which these issues have been addressed in some details for your learning.

Unit 1 Research Design
Unit 2 Meaning and Type of Sampling Method
Unit 3 Bias in Sampling Survey and Sampling Error

You are advice to study each of the unit carefully as you are expected to answer some questions to evaluate your understanding on the various issues as discussed. Possible answers to the questions are provided under each of the unit appropriately.

UNIT ONE RESEARCH DESIGN

Unit Structure

1.1 Introduction

This unit is devoted to research design. A research design is a systematic strategy intended to evaluate proposed causal explanations on the basis of data. We will distinguish between the two major types of design – experimental and quasi-experimental and elaborate on them.

1.2 Learning Outcomes

At the end this unit you should be able to:

- explain what research design entails in statistical inquiry of social phenomenon.
- distinguish between experimental and quasi – experimental research designs.

1.3 What is Research Design?

A research design is the total plan of a given study. It highlights how the study will be executed with the minimum complications. Specifically, research design is to scientific research what a building plan is to building construction. Essentially, a research design maps out the plan, structure and strategy of scientific investigation. In other words, it helps to ensure that research questions are answered easily and accurately, that research objectives are met in an acceptable manner, and that hypotheses are validly and accurately tested.

In statistical inquiry, research design is a systematic plan for empirically evaluating proposed causal relationships. The design specifies a model of proof for testing the validity of these relationships. It is significant to point out that the research design guides the collection, analysis and interpretation of relevant data. There are different types of research design, and they differ in their ability to generate reliable conclusions or inferences concerning causality. In summary, a research design often outlines:

(a) Observations that will be made to answer questions posed by the research as accurately, validly, objectively and economically as possible;
(b) How the observations will be made;
(c) Analytical and statistical procedures to be applied on data so collected; and
(d) If the goal of research is to test hypotheses, how the test will is to be carried out.

The concept of causality has long been a source of controversy among social scientists. This is due to the fact that several different conceptions of the validity or viability of a causal relationship have been proposed and utilized. The two most important of these are internal and external validity and both are virtually mutually exclusive. In other words, the two cannot be attained or maximized in a single study.

Regarding *internal validity*, the question of whether in a given study or research project, the independent (explaining) variable did indeed cause or lead to changes or effects in the dependent (explained) variable is addressed. In essence, internal validity ensures that no extraneous or intervening variable interferes in the process of explanation. All variables and conditions other than those being studied are controlled, and that the way the study is conducted also does not affect what is being studied. A research design that seeks to enhance internal validity enables the researcher to do the following:

- Ensure that variables extraneous to the research environment do not intrude into the research environment;
- Ensure full control of the research environment so that he/she can directly measure the relationships he/she wishes to measure;
- Establish which variable precedes the other in time and
- Eliminate all alternative explanations for the dependent variable.

Obviously, internal validity can be attained and enhanced only when a study is conducted in a controlled, laboratory -type, researcher -created environment. Contemporary history, maturation (in terms of physical and psychological changes), testing, statistical regression, mortality and instrumentation are factors that can hinder the attainment of internal validity. (Read more about these factors)

In the case of *external validity*, a different idea is captured – representativeness of a sample obtained from a population. It is concerned with the issue of whether and to what extent results obtained in a given study can be inferred or generalized to hold true in settings, time periods, and populations different from the ones used in the study. In other words, external validity enhances the probability that a particular study will contribute to the formulation of general laws in the real world. For example, are the findings of a study of NOUN's postgraduate students generalizable to the population of all students in open and distance learning mode? Are the results of election study conducted in year 2000 still applicable today? External validity, thus, touches on the representativeness of research settings and findings and whether it is possible to generalize from such findings to other situations. The more naturalistic a study situation is, the more it is likely to enhance and maximize external validity. Factors capable of hindering the attainment of external validity include; non-representativeness of sample, effect of study procedure and selection biases.

In terms of types of research design, one can make a major distinction between experimental methods and Quasi-experimental methods. These are discussed in the section that follows.

Self - Assessment Exercises (SAEs)1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 6 minutes

1. What does research design entails?
2. What is the difference between internal validity and external validity?

1.4 Types of Research Design

As mentioned in the last section, we have Experimental research designs and Quasi-experimental designs. Let us discuss them one after the other.

1.4.1 Experimental Design

Most researchers in social science field are familiar with the setup of an experiment. A researcher assembles two groups – Experimental and Control – of subjects; to ensure a valid comparison, the groups should be similar as possible. The experimental group receives some treatment or stimulus, whereas the second or control group does not; instead, it serves as a baseline or reference point for evaluating the behaviour of the first group. Before and after administration of the experimental stimulus, both groups are measured on relevant variables, particularly the dependent variable or criterion. By comparing the scores of the experimental with those of the control group, the researcher is able to determine whether the treatment led to a difference in areas of interest (attitude, behaviour, performance). This procedure is called the classical experimental design.

Let us consider this example for a better understanding of the procedure for experimental research design. Suppose that a political science researcher wanted to examine whether the training of Independent National Electoral Commission (INEC) ad-hoc staff reduced the rate of electoral malpractices. One way to do so would be to select two random samples of INEC ad-hoc staff in a given state. The first group of ad-hoc staff would be required to attend a training programme for a period of three weeks (experimental group); the second group of ad-hoc would be left untrained (control group). Data would be collected regarding past records of experience and capacity of both categories of ad-hoc staff with regards to conduct of election, and after the conduct of general elections, data would be collected again to encompass the period of the study. If in this period the rate of electoral malpractices decreased in the experimental (training) group relative to the rate in the control group, then the researcher would have some evidence for inferring those trainings received by INEC ad-hoc staff reduce electoral malpractices.

Conversely, if the data failed to show this pattern, the proposed causal relationship would be rejected.

Classical Experimental Design

Group Treatment	Random Assignment Observation 2	Comparison	Observation 1
Experimental O_{e2}	R_e $O_{e2} - O_{e1}$	O_{e1}	X
Control $Oc2$	R_c $Oc_2 - Oc_1$		O_{c1}

Source: Kenneth J. Meier et.al. 2006

Please note: O stands for observation or measurement, X stands for administration of experimental treatment, R for random assignment, c for control group, e for experimental group, time subscript 1 for pretest and time subscript 2 for post -test.

Classical Experimental Design Post -test only-control group design

Group Treatment	Observation 1	Randomization
Experimental O_{e1}	R_e	X
Control Oc_1		R_c

Source: Kenneth J. Meier et.al. 2006

1.4.2 Quasi-Experimental Designs

The term ‘quasi’ as an appellation is given because these research designs fail to incorporate one or more features of experimental designs. In particular, in many research design it is difficult to control exposure to the experimental stimulus or independent variable. Let us consider a television reorientation program funded by the federal government and intended to promote political participation. To examine the effectiveness of this program, the researcher would ideally want to study the behaviour of two random samples of people: one that viewed the program and another that did not. In this circumstance, it would be relatively easy to determine whether the program led to effective

political participation. In actuality, however, people interested in participation in politics will tend to watch the program, and those not interested in political participation will tend to seek other diversion – and there is little reason to assume that these two groups are random or matched in any sense. (For one thing, the groups differ dramatically in interest in political participation.) Thus, although the researcher may find that those who viewed the program were subsequently more likely to effectively participate in politics than those who did not, it is not clear whether the program – or initial attitude – was responsible for the impact.

A second aspect of the classical experimental research design that is often lacking in quasi -experimental designs is repeated measurement. To appraise whether the independent variable produces changes in the dependent variable, it is very supportive to know respondents' scores on the variables prior to a change in the independent variable. Then one can determine whether this change is accompanied by a change in the dependent variable. In the experiment, this function is served by measurement before and after the experimental treatment, which is intended to affect or change the level of the independent variable in the experimental group (for instance, training of INEC ad-hoc staff is intended to increase the credibility of elections).

The most fundamental difference between quasi-experimental research designs and experimental designs centers on the ability of the researcher to control exposure to the experimental treatment or independent variable. The control is much greater in experimental designs than in quasi-experimental designs. Perhaps the most widely used quasi-experimental design is the cross-sectional study or correlational study. This type of study is based on data obtained at one point in time, often from a large sample of subjects. Most surveys of public opinion or attitudes toward government or civil society organizations are cross sectional studies. Another one is a case study; which is an in-depth examination of an event or location, usually undertaken after something dramatic has occurred (such as the Nigeria civil war or #EndSARS protest of 2020). Although a case study may rest (at least partially) on data obtained from a random sample of respondents, more often the researcher relies on information from carefully selected individuals (informants) and archival records.

There is also a panel study which refers to a series of cross-sectional studies based on the sample of individuals over time; that is a group of individuals is surveyed repeatedly over time. As examination of the effects of the university that followed incoming students until their graduation would be a panel study. Trend studies is the last example of quasi-experimental designs. These studies monitor and attempt to

account for over-time shifts in various indicators, such as gross national product, unemployment, corruption, number of civil society organizations registered by Federal inland revenue service, attitude toward the president among others.

Some common Quasi-Experimental Designs of Research

Designs	Design Diagram
Cross Sectional Studies	*X O*
Case Study	*X O*
Panel Study	O_1 *X* O_2
Trend Study	O_1 O_2 O_3 *X* O_4 O_5 O_6

Source: Kenneth J. Meier et.al. 2006
Note: X represents the independent variable, O stands for observation or measurement and the subscripts 1, 2, 3, 4,.. refer to time points at which relevant variables are measured.

Quasi -experimental designs are relatively strong in demonstrating covariation between independent and dependent variables. Many statistics have been developed for assessing the magnitude of covariation or association between two variables. As long as the independent and dependent variables are measured across a sample of subjects, the researcher can use statistics to assess the degree of covariation.

Self - Assessment Exercises (SAEs) 2

Attempt these exercises to assess what you have learnt so far. This should not take you more than 5 minutes

1. Mention two major groups in experimental design.
2. The most widely used quasi-experimental design is the ……
3. …. a series of cross-sectional studies based on the sample of individuals over time.

1.5 Summary

Research designs involve setting up a research project so that research questions can be answered as unambiguously as possible. The objective of a good research design is to establish causal relationships and to assess their generalizability. The two major types of variables are

independent and dependent. Independent variables are anticipated causes, while dependent variables are the variables thought to be affected by them. A hypothesis formally proposes an expected relationship between an independent and dependent variable.

It is important to point out that there are four criteria identified by social scientists as necessary for establishing a relationship as causal. These are time order, non-spuriousness, covariation and theory. A research design is a program for evaluating empirically proposed causal relationships. This evaluation is based on two factors: Internal validity (did the independent variable lead to dependent variable?) and external validity (can the results obtained in the study be generalized to other populations, times and settings?) There are two major types of research design -; namely experimental research designs and quasi-experimental research designs

In this unit, what research design entails was explained. Both types of research design – experimental and quasi-experimental design were evaluated with regard to the four criteria for causal relationships that define internal validity and with external validity. You were informed that the experimental design has its primary strengths in internal validity and the quasi-experimental design has its strength in external validity

1.6 References/Further Reading

Johnso J. B. and Joslyn, R. A. , 1991, “Political Science Research Methods”, Washington D.C. *Congressional Quarterly* Inc. Chapters 1 -3

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

1.7 Possible Answers to Self-Assessment Exercises

Answers to SAEs 1

1. Research designs entail:

(a) Observations that will be made to answer questions posed by the research as accurately, validly, objectively and economically as possible;
(b) How the observations will be made;
(c) Analytical and statistical procedures to be applied on data so collected; and
(d) If the goal of research is to test hypotheses, how the test will be carried out.

2. While the internal validity ensures that no extraneous or intervening variable interferes in the process of explanation, external validity emphasizes representativeness of a sample obtained from a population.

Answers to SAEs 2

1. Experimental and Control groups
2. Cross-sectional study or correlational study
3. Panel Study

UNIT 2 MEANING AND TYPES OF SAMPLING METHOD

Unit Structure

2.1 Introduction

In this unit, definition of sampling method and its significance will be discussed. This is important, because it has always been difficult to study entire populations. More importantly, in contemporary times, it is no longer necessary to study entire populations. With advancement in the methodology, tools and techniques of social scientific investigation, all that is required now is the study of subsets (called samples) of the larger population drawn up in a scientific manner to enhance their ability to represent or look like the population as closely as possible with regard to the research problem under investigation. Also, the various sampling methods or techniques will be identified and explained.

2.2 Learning Outcomes

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

1. define sampling method and identify its significance.
2. highlight and explain with examples different types of samples that we have.
3. Know and be capable of utilizing sampling scientifically and validly in the research work.

2.3 What is Sampling Method?

Generally speaking, it is the process of selecting a group of people or products or items, or events to be used as a representative or random

sample from a population. Sampling method refers to the way that observations are selected from a population to be in the sample for a sample survey. The main reason for conducting a sample survey is to estimate the value of some attribute of a population. Let us remind ourselves again the difference between population parameter and sample statistic: A population parameter is the true value of a population attribute, while a **Sample statistic** is an estimate based on sample data, of a population parameter. Consider this example. A public opinion pollster wants to know the percentage of voters that are against legislation for same sex marriage. The *actual* percentage of all the voters is a population parameter. The *estimate* of that percentage, based on sample data, is a sample statistic. The quality of a sample statistic (i.e., precision, meticulousness, representativeness) is strongly affected by the way that sample observations are chosen; that is, by the sampling method.

Sampling as a scientific procedure rest on two pillars. These are the principle of randomization, which enables us to draw samples representative of the population; and statistics which enables us to make valid inferences about the sample and from the sample to its population.

2.4 Importance of Drawing a Sample

Let me first and foremost inform you that several reasons why we need to draw sample(s) from our population of study or analysis constitute the importance or significance of sampling. These reasons include the following:

1. It is economical because it is cheaper to work with a sample when the population is large. Assuming the population of politicians in Nigeria is 5 million or more, can we interview all of them? Since population is often too large for us to study, we make use of samples.
2. It saves time and energy because the cost of studying population may be too prohibitive.
3. Where a population is partly inaccessible or unknown, drawing a sample will be more appropriate.
4. Where members of the population can easily be destroyed while studying them, samples are more appropriate.
5. Inferences can be easily drawn from samples than form the entire population.

Self - Assessment Exercises (SAEs) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 8 minutes

1. What is sampling method?
2. Statistic is to sample as parameter is to ----------
3. Statistics can be used to draw an accurate ……. from a …… of data to its parent, the full ……

(A) description, inference, population
(B) parameter, statistic, sample
(C) inference, sample, population
(D) parameter, subset, population

4. One of the advantages of drawing a sample from the population of study is to save energy and time. TRUE OR FALSE

2.5 Probabilistic Sampling Methods

The method of selecting a sample is called sampling procedure or sampling technique. As a group, sampling methods fall into one of two categories – probability sampling techniques or methods and non - probability sampling techniques or methods. Let us first discuss the probability sampling methods.

With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample. In other words, they have equal chance of being selected.

The major types of probability sampling methods are simple random sampling, stratified sampling, cluster sampling, multistage sampling, and systematic random sampling. Let me inform you that the main advantage of probability sampling methods is that they guarantee that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid. **Random sampling** is a procedure for sampling from a population in which

(i) the selection of a sample unit is based on chance and
(ii) every element of the population has a known, non-zero probability of being selected.

Random sampling helps produce representative samples by eliminating voluntary response bias and guarding against under-coverage bias. All probability sampling methods rely on random sampling.

1. *Simple random sample*. This is that technique which ensures that every member of the population has the same chance of being included in the sample. This is considered to be the best sampling

technique. Simple random sampling refers to any sampling method that has the following properties.
- The population consists of N objects.
- The sample consists of n objects.
- If all possible samples of n objects are equally likely to occur, the sampling method is called simple random sampling.

There are many ways to obtain a simple random sample. One way would be the lottery or balloting method. Each of the N population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects n numbers. Population members having the selected numbers are included in the sample.

For example, if you want to select 100 students out of 5000, you can write their matriculation number or name in separate pieces of paper, roll the paper and mix it properly in a container. Then, you will select one piece of paper from the container at a time, mixing the remaining content after each selection until the desired 100 students are selected. By doing that, you have given all the students equal chance of being selected.

1. ***Stratified sample***. With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a probability sample (often a simple random sample) is selected. In stratified sampling, the groups are called **strata**. As an example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents. You should note that this method is often employed where the population is homogeneous.

2. ***Cluster sample***. With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed. Specifically, after defining the population and identifying all possible clusters in the population from the largest to the smallest, then you will successively sample clusters from the very large groups to the large groups to sub-groups to sub-sub-groups etc. until you get to the stage of individual subjects. Note the difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes elements from each stratum. With cluster sampling, in contrast, the sample includes elements only from sampled clusters. This method can be employed where the population is heterogeneous. It is also a very useful method when dealing with a large population or when a list at the macro levels of sampling will be difficult, if not impossible to compile.

3. ***Multistage sample***. With multistage sampling, you select a sample by using combinations of different sampling methods. For example, in Stage 1, you might use cluster sampling to choose clusters from a population. Then, in Stage 2, you might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

4. ***Systematic random sample***. With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first *k* elements on the population list. Thereafter, we select every *Kth* element on the list. This method is different from simple random sampling since every possible sample of *n* elements is not equally likely. For instance, if you have a population of 500 elements, and you need a sample of 50 elements. You need to determine the sampling interval using this formulae:

K = N/n N = Population size n = sample size

Therefore, for our example, we have 500/50 = 10. You then select the first element of the sample at random say number 5 on the population list. The sample list would include the 5^{th}, 15^{th}, 25^{th}, 35^{th}, 45^{th} and so on until you have reached the sample size of 50.

2.6 Non-probabilistic Sampling Methods

With non-probabilistic sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen. It favors the bias or interest of the researcher. You should take note that non-probability sampling methods offer two potential advantages - convenience and cost. The main disadvantage is that non-probability sampling methods do not allow you to estimate the extent to which sample statistics are likely to differ from population parameters. Only probability sampling methods permit that kind of analysis.

There are six major types of non-probability sampling methods. They are:

(i) *Voluntary sample*. A voluntary sample is made up of people who self-select into the survey. Often, these respondents/samples have a strong interest in the main topic of the survey. Suppose, for example, that a weekly newspaper asks readers to participate in an on-line poll. This would be a volunteer sample. The sample is chosen by the readers, not by the survey administrator.

(ii) *Accidental Sample*: An accidental sample is made up of people who are selected by chance without any deliberate plan to select them. For instance, you are to interview residents of a particular place who are living with disability in a particular order but you do not know where exactly they reside. Any one you meet by chance becomes part of your

sample. The first come, first serve basis is always used to select the sample.

(iii) ***Convenience sample***. A convenience sample is made up of people who are easy to reach. Consider this example: A researcher interviews shoppers at a local market. If the market was chosen because it was a convenient site from which to solicit survey participants and/or because it was close to the researcher's home or place of work, this would be a convenience sample.

(iv) ***Purposive or Judgmental sample***: the sample in this case is selected subject to the choice of the researcher/pollster. For example, if you are interested in administering questionnaire to candidates that want to contest the next gubernatorial elections in Nigeria. You can choose any candidate from 36 states in Nigeria and decide the number of contestants you intend to select.

(i) ***Quota sample***: To select a sample using quota sampling method, you need to stratify the population into different strata, after which you apply a non- random sampling in selecting the elements of the sample. Assuming that the sample size is 1500 students and you were asked to use the following quota to select students from the following halls of residence: Independence Hall (30%), Idia Hall (35%), Kuti Hall (20%) and Awolowo Hall (15%)

(ii) ***Snowball sample***: A snowball sample is constituted by reaching other sample elements through initial elements previously included in the sample. In fact, you can use snowball sampling method to get samples of tax evaders, drug addicts, prostitutes and cultists.

Self - Assessment Exercises (SAEs) 2

Attempt these exercises to measure what you have learnt so far. This should not take you more than 8 minutes

1. In not more than two paragraphs, differentiate between probability and non- probability samples.
2. A political consultant is conducting a satisfaction survey, sampling from a list of 1,000 legislators (former and current). The list includes 250 southwest legislators, 250 southeast legislators, 250 northwest legislators, and 250 north-central legislators. The consultant selects a sample of 40 legislators, by randomly sampling 10 legislators from each of selected geopolitical zone. Is this an example of a simple random sample?

(A) Yes, because each legislator in the sample was randomly sampled.
(B) Yes, because each legislator in the sample had an equal chance of being sampled.
(C) Yes, because legislators from each geopolitical zone were equally represented in the sample.
(D) No, because every possible 40-legislator sample did not have an equal chance of being chosen. (E) No, because the population consisted

of legislators from only four geopolitical zones.

2.7 Summary

Probabilistic sampling methods are sampling methods in which the researcher required to utilize the principle of randomization (or chance procedure) in at least one of the stages of the sampling process. On the other hand, the non-probabilistic sampling methods do not apply the principle of randomization in their procedures.

In this unit, we defined sampling method as the process of selecting a group of people or products or items, or events to be used as a representative or random sample from a population. It is significant in political research because it is cheaper to work with samples that the entire populations as well as also saves time and energy. There are two major types of sampling methods or techniques, namely probabilistic and non-probabilistic sampling methods. Examples of probabilistic sampling methods include simple random, systematic, stratified, cluster among others. Accidental, quota, purposive are among the basic non-probabilistic sampling methods.

2.8 References/Further Reading

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers Pvt Limited.

Johnso J. B. and Joslyn, R. A. , 1991, "Political Science Research Methods", Washington D.C. *Congressional Quarterly* Inc. Chapters 1 -3

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

2.9 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1
1. Sampling method refers to the way that observations are selected from a population to be in the sample for a sample survey.
2. Population
3. (C)
4. True

Answers to SAEs 2

1. The method of selecting a sample is called sampling procedure or sampling technique. As a group, sampling methods fall into one of two categories- probability and non-probability. With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample. In other words, they have equal chance of being selected. Regarding non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen.

2. (D) No, because every possible 40-legislator sample did not have an equal chance of being chosen.

UNIT 3 BIAS IN SURVEY SAMPLING AND SAMPLING ERROR

Unit Structure

3.1 Introduction

Having discussed what a sampling method entails, its types and the procedures, it is important to discuss the issue of bias in survey sampling. A good understanding of the nature of bias in selection of samples will enable us to minimize the chance of obtaining unrepresentative samples from our populations. Also, in this unit, you will learn about sampling error and understand how you can reduce it.

3.2 Learning Outcomes

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

- understand the concept of bias in survey sampling.
- explain sampling error and what could be done to reduce it in survey sampling.

3.3 Bias in Survey Sampling

In survey sampling, bias is considered as the tendency of a sample statistic to systematically over- or under-estimate a population parameter. Bias may occur as a result of unrepresentative samples. A good sample should be **representative.** This means that each sample element represents the attributes of a known number of population elements. Bias often occurs when the survey sample does not accurately represent the population. The bias that results from an

unrepresentative sample is called **selection bias**. Some common examples of selection bias include the following:

(i) Under-coverage: This happens when some members of the population are inadequately represented in the sample. Under-coverage is often a problem with convenience samples.

(ii) Nonresponse bias: It may interest you to know that sometimes, individuals selected for the sample are unwilling or unable to participate in the survey. Non-response bias is the bias that results when respondents differ in meaningful ways from non-respondents.

(iii) Voluntary response bias. This occurs when sample members are self-selected volunteers, as in voluntary samples. An example would be call-in television shows that ask for audience participation in surveys on controversial topics (abortion, third term agenda, same sex marriage, Muslim-Muslim presidential ticket etc.). The resulting sample tends to over-represent individuals who have strong opinions.

Random sampling helps produce representative samples by eliminating voluntary response bias and guarding against under-coverage bias. It is not surprising that all probabilistic sampling methods rely on random sampling.

A poor measurement process can also lead to bias. In survey research, the measurement process includes the environment in which the survey is conducted, the way that questions are asked, and the state of the survey respondent. **Response bias** refers to the bias that results from problems in the measurement process. I am quite sure that you wish to know some examples of response bias. They include:

(a) Leading questions. The wording of the question may be loaded in some way to unduly favour one response over another. For example, a satisfaction survey may ask the respondent to indicate where he/she is satisfied, dissatisfied, or very dissatisfied. By giving the respondent one response option to express satisfaction and two response options to express dissatisfaction, this survey question is biased toward getting a dis-satisfied response.

(b) Social desirability. Most people like to present themselves in a favourable light, so they will be reluctant to admit to nasty attitudes or illegal activities in a survey, particularly if survey results are not personal or confidential. Instead, their responses may be biased toward what they believe is socially desirable.

3.4 Sampling Error

As you were told earlier, a survey produces a sample statistic, which is used to estimate a population parameter. If you repeated a survey many times, using different samples each time, you might get a different

sample statistic with each replication. And each of the different sample statistics would be an estimate for the *same* population parameter.
However, if the statistic is unbiased, the average of all the statistics from all possible samples will equal the true population parameter; even though any individual statistic may differ from the population parameter. The variability among statistics from different samples is called **sampling error**. Increasing the sample size tends to reduce the sampling error; that is, it makes the sample statistic less variable. However, increasing sample size does not affect survey bias. A large sample size cannot correct for the methodological problems - under-coverage, nonresponse bias, etc.- that produce survey bias.

Self-Assessment Exercises (SAEs) 1

Attempt these exercises to assess what you have learnt so far. This should not take you more than 7 minutes
1. Indicate whether the following statements are **true** or **false**.
I. Random sampling is a good way to reduce voluntary response bias.
II. To guard against bias from under-coverage, use a convenience sample.
III. Increasing the sample size tends to reduce survey bias.
IV. To guard against nonresponse bias, use a mail-in survey.
2. Identify two sources of selection bias that you know.

3. What is sampling error?

3.5 Summary

In survey sampling, every sample supposed to be representative of the population from which it is drawn. If the samples are unrepresentative of the population, selection bias may occur and if the measurement process is not well controlled, there may be response bias. Random sampling as mentioned in the unit will assist in reducing selection bias. The variability among statistics of different samples from a particular population is referred to as sampling error which can be reduced by increasing the sample size.

What we have done in this unit is explanation of bias in survey sampling and attempt was made to identify types of bias that may occur if our samples are not representative of the population of the study. Bias in survey sampling is different from sampling error which is described in the unit as variability among statistics from different samples of a given population.

3.6 References/Further Reading

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

Obasi, I.N. (1999). *Research Methodology in Political Science*. Enugu: Academic Publishing Company

3.7 Possible Answers to Self-Assessment Exercises (SAEs) 1

1. (I) True (II) False (III) True (IV) False
2. Under-coverage, non-response, voluntary response
3. The variability among statistics from different samples is called **sampling error**.

MODULE 3 DESCRIPTIVE STATISTICS

Descriptive statistics are used simply to describe the sample you are concerned with. They are used in the first instance to get a feel for the data, in the second for use in the statistical tests themselves, and in the third to indicate the error associated with results and graphical output. In this module, we will start our discussion on how to present raw data collected from the field. You will be taught how to manage and organize your collected data by sorting them out using tables and other means. Also, we will focus our attention on measures of central tendency and measures of dispersion and variability. These measures help to describe data fully using different but relevant parameters that are also used in hypothesis testing. To achieve these objectives, this module is thematically structured into five units under which these issues have been addressed in some details for your learning.

Unit1 Data Presentation
Unit 2 Measures of Central Tendency
Unit 3 Describing Data with Averages
Unit 4 Measures of Dispersion or Variability
Unit 5 Describing Variability: Quantitative and Qualitative/Ranked Data

You are advice to study each of the unit carefully as you are expected to answer some questions to evaluate your understanding on the various issues as discussed. Possible answers to the questions are provided under each of the unit appropriately.

UNIT 1 DATA PRESENTATION

Unit Structure

1.1 Introduction
1.2 Learning Outcomes
1.3 Using Tables for Data Presentation
1.4 Diagrammatic/Graphic Presentation of Data
1.5 Summary
1.6 References/Further Reading
1.7 Possible Answers to Self-Assessment Exercises

1.1 Introduction

Recall that in the previous units we identified that there are two branches of statistics-descriptive and inferential statistics. Descriptive statistics deals with presentation and description of collected data

without drawing any inferences, while inferential statistics goes beyond description of data and involves drawing of inferences or conclusion from analyzed data. Here we first dealt with descriptive statistics in this module. Specifically, in this unit, we shall examine how best to organize and present the data. The first section of the unit deals with the use of tables to present data while the second section examines the use of diagrams for the same purpose.

1.2 Learning Outcomes

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

1. manage and organize raw data by tabulating and grouping them as the case may be.
2. present the major information of the data collected on a chart and with the use of diagrams for easy understanding.
3. present frequency distribution table of raw and grouped data

1.3 Using of Tables for Data Presentation

Data organization and presentation/representation are activities carried out by the researcher after data collection. It involves organization of data collected numerically. Arranging numerical data in ascending or descending order of magnitude is called an *array*. The difference between the largest and smallest numbers is called the *range* of the data. There are several methods of presenting data collected. These include classification, tabulation of data, and diagrammatic representation. The last two are emphasized in this section (classification has been discussed under measurement). Meanwhile, classification means sorting or grouping data collected on the basis of similarities or common characteristics.

Tables are a basic means of data presentation. Tabulation is the arrangement and sorting out of data into different categories of rows and columns and counting the number of cases that belong to each category. Tables aid data analysis by displaying facts, figures or information in orderly fashion in small space so that it can be easily seen at a glance the important results. These may be done by hand or machine. There are two types of tabulation. Let us consider them one after the other.

1. Simple Tabulation: Here, data pertaining to one variable only is tabulated, i.e. only one set of numbers is tabulated against a group or class. For example, assuming you collected the individual age of all

registered voters in your local government, the information will be better managed if you arrange it into a table as follows:

Univariate tabulation

Age (class)
1 – 20
21 – 40
41 – 60
61 – 80
81 and above

Table 1

Table 1 above is an example of univariate tabulation showing the age in class interval of certain selected residents in Lagos state.

Bivariate Tabulation

Age (class) in years	No of Registered Voters (frequency) in Millions
1-20	2,500
21-40	12,000
41-60	9,800
61-80	4,600
81 above	1,200

Table 2

In table 2 above, the first column is called the class limits because you have grouped the voters into different age brackets. The second column is called the frequency column which gives the actual number of registered voters in each age bracket.

2. *Complex Tabulation*: At times, you may want to record several variables (Muti-variate) about your respondents on the same tables. In this case you subdivide the table into different categories or parts. For instance, assuming in Table 4.1, in addition to the age bracket, you also have the number that are male and female in each group. You may also ask them whether they are single or married. You can then put all these information together in a complex tabulation as follows:

Table 3: **Complex Tabulation**

Gender/Marital Status

Age (class)	Freq	**Male** Married	Single	**Female** Married	Single
1-20	2,500	0	800	100	1,600
21-40	12,000	3,500	4,500	3,600	400
41-60	9,800	4800	200	4,500	300
61-80	4,600	1,600	0	3,000	0
81 above	1,200	750	0	450	0

You can now see why we call table 3 above a complex tabulation. It contains information on several variables – age, gender and marital status.

1.4 Diagrammatic or Graphic Presentation of Data

This involves the use of symbols, diagrams and pictures to represent our data. The various forms of diagrammatic representation are discussed in this section.

1. *Pictogram*: This is a means of using symbols or pictures to represent data. The symbols or pictures chosen often bear direct relationship with the concept being portrayed.
2. *Bar Chart*: This is a diagrammatic representation used for qualitative data consisting of separated vertical bars of equal width whose heights are drawn proportional to the frequencies of the class they represent. It is a plot of the class limits against the frequencies. Bar graphs can be simple or complex depending on the number of variables. The bar graph shows the number of cases in particular categories, or score on some continuous variables for different categories. For bar graph, you need two main variables: one categorical and one continuous.
3. *Pie-Chart*: A pie chart displays qualitative data pictorially. It consists of drawing a circle and dividing it up into sectors of a particular size corresponding to the frequencies of the classes being considered (see the illustration below).
4. *Histogram*: The histogram is a diagrammatic representation used for quantitative data consisting of continuous, joined vertical bars (meeting at class boundaries). Histogram are used to display the distribution of a single continuous variable e.g. age. The areas

contained by the bars being drawn proportional to the frequencies of the classes they represent. Because the bars must touch one another, you should plot the class boundaries against the frequencies. We shall now use the information provided on table 5 below to illustrate the four concepts discussed in this section.

5. *Scatterplots:* are typically used to explore the relationship between two continuous variables (e.g. Voting age and political participation).
6. *Line Graphs*: allow the researcher to inspect the mean (or sum) scores of a continuous variable across a different values of a categorical variable.

Table 5: **Population of 5 Cities in Nigeria** (Figures are hypothetical)

Cities	**Population in millions**
A	30
B	25
C	15
D	20
E	10

To draw a pictogram, we draw the figure of a person to represent one million people and half the figure to represent half a million people. Thus, the Pictogram of the data in Table 5 appears:

A 30 million

B 25 million

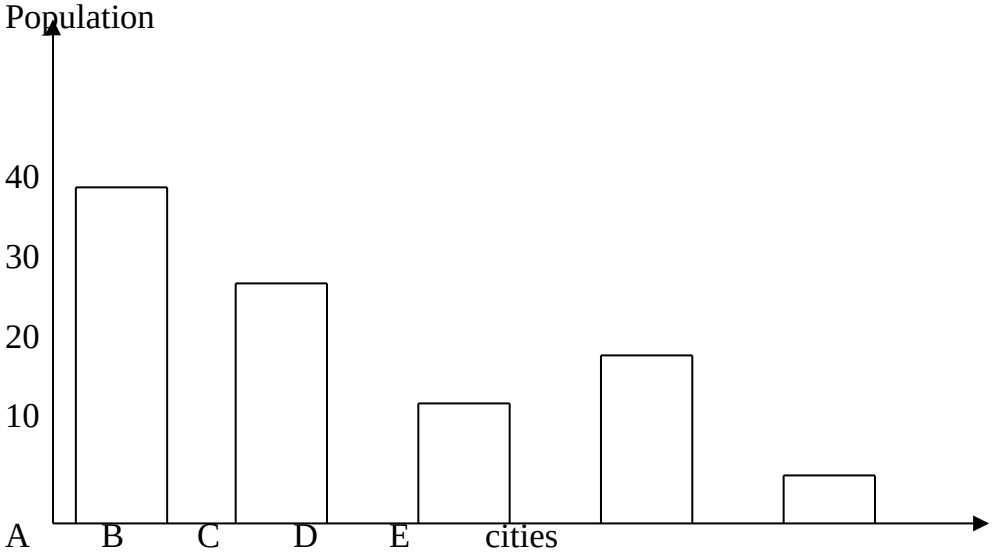
C 15 million

D 20 million

E 10 million

To draw the Bar chart of data on Table 5, scale the frequency column on the vertical axis and the cities or classes on the horizontal axis. Ensure equal interval between the bars and label the chart accordingly.

Bar chart of Table 5



To draw the pie-chart, we first convert the number of persons in each city to degree i.e. unit of measuring an angle in a circle. The conversion is simple, it is just:

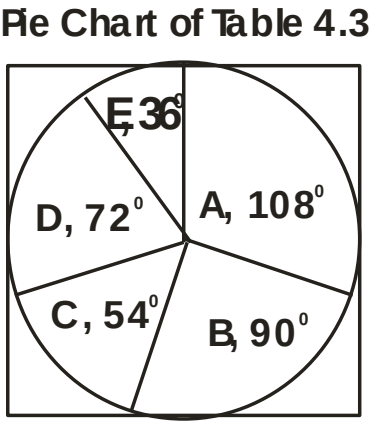
$$\frac{\text{Population of a given city}}{\text{Total population of all 5 cities}} \times 360°$$

Thus,

Column 1	Column 2	Column 3
Column 4 (%)		
A	$\frac{30}{100} x360^0$	= 108°
30%		
B	$\frac{30}{100} x360^0$	= 90°
25%		
C	$\frac{15}{100} x360^0$	= 54°
15%		
D	$\frac{20}{100} x360^0$	= 72°
20%		
	$\frac{10}{100} x360^0$	

E		= 36°
	10%	
	Total	360°

Note that adding column 3 together you must arrive at 360°, likewise addition in column 4 must be 100. Otherwise go back to the computation and check for the mistakes.
Next, draw a circle and divide it into sectors each sector representing each value in column 3.



Self-Assessment Exercises (SAEs) 1
Attempt these exercises to assess what you have learnt so far. This should not take you more than 50 minutes.

Use the table below to draw
(a) Pie chart (b) histogram (c) bar chart

Countries In Africa	No of legislative seats
Togo	50
Ghana	120
Cameroon	100
Nigeria	400
South Africa	250
Gabon	100
Angola	300
	1320

1.5 Frequency Distribution

What we are going to do in this section is to elaborate more on some of the concepts we have encountered in the previous sections. The distribution is a summary of the frequency of individual values or ranges of values for a variable. In other words, by frequency distribution, we

mean a tabular arrangement of data with their corresponding frequencies e.g. Tables 4.1 and 4.2 are frequency distribution tables. Let us now define few other concepts.

1. *Raw Data*: These are data collected or recorded which are yet to be organized or processed.

2. *Class*: By class, it means a collection of items put or grouped together in a particular unit following a definite order. Refer back to Table 4.1, the age bracket 0-20 or 21-40 etc. are the classes. For the second class, 21 is the lower class limit while 0 is the upper class limit. The last class, 81 and above is an open class interval.

3. *Class Boundaries*: For a continuous data, class boundaries (or real limits) are determined simply by subtracting 0.5 from the lower class limits and adding 0.5 to the upper class limit of each class e.g. the class boundary for the class 21-40 are 20.5-40.5, and for the next class, the class boundaries are 40.5-60.5.

4. *Class Interval or Class Width*: This is the size of the class in question. It is the number of items (in units) contained in a particular class e.g. for the class 21-40, the class interval is 20. This is derived by:
Class interval = Upper class limit – Lower class limit
= 40.5 – 20.5
=20

You should now calculate the class interval for all other classes. All must give you 10 except the open class which you don't need to calculate for.

5. *The Class Mark*: This is the mid-point of a given class. It is obtained by adding the lower and upper class limits and thereafter dividing the sum by 2 e.g. for the second class in table 4.1, the class mark is:

$$\text{Class mark (x)} = \frac{21+40}{2}$$

$$= 35.5 \qquad \frac{61}{2}$$

You should also calculate the class mark of all the other classes except the last class.

6. *Cumulative Frequency Curve*: This is the plot of the cumulative frequency column against the classes. To obtain the cumulative frequency column, start from the first class, add the frequency of the

first to that of the second class and write down the sum against the second class. Add on successively until you get to the last class. Cumulative frequency curve is sometimes referred to as the Ogive or the "less than" cumulative curve.

7. *Frequency Polygon*: Plotting the mid-points of the classes against the corresponding frequency gives a frequency polygon.

I will now illustrate how to draw these charts with a good example:
Example 1: The age of 80 registered voters is given below:

Age group	10-19	20-29	30-39	40-49	50-59
No of registered voters	9	17	25	19	10

Use the table above to calculate:
1. The lower and upper class boundaries.
2. The class width.
3. Construct the cumulative frequency table.
4. Plot the Ogive and frequency polygon.
5. Draw the histogram.

Solution

1. See column 4 in (iii) below. The procedure is to subtract 0.5 from the lower class limit and add 0.5 to the upper class limit (of column 1).

2. Class width = Upper – Lower class boundaries
= 19.5 – 9.5
= 10 units

It is the same result for all other classes.

Column 1	Column 2	Column 3	Column 4	Column 5
Age group (class)	No of voters	Cum. Freq (f)	Class boundary	Class mark
10-19	9	9	9.5-19.5	14.5
20-29	17	26	19.5-29.5	24.5
30-39	25	51	29.5-39.5	34.5
40-49	19	70	39.5-49.5	44.5
50-59	10	80	49.5-59.5	54.5

3. To plot the frequency polygon, we need to compute the class mark (see column 5).
Class mark = Lower + Upper class limit

$\equiv \frac{10+19}{2}$

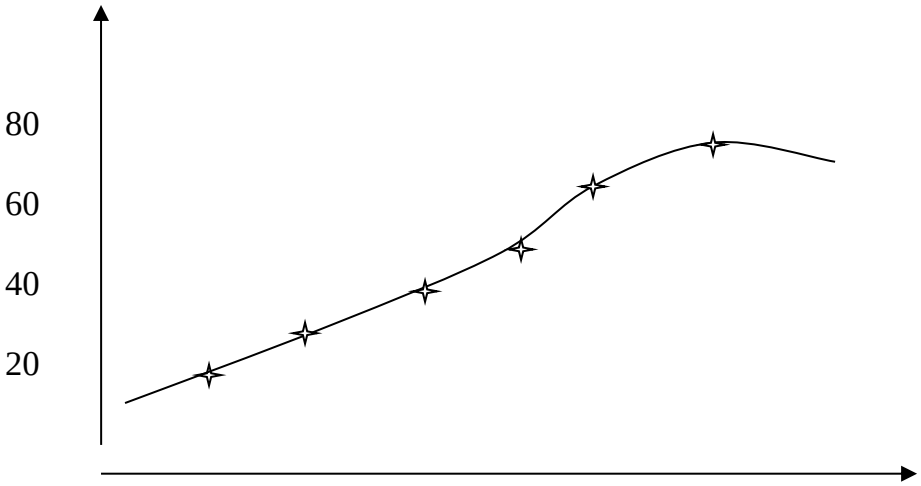
$\equiv \frac{29}{2}$

= 14.5

Do the same thing for all other classes. The result is in column 5.

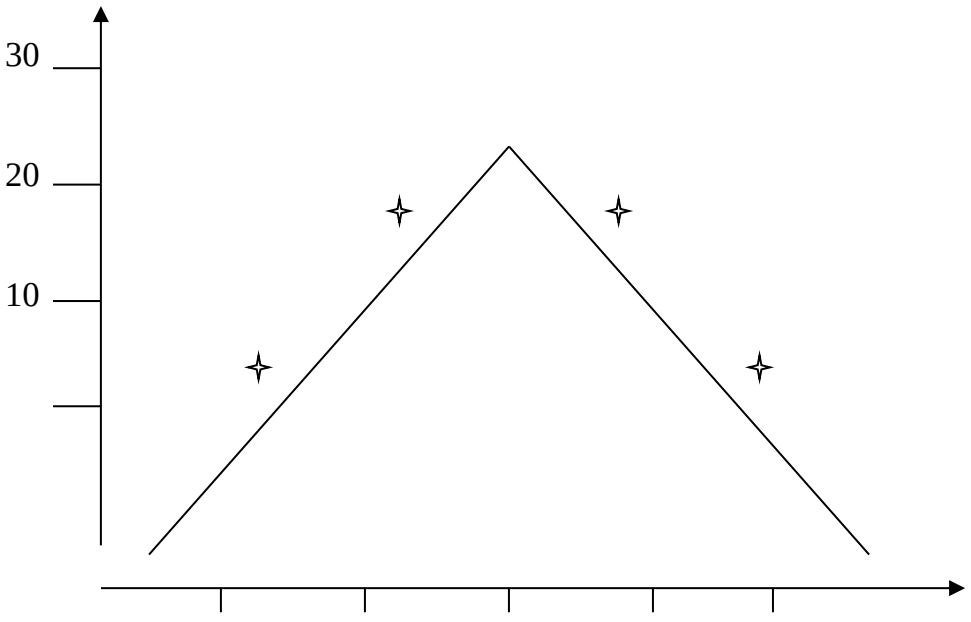
Column 3: **Ogive or cumulative frequency curve of table (iii)**

Cum. Freq



9.5 19.5 29.5 39.5 49.5 59.5

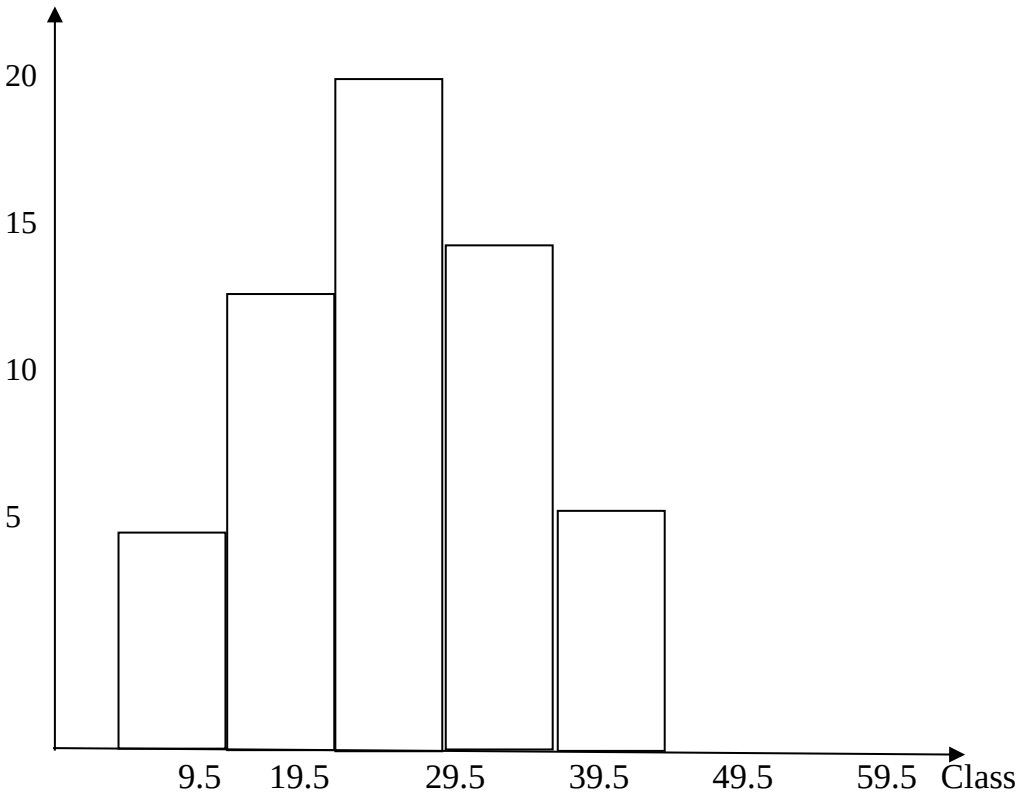
Freq. **The frequency Polygon of the table (iii) Class boundary column 4**



14.5 24.5 34.5 44.5 54.5 Class mark

Freq **Histogram of table (iii)**

25



Self-Assessment Exercises (SAEs) 2

Attempt these exercises to measure what you have learnt so far. This should not take you more than 40 minutes.

1. Consider the age (years) of 20 senators given below:
60, 50, 45, 45, 76, 50, 48, 60, 69, 60, 58, 45, 70, 72, 55, 40, 56, 60, 47, 53.
Use the raw data to present a frequency distribution table with the first-class interval of 40 – 44 years
2. The plot of cumulative frequencies against the class boundaries is known as
(a) Orgeve or less than curve.
(b) Ogive or less than curve.
(c) Ugivi or less than curve.
(d) Ogevu or less than curve.
3. Plotting the midpoints of the class against the corresponding frequency gives
(a) Frequency pentagon.
(b) Frequency hexagon.
(c) Frequency Polygon.
(d) Frequency heptagon

1.6 Summary

Descriptive statistics is nothing more than a fancy term for numbers that summarize a group of data. These data may be the number of arrests each customs officer makes, the amount of refuse collected by local government environmental officers, the number of fund raising events held by a political party in a year, the number of youth corps members assisting INEC in the conduct of election or the size of various governmental departments and ministries. Data are difficult to comprehend when they remain as 'raw data' or in an unsummarized or non-tabulated form.

We have, in the above discussion, explained that descriptive statistics summarize a body of data so that the data can be easily understood. Frequency distributions, percentage distributions and cumulative frequency distributions are three ways to condense raw or ungrouped data into a table that is easier to read and interpret. You are informed in the unit, for instance that a frequency distribution displays the number of times each value or range of values of a variable occurs. To add visual appeal and to increase interpretability, graphic presentations of data are used. Graphic techniques discussed in this unit include pictogram, bar chart, pie chart, histogram and ogive or cumulative frequency curve.

1.7 References/Further Reading

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Adeniyi Gbadegesin, Razaq Oloopoenia and Afeikhena Jerome, (2005), *Statistics for Social Sciences*, Ibadan:IUP

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

1.8 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

(a) PIE CHART

To calculate the pie chart , we first convert each country's legislative seats into degree º, i.e. unit of measuring an angle in a circle.

Therefore,

No of seat of a country **X** 360º
Total no of legislative seats

Thus,

Countries	Column 2	=Degree º	Percentage
Togo	50/1320 * 360º	13.64º	3.8%

Ghana	120/1320 *360º	32.73º	9.1%
Cameroon	100/1320* 360º	27.27º	7.6%
Nigeria	400/1320 *360º	109.09º	30.3%
South Africa	250/1320 *360º	68.18º	18.9%
Gabon	100/1320 *360º	27.27º	7.6%
Angola	300/1320 *360º	81.82º	22.7%
	Total=	**360º**	**100%**

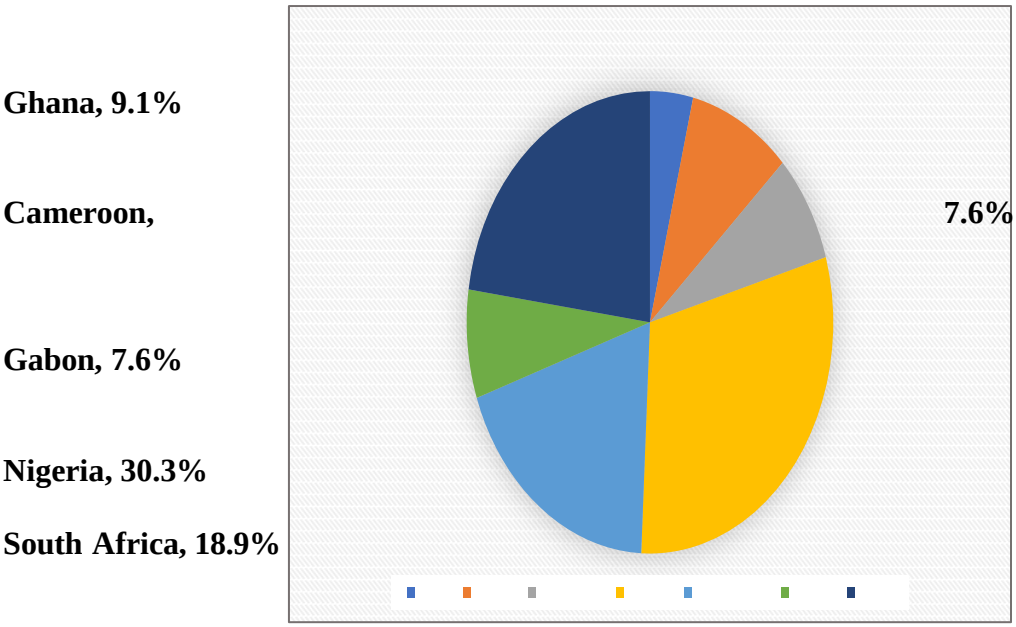
Round Up,

Countries	Column 2	=Degree º	Percentage(%)
Togo	50/1320 * 360º	14º	3.8
Ghana	120/1320 *360º	33º	7.6
Cameroon	100/1320* 360º	27º	9.1
Nigeria	400/1320 *360º	109º	30.3
South Africa	250/1320 *360º	68º	18.9
Gabon	100/1320 *360º	27º	7.6
Angola	300/1320 *360º	82º	22.7
	Total=	**360º**	

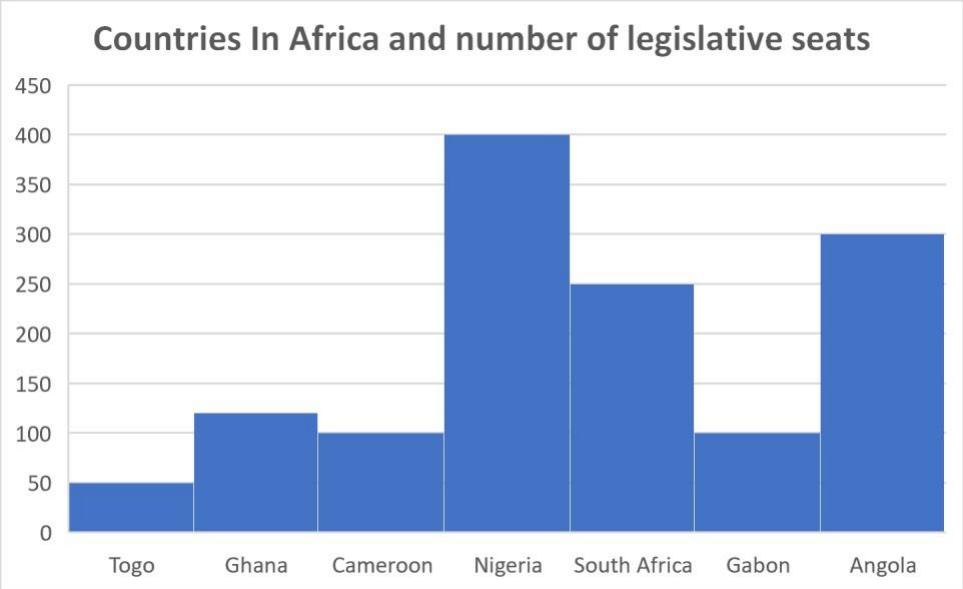
Chart Title

Togo, 3.8%

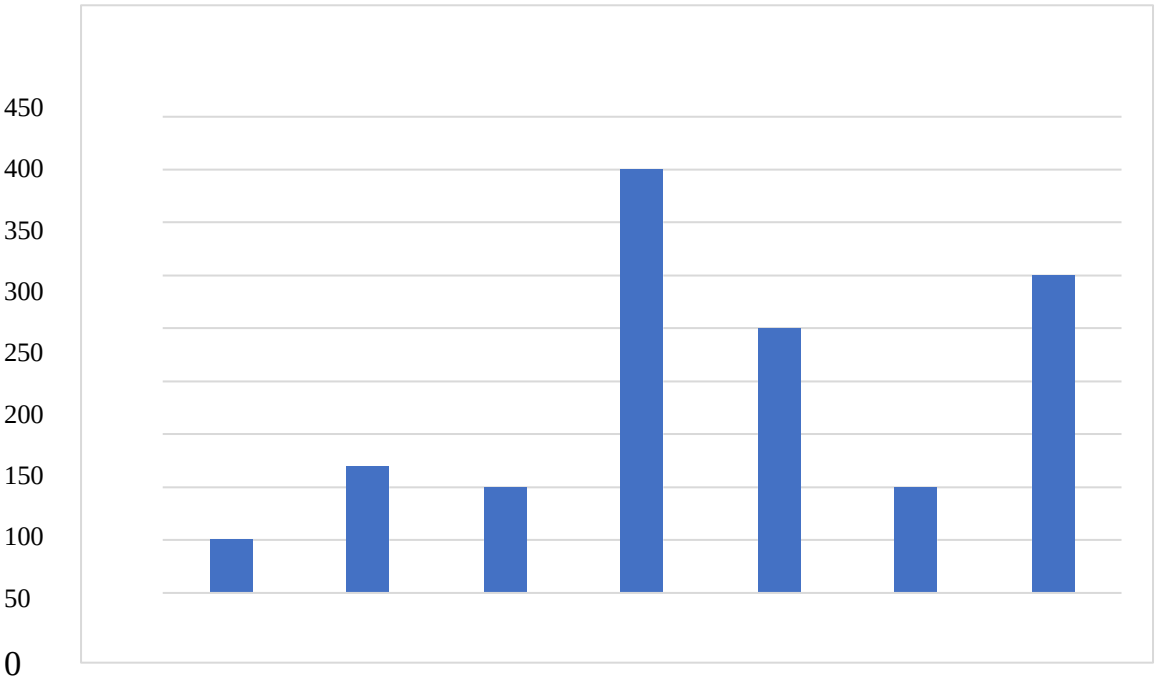
Angola, 22.7%


Ghana, 9.1%
Cameroon,
7.6%
Gabon, 7.6%
Nigeria, 30.3%
South Africa, 18.9%
Togo Ghana Cameroon Nigeria South Africa Gabon Angola

(b) HISTOGRAM


Countries In Africa and number of legislative seats
450
400
350
300
250
200
150
100
50
0
Togo
Ghana
Cameroon
Nigeria
South Africa
Gabon
Angola

(c) Bar Chart



Self-Assessment Exercises (SAEs) 2

1

Column 1	Column 2	Column 3	Column 4	Column 5
Age group (class)	No of Senators (f)	Cum. Freq (Cf)	Class boundary	Class mark
40 – 44	1	1	39.5 – 44.5	42
45 – 49	5	6	44.5 – 49.5	47
50 – 54	3	9	49.5 – 54.5	52
55 – 59	3	12	54.5 – 59.5	57
60 – 64	4	16	59.5 – 64.5	62
65 – 69	1	17	64.5 – 69.5	67
70 – 74	2	19	69.5 – 74.5	72
75 – 79	1	20	74.5 – 79.5	77

2. b
3. c

UNIT 2 MEASURES OF CENTRAL TENDENCY

Unit Structure

2.1 Introduction

The most commonly used descriptive statistics are measures of central tendency. Each measure of central tendency attempts to locate a typical value about which a distribution cluster. It represents an average character of the data. For instance, what was the average starting salary of the graduate students who graduated from Msc. Program of National Open University of Nigeria in year 2021? or On the average, how many federal government employees in Nigeria complained of IPPIS payment platform every month? The most common measures of central tendency are the mean, the mode and the median. As we have learnt in unit 1 of module three, there are grouped and ungrouped data. So also, there are different computation techniques for grouped characteristics and ungrouped characteristics. Therefore, for each concept we discuss here, we shall give the computation technique for ungrouped data separately and that of a grouped data separately. These three measures of central tendency shall be discussed in this unit.

2.2 Learning Outcomes

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

1. compute mode, median and mean as measures of central tendency
2. use and interpret these numerical measures accurately.
3. do better comparative studies using the descriptive statistics as a tool.

2.3 Measures of Central Tendency: Meaning

A measure of central tendency is a number or score or data value that represents the average in a group of data. Three different types of average are calculated and used most often. The first is the mode, or the data value that occurs with greatest frequency; the second is the median, which is the observation that falls exactly in the middle of the group; and the third is mean, which is the arithmetic average of the observations. This unit shows how to calculate the three measures of central tendency for both ungrouped (raw data) and grouped data (data that have been assembled into a frequency distribution) and discusses the use and interpretation of these measures. It concludes with a table that shows the relationship between the measures of central tendency and the different levels of measurement – nominal, ordinal and interval – you learned about in unit 4 of module 1.

2.4 The Mode

The mode is the category, which occurs most frequently. At the nominal data below, the mode (or modal category) is Christianity.

Religion	**No of respondent**
Islam	17
Mysticism	4
Christianity	25
Others	3

It will be wrong for you to say the mode is 25 (which is just the frequency of the modal category).

Another example: The mode of the set of figures 7, 3, 2, 1, 7, 5, 7 is what?

Answer: Mode is 7 because it occurs more frequently than all other figures. Note that if another figure (say 2) occurs as frequent as 7 in the example above, we say the set is bimodal (i.e. mode = 2 and 7). If we have more than two figures with the same frequency, we call it multi-modal distribution or set.

For a **grouped data**, we use the formula:

$$\text{Mode (mo)} = Li + \left[\frac{\Delta 1}{\Delta 1 + \Delta 2} \right] C$$

where Li = lower class boundary of the modal class.
$\Delta 1$ = modal frequency minus the frequency just before it. $\{f_1 - f_0\}$
$\Delta 2$ = Modal frequency minus the frequency just after it. $\{f_1 - f_2\}$
C = Class width (size).

2.5 Median

The median is the value that divides a set of observations into two equal halves. That is, 50% observation is above and below the median. The first task is to arrange the numbers in ascending or descending order. The next task is to locate the $\left(\frac{N+1}{2}\right)th$ position of the middle term

which occupies the position if there are n figures.
As an illustration, let us find the median of Nigerians Defence expenditure given earlier.

Step 1: Arrange in order: 5, 8, 9, 10, 11

Step 2: The Median occupies $\left(\frac{N+1}{2}\right)th$ position

i.e. $\left(\frac{5+1}{2}\right)th = \frac{6}{2} = 3^{rd}\ position$

Note that n is 5 because we have 5 Defence expenditure figures

Step 3: Locate the value that occupies the 3rd position in Step 1. The value is 9.

$\therefore$ The Medium defence expenditure is US $ 9 billion.

Another illustration, assuming in addition to those five years the Defence expenditure for year 2000 is 12 billion dollar. Find the median.
Step 1: Arrange 5, 8, 9, 10, 11, 12

Step 2: Compute $\frac{N+1}{2} = \frac{6+1}{2} = \frac{7}{2} = 3.5^{th}\ position$

Step 3: Locate (3.5)th position in step 1
It is after 9 but before 10. Here values 9 and 10 fall at the middle. What you do is simply add 9 and 10 together and divide the sum by 2.

$\therefore$ Median $= \frac{9+10}{2} = \frac{19}{2} = 9.5$

Therefore, Median Defence expenditure is $9.5b.

To calculate Median for a **grouped data** we use formula

$$\text{Median (Md)} = Li + \left[\frac{\frac{N}{2} - \sum f_{bm}}{f_m} \right] C$$

Where Li = lower class boundary of the median class.
N = total frequency.
$\sum f_{bm}$ = sum of all frequencies before the medium class.
fm = frequency of the median class.
C = class width.

Note that in order to locate the $\left(\frac{N+1}{2}\right)th$ position in this case, we shall use the cumulative frequency column (see example 4.3 below for the calculation procedure).

2.6 Mean

The mean, sometimes called the arithmetic mean of a set of numbers is the sum of the numbers divided by their total frequency. This is usually computed at the interval level by adding together each score and dividing the sum by the total number of cases. It is sometime represented by $\overline{X}$ or simply M.
For an ungrouped data,

$$\text{Mean } \overline{X} = \frac{\text{Sum of all values}}{\text{No of cases}}$$

$$= \frac{\sum xi}{N}$$

where N = total frequency (or $\sum f$)
$\sum$ = is just a symbol for "sum of."

For a grouped data, use

$$\text{Mean} = \overline{X} \frac{\sum fx}{\sum f}$$

Here the numerator says multiply each value of $\overline{X}$ by its corresponding frequency and sum all. The denominator says sum together all frequencies.

Example: From the given data below, calculate the average (mean) Defence expenditure of Nigeria over the five years.

Table 5.1

Year	1995	1996	1997	1998	1999
Defence Expenditure ($b)	5	8	10	9	11

Solution
Since this is an ungrouped data, we use this equation below:

Mean $(\overline{X}) = \frac{\sum Xi}{N}$

$$= \frac{5+8+10+9+11}{5}$$

$$= \frac{43}{5}$$

$$= 8.6$$

Thus, the mean Defence expenditure of Nigeria for the period is US $8.6 billion.

Now you should attempt example 2 before proceeding to the solution.

Example 2: The number of Private Bills passed in year 2001 by nineteen State Houses of Assembly in Northern Nigeria is as given below:

5, 4, 5, 4, 7, 8, 4, 5, 8, 7, 8, 5, 7, 8, 6, 8, 6, 5, 8

What is the average number of bills passed for that year in the North?
Have you solved it? Now proceed to the solution provided below and compare your steps and the answer.

Solution: Since each of the numbers appear more than once; we must multiply the numbers by their respective frequencies (i.e. number of times they appear).
Thus, using equation 4.2

Mean $\overline{X} = \frac{\sum fx}{\sum f}$

Average bills passed $= \frac{5(4)+4(3)+7(3)+8(6)+6(2)}{4+3+3+6+2}$

$$= \frac{20+12+21+48+12}{19}$$

$$= \frac{113}{19}$$

$$= 5.947 -$$

i.e. Approximately an average of six Private Bills were passed in year 2001.

Let us consider this exercise: A referendum was conducted in all the 20 wards of the Bakassi Peninsula to determine where the people would like to belong {Nigeria or Cameroon}. The number of participants in all the wards are a given below:

43	49	39	25	48	54	43	42	45	33
47	51	60	68	34	38	59	31	27	55

Using a class interval of 21-30 etc.

1. Prepare a frequency distribution table.
2. Calculate the mean, mode, median participation.
3. Which of the values in (ii) above best describe the number of participants?

Solution

1	2	3	4	5	6	7
Class	Tally	F	Class mark (x)	Fx	Class boundary	Cumm freq.
21-30	II	2	25.5	51	20.5-30.5	2
31-40	IIII	5	35.5	177.5	30.5-40.5	7
41-50	IIII II	7	45.5	318.5	40.5-50.5	14
51-60	IIII	5	55.5	277.5	50.5-60.5	19
61-70	I	1	65.5	65.5	60.5-70.5	20
	20	20		890		

2. To calculate the Mean, we use the equation below
Since this is a grouped data,

Mean $\overline{X}$ = $\frac{\sum fx}{\sum f}$ i.e. $\frac{\text{sum of column 5}}{\text{sum of column 3}}$

$$= \frac{890}{20}$$

$$= 44.5$$

Thus, the average number of participants at the referendum is 45 people.
To calculate the Mode, we shall use this equation below:

$$\text{Mode} = Li + \left[\frac{\Delta 1}{\Delta 1 + \Delta 2}\right] C$$

The modal class is the class 41-50 because it has the highest frequency.

Therefore Li = 40.5 (see column 6)

$\Delta 1$	=	8 – 5	=	3	(see definition of $\Delta 1$) $\{f_1 - f_0\}$
$\Delta 2$	=	8 – 4	=	4	(see definition of $\Delta 2$) $\{f_1 - f_2\}$
C	=	30.5 – 20.5	=	10	(see definition of C)

$$\therefore Mode = 40.5 + \left[\frac{3}{3+4}\right] x10$$

$$= 40.5 + \frac{30}{7}$$

$$= 40.5 + 4.285$$

$$= 44.785$$

Thus, the most frequent number of participants is approximately 45 people.
To calculate the Median, we shall use equation (4.3). Meanwhile, the median occupies the $\left(\frac{N+1}{2}\right)th$ position

i.e. $= \left(\frac{20+1}{2}\right)th = \frac{21}{2}$ 10.5^{th} position.

Examining the cumulative frequency column (7), we found that 10.5^{th} position falls in the 40.5 – 50.5 class (which is the median class).

$$\therefore \text{Median} = Li + \left[\frac{\frac{N}{2} - \sum f_{bm}}{fm}\right] C$$

$$= 40.5 + \left[\frac{\frac{20}{2} - 7}{8} \right] 10$$

$$= 40.5 + \left[\frac{3}{8} \right] \text{X } 10$$

$$= 40.5 + \frac{30}{8}$$

= 40.5 + 3.75

= 44.25 approximately 44 people or participants.

3. To answer this question, we should bear in mind that the Mean is the only measure of central tendency, which takes all the observations into account in its computation. We are only fortunate in this example that the three measures happen to fall in the same class 41-50. This may not always be the case, the Mean 45 participants is the figure that best describe the distribution.

It is important for you to know the appropriate level of measurement suitable for each measure of central tendency. Check the table below.

Level of Measurement/ Measures of central Tendency	Nominal	Ordinal	Interval
Mode	YES	YES	YES
Median		YES	YES
Mean			YES

Note: Most of the data we deal with in Political Science are statistically measured at nominal, ordinal and interval levels.

Self -Assessment Exercises (SAE) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 45 minutes.

1. The total annual budget for 2001 was ₦500m. Education was allocated only ₦25m. What percentage of total budget goes to education?
(a) 50% (b) 0.5% (c) 500% (d) 5%

2. 16 million people voted in the 1993 Presidential Election while 24 million people voted in the 1999 Presidential Election. What is the percentage change?
(a) 5%(b) 50% (c) 0.5% (d) 500%

3. Calculate the average income of four students whose individual income are 5, 8, 10, 9 (in thousand naira).
(a) ₦80, 000 (b) ₦8, 000 (c) ₦800 (d) ₦800, 000

4. The value that appears most frequently in a distribution is known as the
(a) Mode (b) Median (c) Mean (d) Variance

5. One single value is typical of its group is known as
(a) Mode (b) Median (c) Mean (d) Variance

6. The number of votes cast for a group of candidates in party primaries is given as:
65, 65, 40, 65, 50, 80, 48, 59, 79, 85, 85, 72, 70, 69, 45, 65, 76, 77

(a) Find the mean, mode and the median. Which of them is suitable as a measure of central tendency for this distribution.
(b) Present a frequency table (with an interval size of 10, and the first interval (40- 49)
(c) Compute the mode, median and mean of the grouped data.

2.7 Summary

The measures of central tendency give a value around which other values within the distribution clusters. The three measures of central tendency include the mode, the median and the Mean.

In this unit, you learned how to calculate the measures of central tendency – the mode, the median and the mean. The mean is the typical value that considers all values in the distribution for its computation. The mode is the score that occurs most frequently while the median divides the distribution into two equal halves. Moreover, you were shown through examples that the mean is the best measure of central

tendency in most cases, when you have a normal distribution. In the next lecture, we shall discuss the remaining aspect of Descriptive Statistics i.e. the measures of variation.

2.8 References/ Further Reading

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers Pvt Limited.

Kitchens, L.J. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference* (2nd Ed). USA: Duxburg Press.

Spiegel, M.R. and L.J. Stephens. (2000) *Introduction to Probability and Statistics Schaum's Outline Series* (3rd Ed.). New York: McGraw-Hill.

2.9 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. d
2. b.
3. b
4. a
5. c
6

(a) Find the raw mean, median, and mode

Raw mean =64 participants, raw median = 65 participants, raw mode = 65 participants

(b) Present a frequency table (with an interval size of 10, and the first interval 40-49)

Class	F	Class mark (x)	Fx	Class boundary	Cumm freq.
40-49	5	44.5	222.5	39.5-49.5	5
50-59	2	54.5	109	49.5-59.5	7
60-69	5	64.5	322.5	59.5-69.5	12
70-79	5	74.5	372.5	69.5-79.5	17
80-89	3	84.5	253.5	79.5-89.5	20
	20		1280		

(c) Compute the group mean, mode and median and compare with answers in (a)

Group Mean = 64 participants

Group Mode = Bimodal 70 participants

Group Median = 66 participants

UNIT 3 DESCRIBING DATA WITH AVERAGES

Unit Structure

3.1 Introduction

As shown in the last unit, tables and graphs of frequency distributions are important points of departure when attempting to describe data. However, more precise summaries, such as averages, provide additional valuable information. A young politician might strengthen his resolve to contest in the next general elections upon hearing that, on average, the ages of contestants for the 2023 House of Representatives election in Nigeria is three years less than his present age. Averages consist of numbers (or words) about which the data are, in some sense, centered. We referred to them as measures of central tendency in the last unit. It is important for you to note that the several types of average yield numbers or words that attempt to describe, most generally, the middle or typical value for a distribution. In this unit, the focus is on the types of data and which measures of central tendency—the mode, median, and mean is appropriate to describe them. Each of these measures has its special uses, but the mean is the most important average in both descriptive and inferential statistics.

3.2 Learning Outcomes

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

- explain what describing data entails in political science research
- identify types of data.
- determine appropriate average for describing quantitative data.
- describe qualitative and ranked data with relevant average(s}

3.3 Describing Data in Political Science Research

Descriptive statistics is essentially describing the data collected through methods such as graphical representations, measures of central tendency and measures of variability. It summarizes the data in a meaningful way which enables us to generate insights or meanings from it. As discussed in unit 2 of Module one, the primary goal of descriptive statistics is to provide a clear and concise summary of the data, enabling political researchers or analysts to gain insights and understand patterns, trends, and distributions within the dataset. This summary typically includes measures such as central tendency (e.g., mean, median, mode), dispersion (e.g., range, variance, standard deviation), and the skewness of the distribution.

By employing descriptive statistics, political researchers can effectively summarize and communicate the key characteristics of a dataset, facilitating a better understanding of political data and providing a foundation for further statistical analysis or decision-making processes.

3.4 Types of Data

It is important to note that any statistical analysis is performed on data. Statistical data, in the context of political research, refer to a collection of actual observations or scores in a survey or an experiment. Thus, the precise form of a statistical analysis in most cases, depends on whether data are qualitative, ranked or quantitative. Let us discuss them one after the other.

A. Qualitative:

Qualitative data are not-numerical which can be based on methods such as interviews, ratings of civil servants in promotion examination etc. It can be nominal and ordinal, where nominal data does not contains any order such as the gender, marital status, while ordinal data has a particular order such as performance of a political party in the general election, income status of minister nominees. Generally, qualitative data consist of words (True or False), letters (T or F), or numerical codes (0 or 1) that represent a class or category.

B. Ranked

Ranked data consist of numbers (1st, 2nd, . . . 60th place) that represent relative standing within a group. It is different from qualitative data because we can rank the words, letters or numeric codes used in describing our data. For example, you may want to describe citizens'

responses concerning Abuja Urban Mass Transit System with Strongly Approve, Approve, Neutral, Disapprove and Strongly Disapprove. Also, we can rank the number of visit to University of Ibadan Zoological Garden by public functionaries in year 2022 using numeric codes 0 to 5. 0 represents no visit, while 5 represents five visits. In this sense, you can rank either 0 to 5 in ascending order or 5 to 0 in descending order.

C. Quantitative

Quantitative data consist of numbers (e.g. Legislative Aides allowances of 250, 190, . . . 385 thousands naira) that represent an amount or a count. Quantitative data is in numeric form, which can be discrete that includes finite numerical values or continuous which also takes fractional values apart from finite values. For instance, the number of legislators in US Congress can only take finite values, so it is a discrete variable, while the cost of an electioneering campaign is a continuous variable.

To determine the type of data, focus on a single observation in any collection of observations. For example, the ages reported by 10 presidential candidates in the table 3.1 below are quantitative data, since any single observation, such as 50 years, represents an amount of age. If the ages in Table 3.1 had been replaced with ranks, beginning with a rank of 1 for the youngest candidate of 35 years and ending with a rank of 10 for the oldest of 70 years, these numbers would have been ranked data, since any single observation represents not an amount, but only relative standing within the group of 10.

Presidential Candidates	**Age (years)**
LPP	45
BPP	36
EPP	70
MPP	43
JPP	35
KPP	60
DPP	67
OPP	55
NPP	56
WPP	48

Table 3. 1

Self-Assessment Exercises (SAEs) 1
Indicate whether each of the following terms is qualitative (because it is a word, letter, or numerical code representing a class or category);

ranked (because it is a number representing relative standing); or quantitative (because it is a number representing an amount or a count).

(a) Party Identification
(b) Gender
(c) Income in (£)
(d) State constituency
(e) Number of Votes Cast
(f) Net worth of selected Governors (naira)
(g) |Fourth-place finish

3.5 Appropriate Average for Describing Quantitative Data

As pointed out in the last unit, the Mean plays the role of a balance point because it describes the single point of equilibrium at which, once all scores have been expressed as deviations from the mean, those above the mean counterbalance those below the mean. You can appreciate, therefore, why a change in the value of a single score produces a change in the value of the mean for the entire distribution. The mean reflects the values of all scores, not just those that are middle ranked (as with the median), or those that occur most frequently (as with the mode). In extreme cases, the mean describes the central tendency of a distribution only in the more abstract sense of being the balance point of the distribution.

To determine which average or measure of central tendency to be used in describing our data, we need to determine whether the distribution of data is skewed or not. When a distribution of scores is not too skewed, the values of the mode, median, and mean are similar, and any of them can be used to describe the central tendency of the distribution.

Ideally, when a distribution is skewed, we should report both the mean and the median. Appreciable differences between the values of the mean and median signal the presence of a skewed distribution. If the mean exceeds the median, the underlying distribution is positively skewed because of one or more scores with relatively large values. On the other hand, if the median exceeds the mean, the underlying distribution is negatively skewed because of one or more scores with relatively small values.

You need to take note of the point that the mean sometimes fails to describe the typical or middle-ranked value of a distribution. When this is observed, it should be used in conjunction with another average, such as the median. In the long run, however, the mean is the single most preferred average for quantitative data. In the subsequent units, it will be used almost exclusively. In the next unit, you would see how the mean

serves as a key component in an important statistical measure, the standard deviation. Later, in our discussion on inferential statistics, you would see how it emerges as a significant measure to be used when generalizing beyond actual scores in surveys and experiments.

3.6 Averages for Qualitative and Ranked Data

So far, we have been talking about quantitative data for which, in principle, all three averages can be used. But when the data are qualitative, your choice among averages is restricted. Generally, the mode always can be used with qualitative data. For example, EPP qualifies as the modal or most typical response for the Parties of Presidential Candidates in Table 3.1 above. The median can be used whenever it is possible to order qualitative data from least to most because the level of measurement is ordinal. It's easiest to determine the median class for ordered qualitative data by using relative frequencies. Otherwise, first convert regular frequencies to relative frequencies. Cumulate the relative frequencies, working up from the bottom of the distribution, until the cumulative percentage first equals or exceeds 50 percent. Since the corresponding class includes the median and, roughly speaking, splits the distribution into an upper and a lower half, it is designated as the median or middle-ranked class.

It is important for you to note that when you are finding the median for ordered qualitative data you should avoid a common error that identifies the median simply with the middle or two middlemost classes without regard to the cumulative relative frequencies and the location of the 50th percentile. In other words, do not treat the various classes as though they have the same frequencies when they actually have different frequencies. Also, it would not be appropriate to report a median for unordered qualitative data with nominal measurement, such as the ancestries of Africans. Nor would it be appropriate.

When the data consist of a series of ranks, with its ordinal level of measurement, the median rank always can be obtained. It's simply the middlemost or average of the two middlemost ranks. For example, consider the Table 3.1.1 below displaying the ages of 10 presidential candidates and ranks for the ages, beginning with rank 1 for the youngest candidate (35 years) and ending with rank 10 for the oldest candidate (70 years). Recalling how to find the median when there is an even number of scores, as described in the example on defence expenditure on page 83, assign the average of the two middlemost ranks (5th and 6th), that is, 51.5, as the median rank.

Presidential Candidates	Age (years)	Rank
LPP	45	4
BPP	36	2
EPP	70	10
MPP	43	3
JPP	35	1
KPP	60	8
DPP	67	9
OPP	55	6
NPP	56	7
WPP	48	5

It is not necessary to discuss the mean and modal ranks because they tend not to be very informative.

Self -Assessment Exercises (SAE) 2

Indicate whether the following skewed distributions are positively skewed because the mean exceeds the median or negatively skewed because the median exceeds the mean.
(a) a distribution of test scores of five students on an easy test in POL 803 (over 100): 38, 49, 72, 65, and 83
(b) a distribution of ages of 10 voters (years): 21, 18, 19, 22, 18, 22, 60, 55, 68, and 52
(c) a distribution of loose change carried by five classmates; £0.5, £0.6, £0.7, £3 and £4
(d) a distribution of the sizes of crowds in attendance at a campaign rally; 220, 210, 220, 240, and 230

3.7 Summary

As discussed in this unit, the mode equals the value of the most frequently occurring or typical score and the median equals the value of the middle-ranked score (or scores). It is also shown in the unit that the value of the mean, whether defined for a sample or for a population, is found by summing all the scores and then dividing by the number of scores in the sample or population. It always describes the balance point of a distribution, that is, the single point about which the sum of positive deviations equals the sum of negative deviations.

More importantly, you were told that when frequency distributions are not skewed, the values of all three averages tend to be similar and equally representative of the central tendencies within the distributions. When frequency distributions are skewed, the values of the three averages differ appreciably, with the mean being particularly sensitive to

extreme scores. Ideally, in this case, you were informed to report both the mean and the median. The mean is the preferred average for quantitative data and will be used almost exclusively in subsequent units. It reappears as a key component in other statistical measures and as a well-documented measure in surveys and experiments.

Only the mode can be used with all qualitative data. If qualitative data can be ordered from least to most because the level of measurement is ordinal, the median also can be used. The median is the preferred average for ranked data.

3.8 References/ Further Reading

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

Witte Roberts S. and John S. Witte, 2017, Statistics, (Eleventh ed.) Hoboken, NJ: John Wiley & Sons, Inc.,

3 .9 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

(a) Qualitative
(b) Qualitative
(c) Quantitative
(d) Qualitative
(e) Quantitative
(f) Quantitative
(g) Ranked

Answers to SAEs 2

(a) Negatively skewed
(b) Positively skewed
(c) Positively skewed
(d) Positively skewed

UNIT 4 MEASURES OF DISPERSION OR VARIABILITY

Unit Structure

4.1 Introduction

The last unit discussed measures of central tendency. This unit is the concluding part of our discussion on Descriptive Statistics that we started in the last unit. We shall examine the degree to which a set of figures deviate or spread about their average. The Mean Deviation, the Range and the Variance are some of the measures of dispersion that will be discussed in this unit.

4.2 Lecture Outcomes

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

1. define and calculate the mean deviation.
2. define and calculate the range.
3. define and calculate the variance.
4. deduce the standard deviation from the variance.
5. interpret the values of each of the above measures.

4.3 Measures of Dispersion or Variability: Meaning

Dispersion or variability refers to the spread of the values around the central tendency. These are ways of summarizing a group of data by describing how spread out the scores is. If we have a set of values like the data on Nigerian's Defence expenditure, some of the values will be close to group's average while others will be far away from groups average (in terms of magnitude). The degree to which these values tend to spread about (or deviate from) the mean value is what we call variation, variability or dispersion. In this unit, we shall examine the

following measures of dispersion, the range, the mean, deviation, the variance and the standard deviation.

4.4 Range

The range is simply the difference between the highest score and the lowest score in a distribution. When all scores in the distribution have the same value; the range is zero; that is no variability. The higher the value of the range, the farther apart is the extreme scores of the distribution. Since the range only uses the largest and smallest values, it is greatly affected by extreme values, that is - it is not resistant to change. In essence, a weakness of the range is that it considers only the two extreme cases in its computation.

Example: Find the range of the following income distribution in a private establishment.
₦50,000, ₦20,000, ₦150,000, ₦32,000 and ₦5,000.

Solution:
Income Range = Highest – Lowest income
= ₦150,000 – ₦5,000
= ₦145,000
This shows that there is a high degree of income inequality in the establishment.

For grouped data, the Range is:
Range = Upper limit of the last class boundary – Lower limit of the First class boundary. Consider the table below and find the Range of the scores.

Class	Tally	F	Class mark (x)	Fx	Class boundary
21-30	II	2	25.5	51	20.5-30.5
31-40	IIII	5	35.5	177.5	30.5-40.5
41-50	IIII III	8	45.5	364	40.5-50.5
51-60	IIII	4	55.5	222	50.5-60.5
61-70	I	1	65.5	65.5	60.5-70.5

Solution
The Upper limit of the Last Class Boundary = 70.5
The Lower limit of the First Class Boundary = 20.5
Therefore, Range = 70.5 – 20.5

= 50

4.5 The Mean Deviation, Variance, Standard Deviation and Coefficient of Variation

1. *The Mean Deviation*

The mean deviation is simply the average distance from the mean of all the cases in a distribution.

Now that you are becoming familiar with some of the symbols in this course, we shall introduce some new formulas, which are quite similar to those encountered in the previous. Before _then let me remind you that symbol X stands for numerical value of a variable, symbol X stands for mean or average value of a variable X, N stands for the total number of cases in the distribution. $\sum$ implies summing up.

For an ungrouped data,

Mean deviation (MD) $= \dfrac{\sum \left| X - \overline{X} \right|}{N}$

For computation purposes, use the format below to arrange your work:

1	2	3
X	$\overline{X}$	$\left\| X - \overline{X} \right\|$

1. Put the values of the variables.
2. Insert the mean (which is a constant).
3. Subtract the mean from each value (ignoring negative signs).

The two vertical lines in column 3 means absolute value i.e. ignore negative sign.

For a grouped data, use:

Mean Deviation {MD}= $\sum f \dfrac{\left| X - \overline{X} \right|}{N}$

And the computation table should be arranged as follows:

Class	Class Mark (X)	f	$\overline{X}$	$\lvert X - \overline{X} \rvert$	$f\lvert X - \overline{X} \rvert$

The symbols are as explained above in the ungrouped case. Note that the MD retains the original unit used in measuring the data. We shall take an example to illustrate these concepts soon. Before then, let us discuss the variance and standard deviation.

2. ***The Variance***

The variance is also an index that reflects the degree of variability or the average distance from the mean of all cases in an interval level distribution. However, it relies on the squared distance between X and X. Its derivative, the standard deviation, converts the squared unit to a linear one. The variance is computed by:

Variance { ð²}= $\frac{\sum(X - \overline{X})^2}{N}$ for ungrouped data

Variance { ð² } = $\frac{\sum f\{X - \overline{X}\}^2}{N}$ for a grouped data

3. ***The Standard Deviation***: is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range. It shows the relation that set of scores has to the mean of the sample. The Standard Deviation, which is a derivative of the variance, is the square root of the variance i.e.

Standard Deviation { σ} = $\sqrt{\text{Variance}}$

= $\sqrt{\frac{\sum(X - \overline{X})^2}{N}}$ ungrouped data

$\sqrt{\frac{(X - \overline{X})^2}{}}$

$= \quad \frac{\sum f}{N}$ grouped data

Like the mean deviation, the standard deviation is relatively large if the cases are widely dispersed about mean. If all the cases take on the same value, the standard deviation is zero.

4. ***Coefficient of Variation CV*** $= \sigma / \overline{X}$

It evaluates the proportionality between a mean and a standard deviation.

If CV = 0, there is no variation
If CV = 1, there is a good deal of variation
Illustration:

1. For our Defence expenditure data, 5, 8, 10, 9, 11
Calculate the (i) Mean Deviation.
(ii) Standard Deviation.
(iii)The Variance.
(iv) Coefficient of Variation.

Solution
1. Since this is an ungrouped data, for
(i) Mean deviation, use equation below

Thus MD $= \frac{\sum \left| X - \overline{X} \right|}{N}$

But $\overline{X} = \frac{5+8+10+9+11}{5}$

$= \frac{43}{5}$

$= 8.6$

Therefore MD $= \underline{(5 - 58.6) + (8-8.6) + (10-8.6) + (9-8.6) + (11-8.6)}$

5

$= \underline{3.6 + 0.6 + 1.4 + 0.4 + 2.4}$

5

$= \frac{8.4}{5}$

$= 1.68$
i.e. US $ 1.68 billion

(ii) To get the variance, we use this equation

$$\sigma^2 = \frac{\sum\left(X - \overline{X}\right)^2}{N}$$

$$= \frac{(5\text{-}8.6)^2 + (8\text{-}8.6)^2 + (10\text{-}8.6)^2 + (9\text{-}8.6)^2 + (11\text{-}8.6^2)}{5}$$

$$= \frac{3.6^2 + 0.6^2 + 1.96^2 + 0.16^2 + 5.76^2}{5}$$

$$= \frac{12.96 + 0.36 + 1.96 + 0.16 + 5.76}{5}$$

$$= \frac{21.2}{5}$$

$$= 4.24$$

i.e. US $ 4.24 billion

(iii) Standard Deviation σ = $\sqrt{\text{Variance}}$

= $\sqrt{4.24}$

= 2.059

= US $ 2.06 billion

(iv) Coefficient of Variation

CV = $\frac{\sigma}{\overline{X}}$

= $\frac{2.06}{8.6}$

= 0.239

= 0.24

The social information here is that since the coefficient variation is very small in magnitude, we may conclude that there exists only a little variation in the Defence expenditure within the five years. The values are indeed close to the average.

Self-Assessment Exercises (SAEs) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 30 minutes

1. The difference between the highest and the lowest score in a distribution is called

(a) Mid deviation (b) Range
(c) Standard deviation (d) All of the above

2. When a group of scores have a large variance, then the scores are said to be

(a) Far apart from their median.
(b) Far apart from their mode.
(c) Far apart from their mean.
(d) None of the above.

3. The measure that evaluates the proportionality between a mean and a standard deviation is the

(a) Coefficient of mode.
(b) Coefficient of variation.
(c) Coefficient of proportionality.
(d) Coefficient of mean deviation.

4. The value of the coefficient of variation always lie between

(a) 0 and 1 (b) -1 and +1
(c) -1 and 0 (d) all of the above

5. The number of votes cast for a group of candidates in party primaries is given as:

65, 65, 40, 65, 50, 80, 48, 59, 79, 85, 85, 72, 70, 69, 45, 65, 76, 77

Use the information to calculate:

(a) Standard deviation
(b) Variance
(c) Coefficient of variation and present the social information it conveys

4.6 Summary

The measures of dispersion or variability gives us information about how spread out are data values in a distribution. These measures include range, mean deviation, variance, standard deviation and coefficient of variation. This unit has attempted to show what the measures of dispersion depict. They show the degree to which the figures tend to spread about their average. You were shown in the unit that the smaller the coefficient of variation, the lower the degree of variability.

4.7 References/Further Readings

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers Pvt Limited.

Harry Frank, and S.C. Althoen (1994) *STATISTICS: Concepts and Applications.* (Cambridge) Cambridge University Press.

Kitchens, L.J. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference* (2nd Ed). USA: Duxburg Press.

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Spiegel, M.R. and L.J. Stephens (2000) *Introduction to Probability and Statistics* (3rd Ed), *Schaum's Outline Series*. New York: McGraw-Hill.

4.8 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. b
2. c
3. b
4. b
5.

Class	F	Class mark (x)	Fx	Class boundary	Cumm freq.
40-49	5	44.5	222.5	39.5-49.5	5
50-59	2	54.5	109	49.5-59.5	7
60-69	5	64.5	322.5	59.5-69.5	12
70-79	5	74.5	372.5	69.5-79.5	17
80-89	3	84.5	253.5	79.5-89.5	20
	20		1280		

(a) Compute the group standard deviation and interpret the result
SD = 13.96
(b) Variance = $SD^2 = 13.96^2 = 194.8816$
(e) Determine the coefficient of variation (CV) and present the social information it conveys
Coefficient of variation = 0.2 or 20% which means little variation

UNIT 5 DESCRIBING VARIABILITY: QUANTITATIVE AND QUALITATIVE/RANKED DATA

Unit Structure

5.1 Introduction

Although averages are important (as discussed in units 2 and 3), but they tell only part of the story. In other words a measure of central tendency alone can be misleading. For example, two countries with the same median family income are very different if one has extremes of wealth and poverty and the other has little variation among families. We are interested in the spread or variability of incomes as well as their centres. The simplest relevant numerical description of a distribution consists of both a measure of centre and a measure of spread. In this unit, the importance of variability and how to describe quantitative and qualitative/ranked data with measures of variability or spread will be discussed.

5.2 Learning Outcomes

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

1. explain the importance of variability.
2. describe the spread of quantitative data.
3. describe the variability of qualitative data

5.3 Importance of Variability

Statistics flourishes because we live in a world of variability; no two cities are identical, and a few are really far out. When summarizing a set of data, we specify not only measures of central tendency, such as the

mean, but also measures of variability or spread, that is, measures of the amount by which scores are dispersed or scattered in a distribution. Recalling that measures of variability, include the range, the interquartile range, the variance, and most important, the standard deviation.

Variability assumes a critical role in an analysis of research results. For example, a researcher might ask: Does fitness training improve, on average, the scores of depressed contestants on a mental-wellness test? To answer this question, depressed contestants are randomly assigned to two groups, fitness training is given to one group, and wellness scores are obtained for both groups. Let's assume that the mean wellness score is larger for the group with fitness training. Is the observed mean difference between the two groups real or merely transitory? This decision depends not only on the size of the mean difference between the two groups but also on the inevitable variabilities of individual scores within each group.

5.4 Describing Spread of Quantitative Data

Range, the interquartile range and standard deviation are three major numerical measures of spread or variability. The range is the spread of all the data, and the interquartile range is the spread of the middle half of the data. The standard deviation is more complicated because it is based on the average of the squared distances of the observations from the mean. As in the case of measuring averages, these measures of spread tell us different things and behave differently.

Let me remind you that the range of a distribution is the distance between the largest and the smallest individual values or observations. The quartiles mark the middle half of the data. The **lower quartile** (Q1) is the 25% point, the value that is larger than one-quarter of the observations. The **upper quartile** (Q3) is the 75% point, the value that is larger than three-quarters of the observations. The second quartile is the median, which is larger than half of the observations. Thus, the **interquartile range** (IQR) is the distance between the upper and lower quartiles: IQR = Q3 – Q1. It is the most important spinoff of the range, and it is simply the range for the middle 50 percent of the scores. More specifically, the IQR equals the distance between the third quartile (or 75th percentile) and the first quartile (or 25^{th} percentile), that is, after the highest quarter (or top 25 percent) and the lowest quarter (or bottom 25 percent) have been trimmed from the original set of scores. Since most distributions are spread more widely in their extremities than their middle, the IQR tends to be less than half the size of the range. When you divide the interquartile range into two equal halves, you have **semi-interquartile range**.

Consider this example, The number of votes cast for a group of candidates in party primaries is given as:
65, 65, 40, 65, 50, 80, 48, 59, 79, 85, 85, 72, 70, 69, 45, 65, 76, 77.
Calculate the range, interquartile range and semi-interquartile range.
To obtain these values, arrange in ascending order as follows:
40, 45, 48, 50, 59, 65, 65, 65, 65, 69, 70, 72, 76, 77, 79, 80, 85, 85
Range = 85-40 = 45 votes

The first quartile (Q1) lies between the 4^{th} and 5^{th} score = 54.5
The third quartile (Q3) lies between 13^{th} and 14^{th} score = 76.5
Therefore inter-quartile range (IQR) = Q3 – Q1 = 76.5 – 54.5 = 22 votes
The semi –interquartile range = 22/2 = 11 votes

However, it should be noted that the range, like its companion the midrange is extremely sensitive to outlying observations. It says nothing about the spread of the distribution. The quartiles, like the median, are resistant. They ignore the tails of the distribution. Thus, we can give information about both the body and the tails of a distribution by reporting the five-number summary. The five-number summary of a data set consists of the smallest observation, the lower quartile, the median, the upper quartile, and the largest observation, written in order from smallest to largest.

The standard deviation (s) measures spread by looking at how far the observations are for their mean. It should be used only when the mean is chosen as the measure of central tendency. Standard deviation (s) = 0 only when there is no spread. This happens only when all observations have the same value. Otherwise (s) > 0. As the observations become more spread out about their means, (s) gets larger.

Self-Assessment Exercises (SAEs) 1

Attempt the exercise below to measure what you have learnt so far. This should not take you more than 10 minutes

1. Calculate the range, median, interquartile range and semi-interquartile range of this distribution of scores
20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

5.5 Describing Variability of Qualitative and Ranked Data

For qualitative or nominal data, it is obvious that measures of variability are virtually non-existent. It is probably adequate to note merely whether scores are evenly divided among the various classes (maximum variability), unevenly divided among the various classes (intermediate variability), or concentrated mostly in one class (minimum variability). For instance, if the ethnic composition of the residents of a region is

about evenly divided among several groups, the variability with respect to ethnic groups is maximum; there is considerable heterogeneity. At the other extreme, if almost all the residents are concentrated in a single ethnic group, the variability will be minimum; there is little heterogeneity. If the ethnic composition falls between these two extreme as a result of an uneven division among several large ethnic groups, the variability will be intermediate, as is true of many cities and counties in the United States of America.

If qualitative data can be ordered because measurement is ordinal (or if the data are ranked), then it is appropriate to describe variability by identifying extreme scores (or ranks). For instance, the active membership of an officers' club in the military might include no one with a rank below first lieutenant or above brigadier general.

5.6 Summary

As discussed in this unit, measures of variability reflect the amount by which observations are dispersed or scattered in a distribution. These measures assume a key role in the analysis of research results. The simplest measure of variability, the range, is readily calculated and understood, but it has its shortcomings. The five member summary provides information that the range fails to give. These in symbols, are Minimum, Q1, M (median), Q3 and Maximum. Among measures of variability, particularly the standard deviation occupies the same exalted position as does the mean among measures of central tendency. While measures of variability for qualitative data are virtually non-existent, ranked data can be described with extreme scores (ranks).

5.7 References/Further Readings

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

Moore David S., 2002, The Active Practice of Statistics: A Text For Multimedia Learning, (Fourth Printing), USA W. H. Freeman and Company

Witte Roberts S. and John S. Witte, 2017, Statistics, (Eleventh ed.) Hoboken, NJ: John Wiley & Sons, Inc.,

5.8 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86
Range = 86 – 20 = 66
Q1 = 27
Median = 34
Q3 = 50
Inter-quartile range = 50 -27 = 23
Semi-interquartile range = 11.5

MODULE 4 INFERENTIAL STATISTICS

Inferential statistics are tools used to draw inferences or conclusion from data analyzed not description of data only. We are going to concentrate on correlation and regression as statistical for carrying out inferential analysis. As you will be shown in the course of our discussions in this module, correlation value can be used to determine whether there is relationship or association between two or more variables; the nature of that relationship or association – positive or negative and to determine the strength of the relationship. However, correlation does not determine causality. Regression analysis is another inferential statistical tool that can be used not only to determine association or relationship, but also to predict either the independent or dependent variables depending on which one is given or known. For easy comprehension of the two topics, this module is thematically structured into two units under which these issues have been addressed in some details for your learning.

Unit1 Correlation Analysis
Unit 2 Intuitive Approach
Unit 3 Regression Analysis

You are advice to study each of the unit carefully as you are expected to answer some questions to evaluate your understanding on the various issues as discussed. Possible answers to the questions are provided under each of the unit appropriately.

UNIT 1 CORRELATION ANALYSIS

Unit Structure

1.1 Introduction

Recall that we have dealt with the first aspect of statistics – descriptive statistics in the last module. This having being done, it is

also important to discuss the second aspect of statistics, the aspect we refer to as inferential statistics. This goes beyond mere presentation of data values and the use of graphic to represent our data. Inferential statistics attempt to draw conclusion from our data analysis. In this unit, you will learn about correlation as a measure of association between two variables at the interval level of measurement. Meanwhile correlation is one of the inferential statistical tools. The unit begins with the explanation of what measure of association stands for and the use of scatter diagram and the various forms of association between two variables – independent and dependent. The product moment correlation coefficient and the spearman rank correlation are also fully discussed with examples.

1.2 Learning Outcomes

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

- define correlation as a measure of association between two or more variables;
- use scatter diagram to show association between independent and dependent variables
- calculate correlation coefficient using pearson product- moment correlation formula; and
- calculate correlation coefficient using spearman rank correlation coefficient.

1.3 Measure of Association: Correlation

Correlation is a measure of degree of association between variables. The coefficient of correlation is a single number that tells us to what extent two variables or things are associated or related and to what extent variations in one variable go with variations in the others. In general, the correlation coefficient of a sample is denoted by r, and the correlation coefficient of a population is denoted by ρ or R. It has three advantages:

1. Correlation coefficient indicates whether there is a relationship or no relationship between two variables. When r = 0, it indicates no relationship or association.
2. Correlation coefficient describes the direction and the magnitude of the relationship between two variables. For instance, when r = - 0.4, it indicates negative or inverse relationship and when r = 0.7, it indicates positive or direct relationship.

3. Correlation coefficient gives information on the strength of the relationship between two variables. For instance, a correlation coefficient of 0.9 indicates very strong relationship or association, 0.4 weak association, 0.5 is moderate relationship or association.

If you want to interpret a correlation coefficient, consider the following:

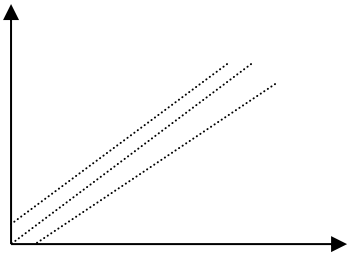
- The sign and the absolute value of a correlation coefficient describe the direction and the magnitude of the relationship between two variables.
- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of a correlation coefficient, the stronger the linear relationship.
- The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
- The weakest linear relationship is indicated by a correlation coefficient equal to 0.
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger. High values on one variable are associated with high values on the other variable and vice –versa.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller. This is also termed as an inverse relationship. It implies that high values on one variable are associated with low values on the other.
- Keep in mind that the Pearson product-moment correlation coefficient only measures linear relationships. Therefore, a correlation of 0 does not mean zero relationship between two variables; rather, it means zero linear relationship.
- Knowing that two variables are correlated does not tell us whether one causes the other.

1.4 Scatter Diagram

If we have data on two variables measured on the interval or ratio scale and we are interested in examining the relationship between them. We can plot the data by putting the dependent variables on the vertical axis and the independent variables on the horizontal axis. The resulting plot in what we call the scatter diagram. It gives a clear picture of the direction of the relationship between the two variables. There are 3 main forms of association between two variables.

1. *Positive correlation or direct association*

Dependent

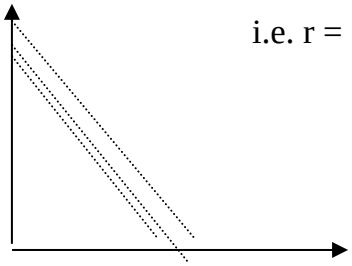


i.e. r = + ve

Independent

2. *Negative correlation or inverse association*

Dependent

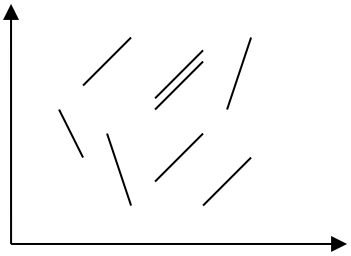


i.e. r = -ve

Independent

3. *Zero correlation or uncorrelated*

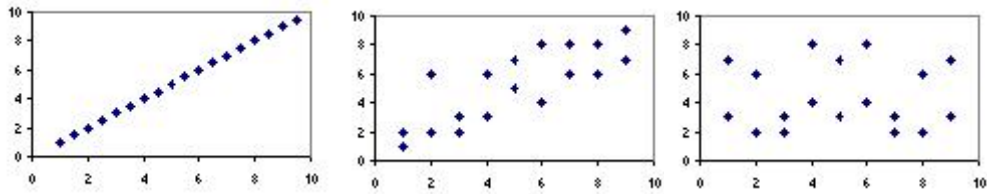
Dependent



i.e. r = O

Independent

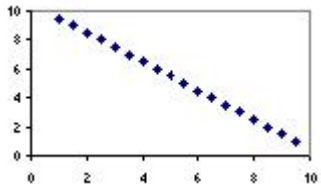
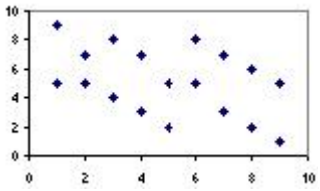
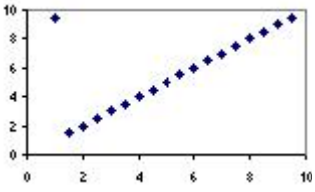
You may also consider how scatterplots can be used to determine correlation coefficient. The scatterplots below show how different patterns of data produce different degrees of correlation.



Maximum positive correlation (r = 1.0)

Strong positive correlation (r = 0.80)

Zero correlation (r = 0)

Maximum negative correlation (r = -1.0) | Moderate negative correlation (r = -0.43) | Strong correlation & outlier (r = 0.71)

You should have observed these points from the scatterplots above.

(a) When the slope of the line in the plot is negative, the correlation is negative; and vice versa.

(b) The strongest correlations (r = 1.0 and r = -1.0) occur when data points fall exactly on a straight line.

(c) The correlation becomes weaker as the data points become more scattered.

(d) If the data points fall in a random pattern, the correlation is equal to zero.

(e) Correlation is affected by outliers (extreme scores). Compare the first scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 1.00 to 0.71).

In the section that follows you will be taught how to calculate the correlation coefficient.

1.5 Pearson Product-moment correlation coefficient

The product moment correlation coefficient otherwise known as the Pearson moment correlation coefficient between X and Y is given by

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

The Pearson's r is based on the linearity assumption between the variables X and Y. For the computation of r, use the table format below.

X	Y	X^2	Y^2	XY
⋮	⋮	⋮	⋮	⋮
ΣX	ΣY	ΣX^2	ΣY^2	ΣXY

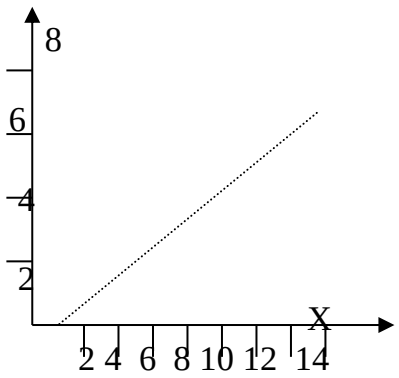
Let us illustrate the use of this table with an example.

Given the data below, calculate the Pearson's correlation coefficient r between X and Y and interpret your result.

X	1	3	4	6	8	9	11	14

Y	1	2	4	4	5	7	8	9

Solution: Although the question does not ask you to plot scatter diagram, but you may plot it to have a rough idea of what the relationship is.
Step 1: Scatter plot of X against Y.



To plot the scatter diagram appropriately, you need a graph sheet.

From the Scatter plot, it is clear that a positive relationship exists between X and Y.
Step 2: Arrange your data in the table format given above.

Revenue ($Mn) X	Expenditure ($Mn) Y	X^2	Y^2	XY
1	1	1	1	1
3	2	9	4	6
4	4	16	16	16
6	4	36	16	24
8	5	64	25	40
9	7	81	49	63
11	8	121	64	88
14	9	196	81	126
56	40	524	256	364

Note that: adding together values in column X gives $\Sigma X = 56$
, , , , , Y gives $\Sigma Y = 40$
, , , , , X^2 gives $\Sigma X^2 = 524$
, , , , , Y^2 gives $\Sigma Y^2 = 256$
, , , , , XY gives $\Sigma XY = 364$

Also note that N = 8 because there are 8 observations.

Step 3: Now apply the formula 11.1 to calculate r.

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

$$= \frac{(8 \times 364) - (56 \times 40)}{\sqrt{((8 \times 524) - (56)^2)(8 \times 256) - (40)^2}}$$

$$= \frac{2912 - 2240}{\sqrt{(4192 - 3136)(2048 - 1600)}}$$

$$= \frac{672}{\sqrt{(1056)(448)}}$$

$$= \frac{672}{\sqrt{473088}}$$

$$= \frac{672}{687.8}$$

$$= 0.977$$

$$= 0.97$$

Comment: The value of the correlation coefficient is very large. This shows that there is a very high linear correlation between X and Y. An increase in one of the variables is accompanied by a proportional increase in the other.

Fortunately, you will rarely have to compute a correlation coefficient by hand. Many software packages (e.g., Excel) and most graphing calculators have a correlation function that will do the job for you.

1.6 Spearman's Rank correlation coefficient (ℓ)

Suppose the values of variables X and Y are measured on a continuous scale and that they are jointly and normally distributed. Instead of using the precise value of the variables, we can rank the data in order of size, importance, preference etc, using the number 1, 2, …… , n. We then calculate the correlation coefficient as follows.

$$r_s = \rho = 1 - \frac{6\sum D^2}{N\left(n^2 - 1\right)}$$

Where d = difference between ranks of corresponding values of X and Y.

N = Number of pairs of value (X,Y).

Let us work this exercise together.
Two different collation officers collated the results of gubernatorial elections of eight parties. The no of votes in ('000) collated are as follows.

Political Parties	APP	BCP	CPP	DSP	ECP	FNP	GNP	HPP
Votes collated by C/ Officer I	46	37	71	64	48	55	35	36
Votes collated by C/Officer II	48	42	68	63	60	58	41	40

(a) Calculate the rank correlation coefficient.
(b) What conclusion can you draw?

Solution
Step 1: Rank the 2 scores using R_1 for rank of no of votes collated by C/O I and R_2 for no of votes ranked by C/O II.

Note: The highest score is marked 8 and lowest is ranked 1.

A	B	C	D	E	F	G	H
C/O I	46	37	71	64	48	55	35	36

C/O II	48	42	68	63	50	58	41	40
R_1	4	3	8	7	5	6	1	2
R_2	4	3	8	7	5	6	2	1
$D= R_1 - R_2$	0	0	0	0	0	0	-1	1
$D^2 = (R_1 - R_2)^2$	0	0	0	0	0	0	1	1

$\sum d^2 = \sum (R_1 - R_2)^2 = 1+1 = 2$

Step 2: Apply the formula for calculating r_s

r_s

$$= 1 - \frac{6\sum D^2}{N\left(n^2 - 1\right)}$$

$$= 1 - \frac{6x2}{8\left(8^2 - 1\right)}$$

$$= 1 - \frac{12}{8(63)}$$

$$= 1 - \frac{12}{504}$$

$$= \frac{504 - 12}{504}$$

$$= \frac{492}{504} = 0.9902$$

The rank correlation coefficient r_s is very high and positive. This shows that there is a very close tie between the votes collated by the two collation officers.

Self-Assessment Exercises (SAEs) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 30 minutes.

1. The main property of the correlation coefficient (r) is that it lies between
(a) O and 1 (b) O and –1 (c) –1 and +1 (d) -1 and O
2. What is the value of r when there is a perfect inverse relationship between two variables?
(a) -1 (b) +1 (c) O (d) 2
3. Pearson's r considers rank of the score while Spearman's r considers the raw scores.
TRUE or FALSE
4. The basic assumption of Pearson's r is based on linearity between the variables.
TRUE Or FALSE
5. The straight line that passes through most of the scatter point in known as
(a) Line of best fit
(b) Line of curvature
(c) Line of correlation
(d) None of the above
6. The following constitute the figures of imports revenue (IR) and exports revenue (ER) in millions of Nigerian Naira (1999- 2007):

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
IR	482	434	463	517	647	664	709	637	771
ER	313	299	385	649	449	457	500	552	596

(a) What is the appropriate measuring scale for these data?
(b) Is there correlation between the imports revenue and exports revenue?
(c) Rank the values of Imports revenue and Exports Revenue and calculate the rank correlation coefficient. What conclusion can you draw?

1.7 Summary

Correlation coefficient is essentially used in statistical analysis to measure association or relationship among variables. It is not being used to determine causality. Correlation also helps to determine the nature or type of association and the strength of the association. The Pearson's

product moment correlation coefficient considers the raw data recorded on an interval or ratio scale. The Spearman's rank correlation coefficient (ρ) considers the rank of the score instead of the raw data. The scatter diagram is the graphical representation of the relationship between two variables. The line of best fit is the single line that passes through most of the scatter point

In this unit you were told that inferential statistics is used to draw conclusions or inferences from data analyzed. Correlation is one of the inferential statistical tools that measures association or relationship between and among variables. You were told that correlation coefficient (r) can be obtained using Pearson Product-Moment correlation formula and when the data are ranked you can use Spearman Rank Correlation formula.

1.8 References/Further Reading

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers PVT Limited.

Clegg, F. (1990) *Simple Statistics: A Course Book for the Social Sciences* (CUP).

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Ott, L. et al (1983). *Statistics: A Tool for the Social Science* (3rd Ed). Boston, Massachusetts: Duxbury Press.

Spiegel, M R. (2000). *Probability and Statistics* (2nd Ed), Schaum's Outline Series. New York: McGraw-Hill.

1.9 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. c
2. a
3. False
4. True
5. a
6.

(a) is there correlation between the imports revenue and exports revenue?

(b)

Yes, Since r = 0.59 which means fairly strong direct or positive relationship

(c) Rank the values of Imports revenue and Exports Revenue and calculate the rank correlation coefficient.

r_s = 0.633 (

(d) What conclusion can you draw?

There is strong relationship between Imports revenue and exports revenue.

UNIT 2 INTUITIVE APPROACH

Unit Structure

2.1 Introduction

In the last unit, we discussed the advantages of correlation analysis and how we can carry out correlation analysis using graphical and computational methods. In this unit, you will learn how to use intuitive approach to determine whether there is positive or relationship between given independent and dependent variables.

1.2 Learning Outcomes

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

- use intuitive approach to determine the nature of relationship that exists between given independent and dependent variables.

2.3 Correlation Analysis: An Intuitive Approach

An intuitive approach in correlation analysis is based on the premise that the nature and types of relationship between two or more variables can be discovered through reasoning like mathematics and logic. Does the familiar saying “To whom much is given, much is expected” accurately describe the exchange between elected legislators and their constituent members? A political researcher suspects that a relationship exists between the amount of money donated by constituent members to support their elected Senator and the number of constituency projects carried out by the Senator after general elections.

Prior to a full-fledged survey and also prior to any statistical analysis based on variability, the researcher obtains the estimates for the current legislative session from five Senators, as shown in Table 2.1 below.

SENATOR	Amount of Donations Received (₦ millions)	Number of Constituency Project Carried out
A	50	10
B	70	12
C	130	14
D	90	18
E	10	6

Table 2.1 Amount of Donations and Number of Constituency Projects

If the suspected relationship does exist between amount donated and number of constituency projects, then an inspection of the data might reveal, as one possibility, a tendency for "much donations" to attract "many constituency projects" and for "little donations" to attract "few constituency projects." More generally, there is a tendency for pairs of scores to occupy similar relative positions in their respective distributions.

Trends among pairs of scores can be detected most easily by constructing a list of paired scores in which the scores along one variable are arranged from largest to smallest.

In Table 2.1a, the five pairs of scores are arranged from the largest (130 million naira) to the smallest (10 million naira) amount of donations received by each Senator. This table reveals a pronounced tendency for pairs of scores to occupy similar relative positions in their respective distributions. For example, Senator E received relatively little donations (10m) and carried out relatively few projects (6), whereas Senator C received relatively much donations (130m) and carried out relatively many projects (14). We can conclude, therefore, that the two variables are related. Furthermore, this relationship implies that "To whom much is given, much is expected." Insofar as relatively low values are paired with relatively low values, and relatively high values are paired with relatively high values, the relationship is positive.

Table 2.1a

SENATOR	**Amount of Donations Received (₦ millions)**	**Number of Constituency Project Carried out**
C	130	14
D	90	18
B	70	12
A	50	10
E	10	6

Thus, **positive relationship** occurs insofar as pairs of scores tend to occupy similar relative positions (high with high and low with low) in their respective distributions.

Table 2.1b

SENATOR	**Amount of Donations Received (₦ millions)**	**Number of Constituency Project Carried out**
C	130	6
D	90	10
B	70	14
A	50	12
E	10	18

Table 2.1c

SENATOR	**Amount of Donations Received (₦ millions)**	**Number of Constituency Project Carried out**
C	130	10
D	90	18
B	70	12
A	50	6
E	10	14

In Tables 2.1b and 2.1c above, each of the five Senators continues to receive the same amount of donations from their constituent members as in Table 2.1a, but new pairs are created to illustrate two other possibilities, which include a negative relationship and little or no relationship. Please take note that in real applications, of course, the pairs are fixed by the data and cannot be changed.

You should take note of the pattern among the pairs in Table 2.1b. Now there is a pronounced tendency for pairs of scores to occupy dissimilar and opposite relative positions in their respective distributions. For instance, although Senator E received relatively little donations (10m), he carried out relatively many projects (18). From this pattern, we can conclude that the two variables are related. Furthermore, this relationship implies that "You give the opposite of what you expect."

Insofar as relatively low values are paired with relatively high values, and relatively high values are paired with relatively low values, the relationship is negative. Therefore, a **negative relationship** occurs insofar as pairs of scores tend to occupy dissimilar relative positions (high with low and vice versa) in their respective distributions.

In terms of little or no relationship, as observed in Table 2.1c above, no regularity is apparent among the pairs of scores. For example, although both Senator A and Senator E received relatively little donations (50m and 10m, respectively), Senator A carried out relatively few projects (6) and Senator E carried out relatively many projects (14). Given this lack of regularity, we can conclude that little, if any, relationship exists between the two variables and that "What you give has no bearing on what you should expect."

Self-Assessment Exercises (SAEs) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 5 minutes.

1. Indicate whether the following statements suggest a positive or negative relationship:
(a) Less densely populated areas have lower crime rates.
(b) Heavier taxes draw less support for ruling political party.
(c) Better-educated people have higher incomes.
(d) More anxious civil servants voluntarily spend more time performing a simple repetitive task.
(e) Politicians who often campaign on TV only perform more poorly in general elections.

2.4 References/Further Readings

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

Moore David S., 2002, The Active Practice of Statistics: A Text For Multimedia Learning, (Fourth Printing), USA W. H. Freeman and Company

Witte Roberts S. and John S. Witte, 2017, Statistics, (Eleventh ed.) Hoboken, NJ: John Wiley & Sons, Inc.,

2.5 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

a. Positive relationship
b. Negative relationship
c. Positive relationship
d. Positive relationship
e. Negative relationship

UNIT 3 REGRESSION ANALYSIS

Unit structure

3.1 Introduction

In the last unit, we considered the measure of association or relationship between two variables. In this unit, we shall discuss a measure of the impact one variable (called the independent) has on another variable (the dependent). We shall also discuss how to make predictions based on the past evidence of the two variables.

3.2 Learning Outcomes

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

- calculate simple regression coefficient and thus fit a regression line.
- predict future values of the variables using the regression line.

3.3 The Regression Analysis

Regression analysis is another inferential statistical tool that can be used not only to determine association or relationship, but also to predict either the independent or dependent variables depending on which one is given or known. The regression coefficient is used to measure the impact of one variable on the other. The relationship between two variables X and Y can be represented by the regression equation. This equation can be used to predict.

Y = a + bX + e

The equation above is called the regression line. In that equation, **Y** is called the dependent variable, **X** is called the independent variable, **a** is called the intercept (constant) and **b** is called the regression coefficient. e, called the error term is often ignored. This equation also signifies that

we are predicting Y from X. For instance, if we are interested in knowing the relationship between income and level of savings, we may use the equation as follows:

Y = a + bX +e

Y = Amount of savings

X = Income level

a = constant (amount saved when income level is zero)

b = The factor by which saving will increase given a unit increase in income.
e = Error term (usually ignored)

How do we compute **a** and **b**?

The procedure is to compute **b** first, and later on compute **a**.
The computation formula is as given below:

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

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

Where $\bar{Y} = \frac{\sum Y}{N}$ that is the mean of Y

$\bar{X} = \frac{\sum X}{N}$ That is the mean of X

Let us consider the exercise below:

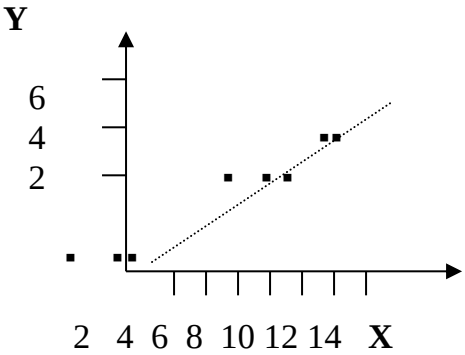
The income and savings of some local government employees is as given below:

Income X '(000)	7	5	10	15	13	14	8	11
Saving Y '(000)	1	1	4	6	5	4	2	4

(a) Construct a scatter diagram
(b) Find the least square regression line of Y and X
(c) How much is an employee that collects #20,000 likely to save?

Solution

(a) Scatter plot of X and Y



We can draw a straight line to pass through some of the point, although there are outliers.

(b) the required regression line is:

$Y = a + bx$

Step 1: In order to calculate **a** and **b**, arrange your computation in a table as below:

X	Y	XY	X^2
7	1	7	49
5	1	5	25
10	4	40	100
15	6	90	225
13	5	65	169
14	4	56	196
8	2	16	64
11	4	44	121

The sum of each of the columns gives:

$\Sigma X = 83$, $\Sigma Y = 27$, $\Sigma XY = 323$,

$\Sigma X^2 = 949$

N = Number of observations = 8 employees

Step 3: Compute the value of b:

b

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

$$= \frac{(8x323) - (83x27)}{(8x949) - (83)^2}$$

$$= \frac{2584 - 2241}{7592 - 6889}$$

$$= \frac{343}{703}$$

$$= 0.4879$$

Step 4: Compute the value of a:

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

$$= \frac{\sum Y}{N} - b\frac{\sum X}{N}$$

$$= \frac{27}{8} - (0.4879)\frac{83}{8}$$

$$= 3.375 - (0.4879)(10.375)$$

$$= 3.375 - 5.062$$

$$= -1.687$$

Step 5: Fit the regression line:
The regression line between X and Y is Y = a + bx

$\longrightarrow$ Y= - 1.687+ 0.4879X

(c) You are required to find Y when X = 20. All you need do is to put 20 in place of X in the regression line.

i.e. Y = - 1.687 + 0.4879 X
= - 1.687 + 0.4879(20)
= - 1.687 + 9.758
= 8.071
= ₦8,071,000

Interpretation

The regression line Y = -1.687 + 0.4879X signifies that, when an employee receives a salary of zero, that is when X = 0, he / she will be indebted by a sum of ₦1,687: 80k. Remember a is the intercept of the

regression line. Also remember that the figures are given in thousand naira in the question. Furthermore, the regression coefficient b = 0.4879 shows that an employee's savings will increase by a factor of ₦487.90k whenever the salary increases by a factor of ₦1,000.

The Coefficient of Determination (r^2)

The coefficient of determination (r^2) measures how much variation in the dependent variable that is explained by the independent variable. It's simply the power of explanation of the dependent variable (Y) by the independent variable (X). Its value lies between zero and positive one.
i.e. $0 \leq r^2 \geq 1$

The closer the value of r^2 to 1, the better the goodness of fit (i.e. the more the independent variable X explains the variation in Y.
It is computed as follows:

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

Where n = no of cases
$\sum XY$ = Sum of all the product of X and Y
$\sum X$ = Sum of all data values of X
$\sum Y$ = Sum of all data values of Y
$\sum X^2$ = Sum of all the square of X
$\sum Y^2$ = Sum of all the square of Y

Self-Assessment Exercises (SAEs) 1

Attempt these exercises to measure what you have learnt so far. This should not take you more than 25 minutes.

1. In the straight-line equation Y = a + bx, we say.
(a) Y depends on a
(b) Y depends on b
(c) Y depends on X
(d) None of the above

2. In the regression equation Y = a + bX, a is called the
(a) Intercept (b) coefficient (c) Predictor (d) None of the above

3. In the regression equation Y = a + bx, the term that measures the change in the dependent variable for a unit change in the independent variables is
(a) Y (b) a (c) b (d) X

4. In the regression line Y = a + bx, if the value of a = 5, b = 3 and X = 25, predict the value of Y.
(a) 40 (b) 100 (c) 75 (d) 80
5. In the regression line Y = a + bx
(a) = #40, (b) = #6 and Y = #200.
Predict the value of X.
(a) #240, (b) #40 (c) #200 (d) #400
6. If $\Sigma X = 234, \Sigma Y = 760, \Sigma XY = 25800, \Sigma X^2 = 11{,}336$ and N = Number of observations = 8 countries. X = population of people (in millions) Y = Number of legislative seats
Find the least square regression line of **Y = a + bX** Use the answer to predict the number of legislative seats for a country in Africa, if her population is 50 million.

3.5 Summary

The regression coefficient b measures the impact of the independent variable X on the dependent variable Y. The regression equation can also be used to predict future values of Y when the value of X is known and to predict values of X, when the value of Y is known. The amount of variations in Y that is explained by the independent variable X is measured by r^2 called the coefficient of determination. Its values lie between 0 and 1.

In this unit, we have explained how you perform the regression analysis as another inferential statistics that you can use to measure the impact of the independent variable X on the dependent variable.

3.6 References/Further Reading

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers PVT Limited.

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Ott, L. et al (1983). *Statistics: A Tool for the Social Science* (3rd Ed). Boston, Massachusetts: Duxbury Press.

Spiegel, M R. (2000). *Probability and Statistics* (2nd Ed), Schaum's Outline Series. New York: McGraw-Hill.

Wright, D.B. (1997) *Understanding Statistics: An Introduction for the Social Sciences* (Sage).

3.7 Possible Answers to Self-Assessment Exercises (SAEs)

Answers to SAEs 1

1. c
2. a
3. c
4. d
5. b
6. If $\Sigma X = 234$, $\Sigma Y = 760$, $\Sigma XY = 25800$, $\Sigma X^2 = 11{,}336$ and N = Number of observations = 8 countries. X = population of people Y = Number of legislative seats

(a) Find the least square regression line of $Y = a + bX$

(b) Predict the number of legislative seats for a country in Africa, if her population is 50,000,000

(a) b= 0.7948 a =71.75

Regression line Y= 71.75 + 0.7948X

(b) 111.49 seats approximately = 111 seats

MODULE 5 TESTING HYPOTHESES AND COMPUTER APPLICATIONS IN POLITICAL SCIENCE RESEARCH

In the last module, you will be taught how to test research hypotheses using the manual methods and how to use computer applications or software (particularly Excel and SPSS). To achieve the set objectives, this module is thematically structured into five units under which these issues have been addressed in some details for your learning.

Unit 1 Meaning and Types of Hypotheses
Unit 2 Statistical Tools for Testing Hypothesis in Political Science Research: T test
Unit 3 Statistical Tools for Testing Hypothesis in Political Science Research: Chi-Square Test
Unit 4 Elementary Probability Theory
Unit 5 Computer Applications in Data Analysis

You are advice to study each of the unit carefully as you are expected to answer some questions to evaluate your understanding on the various issues as discussed. Possible answers to the questions are provided under each of the unit appropriately.

UNIT 1 MEANING AND TYPES OF HYPOTHESES

Unit Structure

1.1 Introduction

A statistical analysis of political phenomenon is meaningless if there is no pre-stated research hypotheses. It is the hypothesis that will serve as the searchlight while the theory serves as the footpath to a successful research outcome. In the light of this, this unit will examine what a good hypothesis entails, types of hypotheses, how to formulate it, and how to test it.

1.2 Learning Outcomes

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

- Define a hypothesis and highlight the types of hypotheses.
- Formulate a good hypothesis and discuss the procedures to be taken in testing it.
- Distinguish between Type 1 and Type 2 error in hypothesis testing.
- demonstrate your research competence in the area of hypothesis testing.

1.3 Meaning and Components of a Good Hypothesis

Decisions that are made about a population on the basis of sample information are called statistical decisions. In attempting to reach decisions, it is useful to make assumptions (or guesses) about the larger population involved. Such assumptions, which are open to be accepted or rejected, are called statistical hypotheses. A hypothesis is a testable proposition or conjectural statement made by a researcher at the commencement of the research work which represents his/her beliefs or assumption about the relationship that exist within the research variables. It is a tentative answer to a research problem that is being studied. Hypotheses are to be tested against the agreement between the implied result and the existing body of knowledge. The main function of the hypothesis is to sharpen or concentrate attention to the problem and determine the direction in which the solution to the problem can be found.

A hypothesis serves as a link between the world of theory and the world of reality. Sources of hypotheses include, researcher's experience, observation, knowledge of literature, findings of completed studies, creativity etc. More importantly, every hypothesis must be clear, be specific, testable and should be value – free.

Components of a "Good" Hypothesis

It has been established that a good hypothesis for political analysis must:

1. be politically relevant.
2. be plausible or it must make a good deal of sense.
3. be positive and non-tautological. By positive, we mean simply that, as a matter of technical style, they must be stated as propositions rather than asked as questions. "Are women more likely to vote than men?" is a question and not a hypothesis. By tautology we mean statements that are true by definition: e.g.

"Rich people have more money than poor people" or "Belligerent nations are more likely to go to war than peaceful nations" are tautological statements.

4. be addressed to a single set of cases e.g. "Military regimes are more likely to spend higher on Defence than civilian regimes". In this hypothesis, the dependant variable is Defence expenditure. The independent variable is regime type. Cases are the nation states. What component 4 is saying is that the states whose Defence expenditure are considered must be the same set of states whose regime type are considered.
5. encompass more than one case and the cases must fall into more than one category on the independent variable. Note that a variable may be independent in some hypotheses and dependent in others.

1.4 Types of Hypotheses

For the purpose of our discussion, we limit ourselves to the two main hypotheses – viz – The Null Hypothesis and The Alternative Hypothesis.

The Null Hypothesis

A proposition that is formulated for the sole purpose of rejecting or accepting it, based on the result of a statistical test, is called a Null hypothesis. The Null hypothesis is often denoted by H_O.

The Alternative Hypothesis

Any hypothesis that differs from a Null hypothesis is called an alternative hypothesis. It is in fact a negation of the Null hypothesis. It is sometimes called the research hypothesis, denoted by H_1.

Thus, we can denote a hypothesis as a proposition about an assumed relationship between two or more variables.

If after conducting a statistical test, we discover that the observed result is different remarkably from the expected result under the hypothesis, then we would be inclined to reject or accept the hypothesis. Procedures that enable us to determine whether observed sample differ significantly from the result expected and thus help us decide whether to accept or reject hypothesis, are called test of hypotheses, test of significance or decision rules.

1.5 Types of Error in Hypothesis Testing

There are two type of error that could be made when hypotheses are being tested. If we reject a hypothesis when it should be accepted, we say that a Type I error has been made. On the other hand, if we accept a hypothesis when it should be rejected, we say that a Type II error has been made. In either case, a wrong decision or error in judgement has occurred.

	Accept	Reject
H_O True	Right Decision	**Type I Error**
H_O False	**Type II Error**	Right Decision

** Note that Type I error is more serious than Type II error.
In summary, Type I error: A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the significance level. This probability is also called alpha, and is often denoted by α.

Type II error: A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by β. The probability of not committing a Type II error is called the Power of the test.

1.6 Steps Involved in Conducting Hypothesis Testing

Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. Testing of the hypothesis is a crucial activity because until a hypothesis is tested and checked against available data, it is nothing more than a guess.
Before you proceed further, perform the following tasks: In the few hypotheses given below, you should identify the dependent, independent variables, and the case involved. The solution to the first one has been provided. You can follow the same format.

1. "States with higher per capita income spend a larger proportion of their budgets on education than states with lower per capita income".
Independent variable = per capital income
Dependent variable = proportion of budget devoted to education
Cases = States

2. "States with higher rate of unemployment are likely to also have higher crime rate."

3. "Nations with unequal income distribution are less likely to be democracies than nations with more equal income distribution."

Solution to the second hypothesis:
Independent variable = rate of unemployment
Dependent variable = crime rate
Cases = States
Solution to the third hypothesis:
Independent variable = income distribution
Dependent variable = democratic status
Cases = Nations

You need to follow the following steps when testing a hypothesis:

(a) Formulation of a hypothesis: This involves:
1. Making of a Null hypothesis (H_O).
2. Making of an Alternative hypothesis (H_1).

(b) Set up a suitable level of significance denoted by **α** e.g. 5% or 1%. In testing a given hypothesis, the maximum probability with which we would be willing to risk a Type I error is called the level of significance. It is conventional among researcher to adopt either 0.05 (5%) or 0.01 (1%). If 0.05 significance level is chosen in designing a decision rule, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted; that is, we are about 95% confident that we have made the right decision.

(c) Choose a suitable test statistic: Selecting an appropriate statistical technique depends on a number of considerations, such as number of variables involved (two or three), the type of data (interval, ordinal, nominal, or ratio), the size of the sample (large or small) and whether the samples are independent or selected. It could be a

- T – test
- Z – test
- Chi – Square test
- Correlation
- F – test etc.

(d) Determine the degree of freedom, depending on the test-statistics you are using. You will learn more about this in our next lecture.

(e) Calculation or performance of various computations necessary for the application of selected statistical test or technique. Evaluate your result by comparing computed value with tabulated value.

(f) Making of decisions based on (c), (d) and (e) above. It involves the acceptance or rejection of null hypothesis. This depends on whether the computed value of the statistical test falls in the region of acceptance or rejection at a given level of significance. For instance;

If calculated value > tabulated value, REJECT H_O
If calculated value < tabulated value, ACCEPT H_O
It is important for you to note that the analysis plan includes decision rules for rejecting the null hypothesis. It is common for statisticians to describe these decision rules in two ways - with reference to a P-value or with reference to a region of acceptance.
(A) P-value. The strength of evidence in support of a null hypothesis is measured by the **P-value**. Suppose the test statistic is equal to *S*. The P-value is the probability of observing a test statistic as extreme as *S*, assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.
(B) Region of acceptance. The **region of acceptance** is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected. The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level.
(g) Interpret your result by giving the relevant social implications.

Let us consider the following methods in hypothesis testing.

Method of use of Correlation Table

Let me remind you that correlation analysis is used to determine the existence or not of a relationship between two or more variables. The correlation coefficient also shows both the strength of a relationship (if any) and its direction (positive or negative). If there is relationship, you may want to know if it is significant, and this is where the correlation (Pearson) table above comes in.

To do this test, you need to compute the correlation coefficient (let us assume that the correlation coefficient of number of constituency seats and total number of registered voters in 20 states is 0.98), set out the H_O and H_I, establish the Significance Level, degree of freedom (d. f.) through (n – 2). The next thing to do is to check out the critical value under the level of significance from the statistical table. Your hypothesis testing layout should look like the one below:

H_O: There is no significant relationship between the total number of registered voters and constituency seats
H_1: H_O is false
α = 0.05 (one tailed test)
d. f. = (n-2) = (20 -2) = 18
r = 0.98

Critical Value (at d. f. of 18 under 0.05) = 0.378 (Check Pearson correlation table in your Statistical table)

Decision and Interpretation: Since computed value of r 0.98 is greater than the critical value of 0.378, H_O is rejected and H_1 accepted.
Conclusion: There is significant relationship between the total number of registered voters and the constituency seats.

Method of use of Probability (Z score test) Table

The standard normal probability table is used to calculate probability – involving one (or a subset of scores) whose mean $\bar{X}$ and standard deviation 'S' have been computed. To do this, you subtract the mean of the set or group from the score, divides the difference or deviation by the standard deviation to get the Z score, or standard score. The next thing is to check out the probability of the wanted scores at the computed Z score and under the chosen level.

For example, a political science lecturer conducted a snap test to 20 students who get various scores. He computed the mean of the group to be 50, and the standard deviation to be 15. Now he wants to know the probability of a student scoring 70%. This is how he finds it.
Given N =20, $\bar{X}$ = 50 S = 15 Significance level = 0.05
Conversion of 70 to Z score through

$$\frac{X - \bar{X}}{S} = \frac{70 - 50}{15} = \frac{20}{15} = 1.3$$

To get probability, he enters the table at Z of 1 .3 under 0.05 = 0.0885
Conclusion: The probability of a student scoring 70% = 0.0885 or 8.9%

Self-Assessment Exercise

1. What is a hypothesis?
2. A proposition formulated for the sole purpose of rejecting or nullifying based on the result of a statistical test is called (a) a first hypothesis (b) a good hypothesis (c) a rejected hypothesis (d) a null hypothesis
3. If a Null hypothesis is accepted when in fact it is false, the researcher is said to have committed a (a) Type I error (b) Type II error (c) Right decision (d) Wrong decision.
4. Rejecting a Null hypothesis when in fact it is false is a (a) Type II error (b) right decision (c) wrong decision (d) Type I error.

1.7 Summary

In this unit, we defined what a hypothesis is and highlight the types and components of a good hypothesis. A hypothesis is a proposition about an assumed relationship between two or more variables. A Null hypothesis is often made for the sole purpose of accepting or rejecting it,

based on the result of a statistical test. A good hypothesis must be politically relevant, plausible, non – tautological and addressed to a single set of cases. Type I and Type II errors are those errors committed when a wrong decision is made. The level of significance is the maximum probability with which the researcher is willing to risk a Type I error.

You were informed that there are two types of errors as a result of wrong decisions we make when testing our hypotheses. Steps or procedures to be taken when hypotheses are being tested were explained in the unit.

1.8 References/Further Reading

Aggarwal, Y.P. (1998). *Statistical Methods: Concepts, Application and Computation*. New Delhi: Sterling Publishers, Pvt Limited.

Champney, L. (1995). *Introduction to Quantitative Political Science*. New York: Harper Collins College Publishers.

Kitchens, L.J. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference* (2nd Ed.). USA: Duxbury Press.

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

UNIT 2 STATISTICAL TOOLS FOR TESTING HYPOTHESIS IN POLITICAL SCIENCE RESEARCH: T – TEST

Unit Structure

2.1 Introduction

In the last unit, we explained the meaning of a hypothesis and what hypothesis testing entails. There are different test-statistics which serve as relevant tools in hypothesis testing. They are the T-test and the Chi-square test. In this unit, our discussion will focus on the T-test or student T. It is one of the most commonly used techniques for testing a hypothesis on the basis of a difference between sample means. You can use the t- test to determine a probability that two populations are the same with respect to the variable tested.

2.2 Learning Outcomes

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

- define and explain what t test means.
- apply the T -test as a tool for test of significance.

2.3 Main Content

2.3.1 The T -Test Statistic

It is important for you to note that the t- statistic was introduced in 1908 by William Gosset, a chemist working for the Guinness Brewing Industry in Dublin, England. The Student's **t** (Student was the pen name of Mr. Gosset) is used to estimate one mean, and comparing two means for matched or unmatched data. To use it in estimating one mean, you have to compute the sample mean, the student t statistic, and set out your H_O and H_I, Significance Level, degree of freedom (d. f.) through n-1 (where n is the number of cases). The next thing to do is to enter the

student's t table at the d. f. value, and checks out the critical value under the level of significance.

For example, a political science researcher wishes to investigate if countries of the world are exceeding an established 3% maximum level for defence expenditure as a percentage of Gross National Product (GNP). He collects data on a sample of 10 countries and gets a mean of 7.6% and t of 3.3. He now moves on to test for significant difference between the hypothesised mean and computed mean.

This is the appropriate layout for you to consider.
H_O: The countries are not exceeding the maximum 3% level
H_1: H_O is false
$\alpha = 0.05$
d. f. = n-1 (10 -1) = 9
t = 3.3
Critical Value (at d. f. of 9 under 0.05) = 1.833 (Check T table in your Statistical table)
Decision and Interpretation: Since t value of 3.3 is greater than the critical value of 1.833, H_O is rejected and H_1 accepted.
Conclusion: The Countries are exceeding the maximum 3% level.

Let us go into details on how to compute the t -value.
(a) It may be of interest to test whether the difference between the sample mean and the population mean is so significant that we cannot attribute it to chance factors or sampling variation.
(b) Likewise, we may have 2 different samples drawn from the same population, and our interest is to test if there is any significant difference between the means of the two samples. When we are confronted with problems like the above, the T-test is to be employed in carrying out the test. It operates by computing a critical value T (or t calculated) and comparing it with a table (or t tabulated) value at a specified degree of freedom and level of significance. T calculated is computed as follows:

For the scenario (a) above, we use the formula

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

for a small sample, when only the sample variance is known.

or $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

for a large sample and when the population variance is known

Where $\overline{X}$ = Sample
μ = Population mean
S = Sample standard deviation
σ = Population standard deviation
n = Sample size

For scenario (b) above, we use the formula:

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

for small samples

Where $\sigma = \sqrt{\frac{N_1 S1^2 + N_2 S2^2}{N_1 + N_2 - 2}}$

And for a large sample we use:

$$T = \frac{\overline{X_1} - \overline{X_2}}{SE_D}$$

Where $SE_D (\sigma_D) = \sqrt{\frac{\sigma_1^2}{N1} + \frac{\sigma_2^2}{N2}}$

Symbols:

$\overline{X}_1$ = Means of sample 1

$\overline{X}_2$ = Mean of sample 2

N_1 = Number of Observation in sample 1

N_2 = Number of Observation in sample 2

σ = pooled Variance

S_1^2 = Sample 1 Variance

S_2^2 = Sample 2 Variance

SE_D (σ_D) = Standard error of mean difference.

A sample is large when N > 30.

Determining the degree of freedom:
(i) for independent samples, df = N_1 + N_2 - 2.
(ii)for one sample case, df = N - 1

Decision Rule
If t calculated > t tabulated, Reject H_o and Accept H_1
If t calculated < t tabulated, Accept H_o and Reject H_1
Let us now use an example to illustrate what we have discussed so far.

Exercise1: A Times survey reveals that the ages of Post UTME candidates follow normal distribution with mean 23years and standard deviation 7years. However, a randomly selected sample of 50 UTME candidates shows that their average age is 25. Do we have enough evidence to conclude that the Times survey was low?

Solution
Let μ = true population means of JAMB candidates
Step 1: Set up your hypothesis
H_o : μ = 23 years
H_1 : μ ≠ 23years

Step 2: Compute your z statistics

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{25 - 23}{\frac{7}{\sqrt{50}}}$$

$$= \frac{2}{\frac{}{7}} \quad x \quad 7.07$$

$$= 2.02$$

Step 3: Compare this value with the table value at 5%

Z (5%) = (-1.96 --------- +1.96)

Since Z calculated falls outside this range, we reject H_O and conclude that the true population mean μ is something greater than 23 years. Thus, the value given by the Times Survey is too low.

Example An arithmetic test is given to 2 sets of students. The first group consists of 30 students who attend evening classes after school hours while the second group consists of 40 students who do not attend any extra lesson. Their mean and standard deviation are:

	N	M	SD
Group 1	30	20.5	4
Group 2	40	16.2	5

Is the difference in their average score due to chance?

Solution

Step 1: Set up your hypothesis.

Ho : $\overline{X}1 = \overline{X}2$ difference is due to chance

H1 : $\overline{X}1 \neq \overline{X}2$ difference is not due to chance.

Step 2: Calculate SED

$$SE\ D\ (\sigma D) = \sqrt{\frac{\sigma 1\ 2}{N1} + \frac{\sigma 2\ 2}{N2}}$$

$$= \sqrt{\overline{42} + \overline{52}}$$

$$= \sqrt{\frac{16}{30} + \frac{25}{40}}$$

$$= \sqrt{\frac{640}{30} + \frac{750}{40}}$$

$$= \sqrt{\frac{1390}{120}}$$

$$= \sqrt{11.58}$$

$$= 3.4$$

$$df = N1 + N2 - 2$$

$$df = 30 + 40 - 2$$

$$= 68$$

Step 5: Check the table value of T
T tabulated (0.05, df 68) = 1.67

Step 6: Compare calculated t with tabulated t.
Since t calculated < t tabulated
i.e. 1.26 < 1.67
We accept H_o and conclude that the difference between the two means is only due to chance.

Self-Assessment Exercise

1. Which of the statements below is *not true*?
(a)T – test can be used to test for association
(b)T- test can be used to test a hypothesis on the basis of a difference between sample means
(c) Another name for T – test is Student T
(d) T – test can be used to test hypothesis

2. The Ministry of Health of Oyo state wants to monitor the level of industrial pollution of the Ogunpa River. Ten sites were randomly selected and the amount of toxic wastes in parts per million (P/Mn) found to be as follows:

Site	A	B	C	D	E	F	G	H	I	J
P/Mn	5	10	9	6	7	8	8	7	4	11

(i) If the safety level was set at 7 parts per million, estimate the difference between the means
(ii) Assuming a critical value of 1.833 (0.05,df =9) use T test statistic to test the null hypothesis (Ho)that industries along Ogun river are obeying the 7 P/Mn
(iii) Advise the Ministry accordingly.

2.4 Summary

In this unit, you have been shown through the procedures and exercises how the T-test can be employed when testing whether difference between two means is significant or due to chance factors. The T – test or the Student's **t** is used to estimate one mean, and comparing two means for matched or unmatched data. To use it in estimating one mean, you have to compute the sample mean, the student t statistic, and set out your H_O and H_I, Significance Level, degree of freedom (d. f.) through n-1 (where n is the number of cases)

1.5 References/Further Reading

Gupta, C.B. (1982). *An Introduction to Statistical Methods*. Delhi: Vikas Publishing House Limited.

Kitchens, L. J. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference* (2nd Ed). USA: Duxbury Press.

Ott. L. et al (1983). *Statistics: A Tool for the Social Science* (3rd Ed). Boston, Massachusetts: Duxbury Press.

UNIT 3 STATISTICAL TOOLS FOR TESTING HYPOTHESIS IN POLITICAL SCIENCE RESEARCH: CHI – SQUARE TEST

Unit structure

3.1 Introduction

In this unit, we shall be discussing Chi –Square as one of the commonest statistical tools for testing our research hypothesis. The procedures for testing hypothesis with chi-square statistic will be explained with examples in this unit.

3.2 Learning Outcomes

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

1. Define and explain what a test of significance means.
2. Use the Chi – square (X^2) test as a tool for test of significance.

3.3 Main Content

3.3.1 The Chi – Square Test Statistic

Chi – Square statistic is a statistical calculation used to test how well the distribution of a set of observed data matches a theoretical probability distribution. It is denoted by χ^2 and sometimes referred to as a goodness of fit test. It is often used to test for goodness of fit and also for independence or association between variables. It operates on categorical data, which are recorded based on frequency count. For instance: if 1,200 secondary school students are randomly selected and asked what profession they would like to choose in their adulthood. We may get a data like the one below.

Table 13.1

	Lecturer	Medical Doctor	Banker	Entertainer
No of Students	100	500	400	200

In the table above, the categories are the profession while the data are the number of children recorded. The task of the χ^2 - test therefore is to test whether the children are evenly distributed across the professions. Since we have 1,200 students and four professions ordinarily, we expect each profession to have ¼ of 1200 or 300 students and the observed frequencies are the ones in the table.

What χ^2 does is to compare the observe frequency with the expected frequency. If a large difference exists between observed and expected, the χ^2 will have a large value and vice versa. If you want to know about the "goodness to fit" between the observed and expected, you may ask this question: Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. How much deviation can occur before you, the investigator, must conclude that something other than chance is at work, causing the observed to differ from the expected. The calculated value is equal to the sum of the squares of the differences divided by the expected values. Statistically, it is computed by the formula:

$$X^2 = \frac{\sum (o-e)^2}{e}$$

With (K-1) degree of freedom
If there are more than one group in the contingency table, the degree of freedom will be:

d.f. = (r-1) (c-1)

Where r = no of rows

C = no of columns

The chi-square test is always testing what social scientists refer to as the null hypothesis, which states that there is no significant difference between the expected and observed result.

To use the χ^2 table, you need first computes the Chi –Square statistic, and set out your H_O and H_I, establish the Significance Level, degree of

freedom (d. f.) through (r-1) (c-1). The next thing to do is to enter the chi- square table at the d. f. value, and checks out the critical value under the level of significance. You need to know that Chi-square requires that you use numerical values, not percentages or ratios.

For example, assuming you want to find out whether there is significant difference between the opinions of rich and poor voters towards independent candidature in Nigeria's elections. From your data you compute a chi –square value of 12.8. If you chose a significant level of 0.05 and your table has two rows and two columns, your hypothesis testing layout should look like the one below:

H_O: There is no significant difference between the opinions of men and women

H_1: H_O is false

α = 0.05 (The relative standard commonly used in social research is $p > 0.05$. The p value is the **probability** that the deviation of the observed from that expected is due to chance alone (no other forces acting). In this case, using $p > 0.05$, you would expect any deviation to be due to chance alone 5% of the time or less.)

d. f. = (r-1) (c-1) = (2 -1) (2-1) = 1

χ^2 = 9.8

Critical Value (at d. f. of 1 under 0.05) = 3.841 (Check Chi –Square table in your Statistical table)

Decision and Interpretation: Since chi –square value of 12.8 is greater than the critical value of 3.841, H_O is rejected and H_1 accepted.

Please Note: the Rule for decision making

(i) If the calculated Chi –Square value is > than Critical Value from the table, we accept Alternative hypothesis (H_1) and reject Null hypothesis (H_0).

(ii) If the calculated Chi –Square value is < than Critical Value from the table, we reject alternative hypothesis (H_1) and accept Null hypothesis (H_0).

For our example above:

Conclusion: There is significant difference between the opinion of rich and poor voters towards independent candidature in Nigeria's elections.

Let us discuss the computation of chi –square value in details.

Example: Sixty respondents were asked what their opinion is about convocation of Sovereign National Conference. The data was obtained:

Please note that the P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than calculated Chi – square.

(i) If the p value for the calculated Chi -Square is $p > 0.05$, accept the hypothesis. 'The deviation is small enough that chance alone accounts for it. A p value of 0.6, for example, means that there is a 60%

probability that any deviation from expected is due to chance only. This is within the range of acceptable deviation.

(ii) If the p value for the calculated Chi –Square is $p < 0.05$, reject the hypothesis, and conclude that some factor other than chance is operating for the deviation to be so great. For example, a p value of 0.01 means that there is only a 1% chance that this deviation is due to chance alone. Therefore, other factors must be involved

	Hausa	Igbo	Yoruba	Total
Support	17	8	5	30
Against	3	12	15	30
Total	20	20	20	60

Test the hypothesis that:

H_o : Ethnic affiliation and opinion on SNC are independent

H_1: Ethnic affiliation and opinion on SNC are dependent.

Solution: The above given figures are called the observed frequencies.

Step 1: Calculate the expected frequencies thus:

	Hausa	Igbo	Yoruba	Total
Support	$\frac{20x30}{60}=10$	$\frac{20x30}{60}=10$	$\frac{20x30}{60}=10$	30
Against	$\frac{20x30}{60}=10$	$\frac{20x30}{60}=10$	$\frac{20x30}{60}=10$	30
Total	20	20	10	60

Step 2: Tabulate the frequency columns as follows for easy computation.

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
17	10	7	49	4.9
8	10	-2	4	0.4
5	10	-5	25	2.5
3	10	-7	49	4.9
12	10	2	4	0.4
15	10	5	25	2.5
				15.6

Step 3: Compute χ^2

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

= 15.6 see the procedure in step 2 above.

Step 4: Compare calculated χ^2 with tabulated χ^2 df = (2-1) (3-1) = 2, since we have 2 rows and 3 columns.

χ^2 tabulated (5%, df 2) = 5.99

Step 5: Decision-making
Since χ^2 calculated is greater than χ^2 tabulated i.e. 15.6 > 5.99, we reject the H_o and accept H_1 and conclude that ethnic affiliation and opinion on the convocation of Sovereign National Conference are dependent i.e. Whether a respondent is in support or against SNC depends on his/her ethnic affiliation.

To determine the strength of the relationship, Calculate crammer's V statistic :

$$V = \sqrt{\frac{X^2}{N(n-1)}}$$

X^2 = Chi Square estimate
N = no of cases or respondents
N = no of rows or no of columns (whichever is the smallest)
Therefore $V = \sqrt{\frac{15.6}{60(2-1)}}$

= 0.51 (There is moderate or not very strong relationship between ethnic affiliation and opinion on the convocation of Sovereign National Conference.

Self-Assessment Exercise

There have been talks about application for a review of UN International Court of Justice (ICJ)'s decision which ceded Bakassi Peninsula to Cameroun in 2002. A political Scientist who wants to know the views of 30 Nigerians and 30 Cameroonians on the matter carried out a survey. The question asked the respondents is "Do You approve of the idea for a review of ICJ's decision? Data below were collected:

	Nigerians	Cameroonians	Total
YES	28	4	32
NO	2	26	28
Total	30	30	60

(a) Construct the relevant tables and calculate the Chi-square χ^2

(b) Test the null hypothesis 0.05, d.f. 1 using a critical value of 3.84
(c) What percentage of the total sample approves the idea for a review of ICJ's decision?

3.4 Summary

We have discussed Chi -Square test statistic as one of the use tools in hypothesis testing. It is denoted by χ^2 and we have shown you through the exercises how you can use it to test for goodness of fit and also for independence or association between variables. We can conclude that the chi-square test is a goodness of fit test, which operates on categorical data recorded based on frequency counts.

3.5 References/Further Reading

Gupta, C.B. (1982). *An Introduction to Statistical Methods*. Delhi: Vikas Publishing House Limited.

Kitchens, L. J. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference* (2nd Ed). USA: Duxbury Press.

Ott. L. et al (1983). *Statistics: A Tool for the Social Science* (3rd Ed). Boston, Massachusetts: Duxbury Press.

Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

UNIT 4 ELEMENTARY PROBABILITY THEORY

Unit structure

4.1 Introduction

Although the uncertainty principles create a kind of complexity in making inferences in social science in general and in political science research in particular, there is a need to develop a tool that will enable us make inferences. Probability is the tool that enables inference-making even at the face of uncertainty. It refers to the proportion or fraction of times that a particular event is likely to occur. In this unit, we will be examined the basic principles of probability and random variables as it relates to social events.

4.2 Learning Outcomes

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

- define Probability
- apply the principles of probability in making predictions.
- compute the probability of a given event.

4.3 Main Content

4.3.1 Definition of Probability

The probability of an event can be determined in several ways. We can *guess* that if a coin is truly fair, heads and tails should be equally likely to occur whenever the coin is tossed, and therefore, the probability of heads should equal 0.5, or ½ or 50%. Similarly, ignoring the slight differences in the lengths of the months of the year, we can *guess* that if a couple's wedding is equally likely to occur in each of the months, then the probability of a June wedding should be 0.08 or 1/12 or 8.33%.

On the other hand, we might actually *observe* a long string of coin tosses and conclude, on the basis of these observations, that the probability of heads should equal 0.5, or 1/2. Or we might collect extensive data on wedding months and *observe* that the probability of a June wedding actually is not only much higher than the speculated 0.08 or 1/12 or 8.33%, but higher than that for any other month. In this case, assuming that the observed probability is well substantiated, we would use it rather than the erroneous guessed probability.

As we have mentioned in the introductory section of this lecture, virtually all human activities are uncertain events. For instance, a candidate in a presidential election will think of the possibility of success or failure. In an election, a candidate will either win or lose. Our task here is to predict the chances of winning or losing the election.
By simple definition, the probability of an event E is given by:

$$P(E) = \frac{\text{Total number of occurrences of event E}}{\text{Total sample space}}$$

So how do we define an event, an outcome, and a sample space?
Events and Outcomes
In statistical terms, an event is an outcome of an experiment. Thus, success is an event in an election so also failure is an event. In fact, the expected outcome of any election is the set:

A = {win, lose}
Likewise in a war game, the expected outcome of a war is the set B = {win, lose}. Thus, in the war experiment, win is an event and lose is another event.
There are some experiments that have more than two possible outcomes. The collection of all possible outcomes in an experiment is called the sample space.
Thus, in numerical term, the sample space of an election is 2 (i.e. win or lose).

Likewise, the sample space of a war game is 2. By using our definition for probability of an event {E}, equation above,
If E is the event that a candidate wins in election, then

> * Here, we assume that the game is fair. We also assume the two events have equal chance of occurrence.

P (E) = ½
Since there is only one favourable way that "win" can occur and the total sample space is 2. Let us illustrate the concept of probability with another example.

Example: A Senate Committee on Education consists of 5 AD Senators, 7 ANPP Senators and 8 PDP Senators. What is the probability that a Senator selected at random from the committee is a PDP Senator?
Solution
Total sample space = (5 + 7 + 8) Senators
= 20 Senators
Number of outcomes favourable to PDP Senators = 8
Let E = event of selecting a PDP Senator, then using equation 8.1
P (E) = 8/20
= 0.4 or 40%
This is a fairly high probability.
It is important for you to take note that probability is always between 0 and 1. It is just a guide to make decision. Probability does not tell us exactly what will happen, it is just a director. For instance, if you toss a coin 10 times, how many Tails will come up? Probability says that Tails have a ½ chance, so we can expect 5 Tails. But when we actually try it we might get 4 Tails, or 7 Tails ... or anything really, but in most cases it will be a number near 5.

4.3.2 Elementary Probability Theory

Complementary Events
Two events are said to be complementary if they both exhaust all possible outcomes.
For instance, if the event of winning a war is ½ and the event of losing the war is also ½, then, the event "winning the war" + "Not winning the war":

= ½ + ½
We also assume a fair game here.

= 1 which exhausts all possible outcomes in that experiment

As a rule, if A^1 is the complement of event A, then

$P(A)^1 = 1 - P(A)$

Mutually Exclusive Events
Two events are said to be mutually exclusive if the occurrence of one prevents the occurrence of the other. That is, the two events cannot occur at the same time.
For example, in a single election, the events
W = Candidate A wins and
L = Candidate A loses

Are mutually exclusive events. A candidate cannot win and lose in a single election all together.

Thus P (W U L) = P (W) + P (L)
We call this equation the addition rule of probability.

The **addition rule** tells us to add together the separate probabilities of several mutually exclusive events in order to find the probability that any one of these events will occur. Stated generally, the addition rule reads:

Pr(*A* or *B*) Pr(*A*) Pr(*B*) where Pr( ) refers to the probability of the event in parentheses and A and B are mutually exclusive events.

Independent Events
Two events are said to be independent if both can occur together at the same time. That is the occurrence of the one does not prevent the occurrence of the other.

e.g. Let us define two events as follows: A = Nigeria wins the Bakassi war. B = Nigeria wins the next African Nations Cup.
Then A and B are said to be independent events since both events can occur at the same time. Therefore:
P (A ∩ B) = P (A) **.** P (B)

We call this equation the multiplication rule of probability.
Whenever you must find the probability for two or more sets of independent events connected by the word *and,* use the multiplication rule. The multiplication rule tells us to multiply together the separate probabilities of several independent events in order to find the probability that these events will occur together. Stated generally, for the independent events *A* and *B,* the multiplication rule reads:
Pr(*A* and *B*) [Pr(*A*)][Pr(*B*)] where *A* and *B* are independent events.

If W and L are *not mutually exclusive,* then the equation becomes

P (W U L) = P (W) + P (L) - P (W ∩ L)
Example: Suppose it is known that the probability of winning the Bakassi War is 0.58 and the probability of playing at the finals of the next African Nations Cup is 0.61. What is the probability of winning the war and playing at the finals?

Solution: Since the two events are independent, we shall use equation (8.4)
Let A = event of winning the war. Therefore P (A) = 0.58
Let B = event of playing at the finals. Therefore P (B) = 0.61

Then:
P (A ∩ B) = P (A) P(B)
= (0.58) (0.61)
= 0.35 which is a weak probability.

Self- Assessment Exercises (SAEs1)
Attempt these exercises to measure what you have learnt so far. This should not take you more than 10 minutes.

1. If the probability that an incumbent governor will win an up - coming election is 4/5, what is the probability that he will lose the election?
(a) 3/5 (b) 4/5 (c) 1/5 (d) 2/5

2. The collection of all possible outcomes in an experiment is called
(a) Sample size (b) Sample mean
(c) Sample space (d) Sample half

3. The outcome of an experiment is called
(a) Science (b) Event (c) Sample space (d) None of the above

4. A function that associates a real number to every element (event) in the sample space is called
(a) A random variable (b) A dependable variable
(c) An independent variable (d) None of the above

5. The probability of any given event can only have values that lie between:
(a) -1 and +1
(b) 0 and +10
(c) 0 and 1
(d) None of the above

4.4 Summary

You learnt in this unit that probability is a powerful tool in inference making at the face of uncertainty. It is defined as the ratio of the number of favourable outcomes of an event to the total number of equally likely sample space. An event is an outcome of an experiment. The probability of an event can only lie between zero and one.

4.5 References/Further Reading

Gupta, C.B. (1982). *An Introduction to Statistical Methods*. Delhi: Vikas Publishing House Pvt Limited.

Ott, L. et al. (1983). *Statistics: A Tool for the Social Sciences* (3rd Ed). Boston, Massachusetts: Duxbury Press.

Spiegel, M.R et al. (2000). *Probability and Statistics* (2nd ed), Schaum's Outline Series. New York: McGraw-Hill.

Witte Roberts S. and John S. Witte, 2017, Statistics, (Eleventh ed.) Hoboken, NJ: John Wiley & Sons, Inc.,

4.6 Possible Answers to SAEs

1. c
2. c
3. b
4. a
5. c

UNIT 5 COMPUTER APPLICATIONS IN DATA ANALYSIS

Unit structure

5.1 Introduction

Our discussion so far in this course has been centered on using manual calculators, pen, pencil, graphs and sheets of paper in estimating parameter values. However, if we have a large volume of data before us, which is usually the case with survey research, it may be tedious to use manual calculators. Thus, in this last unit, we introduce the use of computer applications in Data analysis.

5.2 Learning Outcomes

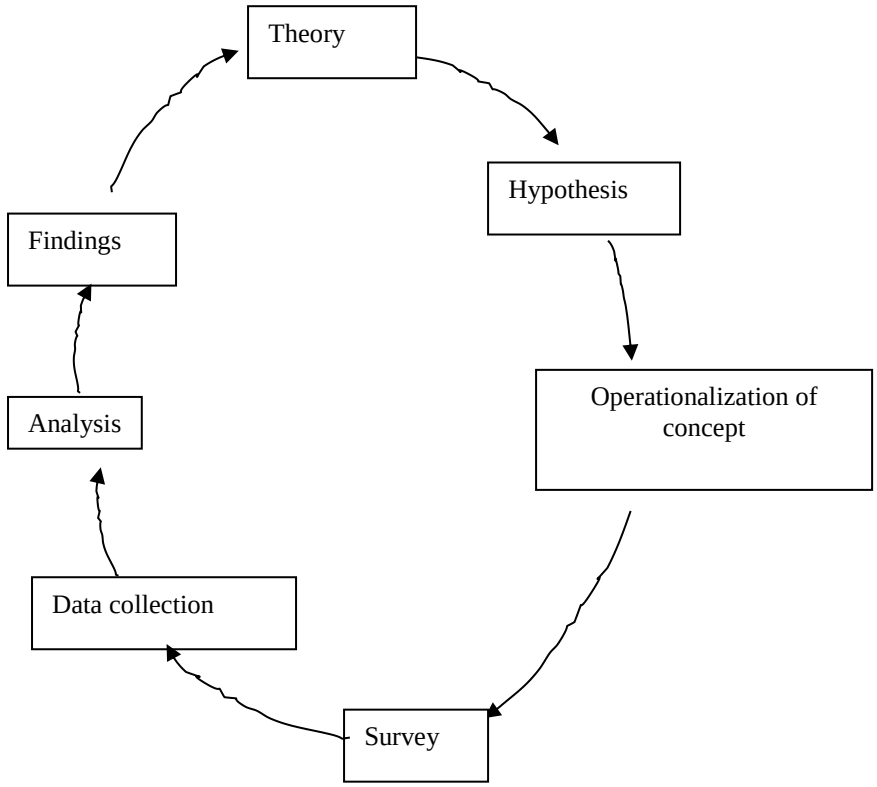
At the end of this unit, you will be able to:

- prepare a good questionnaire for survey research.
- code information collected by the questionnaire method.
- prepare data for analysis and use a computer package or application to analyze your data.
- have more knowledge about using statistical package for the social sciences or what is now known as statistical product and service solutions (spss) and or any other statistical packages.

5.3 Setting the Stage in the Use of Computer Applications for Data Analysis

Let me remind you of some principles we discussed in the earlier part of this course. We discussed that in any qualitative research work, it is

always rewarding to have specific guidelines for the research to be meaningful. Many scholars have suggested the following basic steps in qualitative research.



Source: Atoyebi (2003) P. 128

The starting point is the theory. The next stage is the setting up of relevant hypotheses. The hypotheses are propositional statements hence they are made up of concepts. These concepts must be measured. You must develop measures for these concepts that is learn how to operationalize them. A faulty operational definition will produce a wrong measure. You may do a pre survey test to try your measure. Then you go out to the field to conduct your survey by carefully selecting your sample. Collect your data through the various devices we have discussed in this course. Before the data can be analyzed it must be prepared in a computer ready format (that is coding). This stage will be discussed further in this lecture. The result of your analysis will validate or disprove your hypothesis.

It is very important for you to be computer literate for you to make good use of its application in your analysis of political phenomenon. We are making a very big assumption here that you are computer literate. Let

me tell you point blank that you do not have to be a computer "guru" in order to use statistical packages, since you can hire the service of a data analyst. Nonetheless, knowledge of computer operating systems is desirable. The skills you need are not limited to what we listed below. As a beginner, you should be able to:

1. Put on the computer.
2. Open a worksheet.
3. Create a file.
4. Save a file.
5. Retrieve a file.
6. Print a file.
7. Use the keyboard.
8. Use the mouse.
9. Have knowledge of what statistical tool you wish to use.

For items 2 and 6 you need to use the mouse to click. No special skill is required.
What you will discover about the computer is that the windows environment is interactive, so even when it appears you cannot go further, the Help Window will come to your rescue.

For coding of data, the data file of the package you are using is divided into rows and columns. The interjection of these rows and columns create cells in which numerical data can be typed. A typical data file appears like below:

	1	2	3	4	5	6	7
1							
2							
3							
4							

Table A

The rows and columns are numbered serially and there are several rows and several columns on a worksheet. The boxes are called cells. By coding we mean specifying what figure goes into what cell for which variable.
Let us use a typical questionnaire to illustrate coding technique.

QUESTIONNAIRE

This template is designed to measure the relationship between ethnicity and voting behaviour (not in details)

Que 1: Age:

16-25	26-35	36-45	46-55	> 56

Que: 2 Religion

Christianity	Islam	Traditional	Others

Que 3: Ethnic group | Yoruba | Hausa | Igbo | Others |

Que 4: Occupation | Employed | Unemployed | Student |

Que 5: Gender | Female | Male | others |

Que 6:
What party do you belong to?

| APC | LP | PDP | No Party |

Que 7:
What Party would you vote for in the next presidential election?

| APC | LP | PDP | No Idea |

You should note that in question (1) age is measured on interval scale but by grouping the respondents into age brackets, you have reduced it to an ordinal level. Question 2-6 are all measured on Normal scale.
Your coding guide will appear as follows:

Que 1:	Age:	1	=	16-25
		2	=	26-35
		3	=	36-45
		4	=	46-55
		5	=	Above 56
Que 2:	Religion	1	=	Christianity
		2	=	Islam
		3	=	Traditional
		4	=	Others
Que. 3	Ethnicity	1	=	Yoruba
		2	=	Hausa
		3	=	Igbo
		4	=	Others
Que 4:	Occupation	1	=	Employed
		2	=	Unemployed
		3	=	Student
Que 5:	Gender	1	=	Female
		2	=	Male
		3	=	Others

Que 6: Pol. Party	1	=	APC
	2	=	LP
	3	=	PDP
	4	=	No Party
Que. 7	1	=	APC
	2	=	LP
	3	=	PDP
	4	=	No Idea

Now assuming you pick the first questionnaire already filled and returned from the respondent, and the person is 37 years old, a Muslim, a Yoruba woman, employed, a member of APC, would prefer to vote for APC in the next election. Then you will feed in the data into your data file on the computer. Thus, the table will become:

	V1	V2	V3	V4	V5	V6	V7	
1	3	2	1	1	2	1	1	
2								
3								

Table B

Notice that the title of the column have been changed to V1, V2, …….V7. What we have done is define our variables (i.e. give them names that the computer will use to recognize them. Instead of using V1 for Ques 1, you may decide to use Q1. It is flexible. Just click on DEFINE VARIABLE and the dialog box that appears will be interactive. You must also specify the value labels and the column size of each of your variables. All these are contained in the dialog box, just click as desired.

Now we want to assume you have treated all your returned questionnaire and that you have key in the data into your data file (table A above). A very important information at this stage is that you should save your working files as you go along. Give it a suitable file name that you can easily remember.

The next stage is to select appropriate statistical test. This stage is the most technical stage in the sense that the computer operates on the principle of 'garbage in garbage out' (GIGO). The computer does not understand what you intend to do. You must instruct the computer accordingly. If you desire to run a T-test analysis but you wrongly click on Regression, you will get a wrong output. This is why we emphasize that a good knowledge of what statistical test you wish to carry out, in order to test your hypothesis, is very important. This must be decided even before you sit in front of the computer.

Once you know what test to carry out and what instruction to give to the computer, you have completed your own side of the contract. The computer will carry out all the instruction as required. You don't need to know the formula for calculating regression coefficient. Just click STATISTICS or ANALYSE, and from the list of tests that comes out pick REGRESSION. The dialog box that drops out will prompt you to specify what your dependent and independent variables are. After specifying all those, click OK and your Regression output will come up in few micro-seconds. Save the output file or print it as desired.

5.3.2 Using Excel Sheet

The following procedures should be followed:

1. Open an excel worksheet
2. Click on File
3. Go to option and then click
4. Look for side Menu and Click on Add-ins Box (Analysis Tool pack)
5. Check Add-ins: Analysis tool pack and Analysis Tool pack + VBA
6. Click on Data on the Excel worksheet
7. Click on any of the analysis tools

Anova: two factors without replicate
Correlation
Descriptive statistics etc.

8. Insert Input range by (i) highlighting the data set you want to analyse,

(i) New workbook
(ii) Summaries then Click OK.

Consider this data set

X	Y
40	50
45	40
40	30
35	32
90	86
65	62
37	40
86	95
40	40
50	55
56	54
28	27
85	80
59	56

62	70
38	50
70	72
72	69
40	45
50	48

Anova
Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	20	1088	54.4	344.7789
Column 2	20	1101	55.05	355.7342

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.225	1	4.225	0.012063	0.913122	4.098172
Within Groups	13309.75	38	350.2566			
Total	13313.98	39				

Correlation

	Column 1	*Column 2*
Column 1	1	
Column 2	0.950938	1

Descriptive Statistics
(A) Data set X

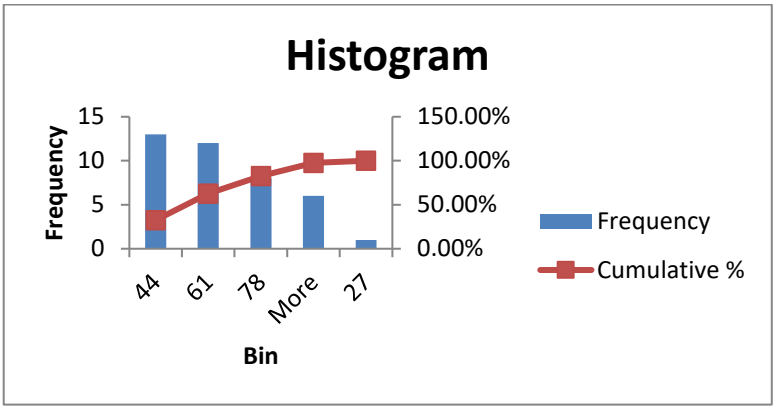
Column1	
Mean	54.4
Standard Error	4.151981
Median	50
Mode	40
Standard Deviation	18.56822
Sample Variance	344.7789
Kurtosis	-0.72941
Skewness	0.620973
Range	62
Minimum	28
Maximum	90
Sum	1088
Count	20
Largest(1)	90
Smallest(1)	28
Confidence Level(95.0%)	8.690196

(B) Data set Y

Column1	
Mean	55.05
Standard Error	4.217429
Median	52
Mode	40
Standard Deviation	18.86092
Sample Variance	355.7342
Kurtosis	-0.40639
Skewness	0.508273
Range	68
Minimum	27
Maximum	95
Sum	1101
Count	20
Largest(1)	95
Smallest(1)	27
Confidence Level(95.0%)	8.827181

Histogram

Bin	*Frequency*	*Cumulative %*	*Bin*	*Frequency*	*Cumulative %*
27	1	2.50%	44	13	32.50%
44	13	35.00%	61	12	62.50%
61	12	65.00%	78	8	82.50%
78	8	85.00%	More	6	97.50%
More	6	100.00%	27	1	100.00%



5.3.3 Using Computer Applications (SPSS)

SPSS (originally "statistical package for the social sciences," now "statistical product and service solutions") is a powerful program designed to allow users to perform a very wide range of data analysis. Data analysis is the language of research. In many fields, research is critical for human progress. Therefore, as long as there is research, there will be the need for data analysis. There are different versions of SPSS which are similar, although there are some differences because of additional functions in the higher version. It is recommended that you use the guide whilst sitting at a computer that is running latest SPSS version.

SPSS is one of the most widely used and powerful data analysis software. However, there are other data analysis software that are in use. Some examples are;
MATHEMATICA
STATISTICA
R
EPILNFO
Eviews
Maple
STATA
NCSS
CSpro
SAS
Mini tab

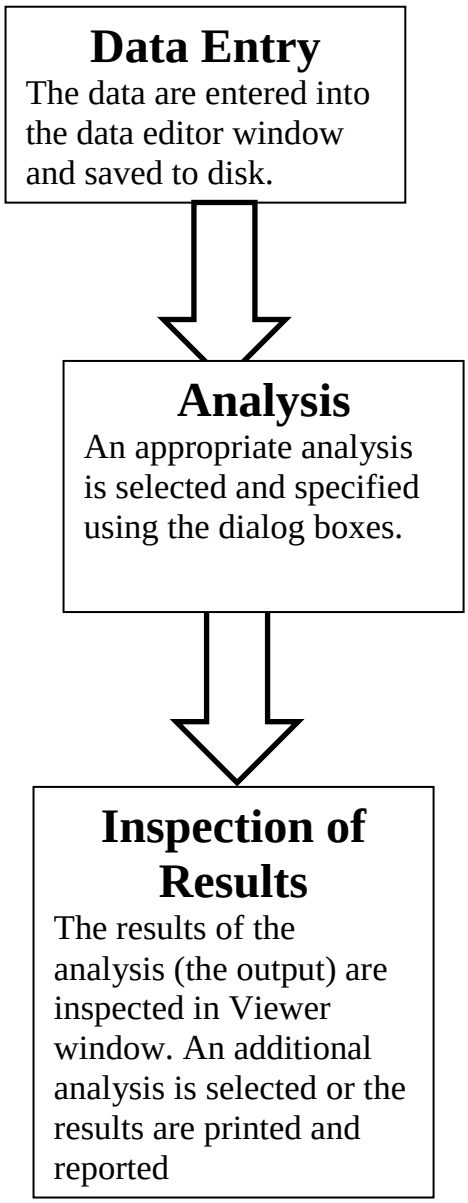
This guide below is designed to help you analyze data on your own. As mentioned earlier, a basic knowledge of statistics and a general acquaintance with the use of computer is, however, required for this guide to be effective in guiding you on how to conduct analysis with SPSS. The type of procedure to use and what the output mean will be meaningful provided that you have at least a rudimentary knowledge of statistics. Undoubtedly, the guide should provide individuals with limited statistical background ways of using SPSS to conduct statistical operations.

SPSS is useful for the following

- Data entry and data clearing.
- Descriptive and statistics analysis and output.
- Parametric and non-parametric test-tests of relationship and difference.
- Data division based on factors and groups.
- Quantitative research and with observed variables.
- Qualitative research with coded themes.
- Market research and trends.
- Data management and documentation.
- Predictive analysis
- Health statistics
- Surveys
- Data mining among others.

An overview of the structure of thus guide follows. Firstly, data analysis is considered by introducing you to the windows and buttons you will use when analyzing your data with SPSS. Here, how to start and exit SPSS. Create and save a data file and hoe to obtain some simple descriptive statistics are described. Secondly, crosstabulations and chi-square statistical procedure in SPSS is described.

There are three basic steps involved in data analysis using the SPSS. Firstly, you must enter the raw data and save to a file. Secondly, you must select and specify the analysis you require. Thirdly, you must examine the output produced by SPSS. These steps are illustrated below. The special windows used by SPSS to undertake these steps are described next.



SPSS utilizes several different window types. However, new users of SPSS only need to be familiar with two of these windows., the Data editor window and the viewer editor window. We will be using two windows in this guide. The other window types are explained very briefly below.

The Data Editor window

The data editor window (or data window) is the window you see when you start up SPSS. This spreadsheet-like window is used to enter all the data that is going to be analyzed. You can think of this window as containing a table of all your raw data. We will examine this data in detail when SPSS started us.

The viewer window

The viewer window is used to display the results of your data analysis. For this reason, we will sometimes refer to it as the output window. We will examine this window in more detail when we perform our first simple analysis.

Other windows used in SPSS

The syntax editor window is used to edit special program files called syntax files.
The chart Editor window is used to edit standard (not interactive) charts or graphs.
The pivot table editor window is used to edit the table that SPSS uses to present the results of your analysis.
The text output editor is used to edit the text elements of the output shown in the viewer window.

To get started, move the mouse pointer over the SPSS icon and double click on it (i.e., press the left- hand mouse button twice in rapid succession). After a brief delay you will see the Data Editor window as shown below. If you do not have an SPSS icon on your desktop then click on the start button at the bottom left- hand corner of the screen, then select Programs and then either SPSS 23.0 for windows or any other versions on windows.

3.4 Using SPSS for Data Analysis

Click on the link to have to detailed information and pictures on how to use SPSS for data analysis.

https://med.und.edu/daccota/_files/pdfs/berdc_resource_pdfs/data_analysis_using_spss.pdf

To use SPSS, after your computer system or other electronic device is open, you move the mouse pointer over the SPSS icon (already downloaded on the device you are using) and double click on it. After a brief delay you will see the **Data Editor window**. If you do not have an SPSS icon on your computer system or device, then click on the Start button at the bottom left -hand corner of the screen, then select **Programs** and then the SPSS 23.0 for Windows or any higher version of it that you have installed or downloaded from the internet.

After opening the full screen image of the Data Editor, with a detailed breakdown of the toolbar buttons below it, you will see the Menu bar (the various commands) and the Tool bar located at the top of the screen and are described below. Take note: when you start SPSS, there is no data in the Data Editor. To fill the Data Editor window, you may type data into the empty cells or access an already existing data file.

Toolbar

The toolbar icons are located below the menu bar at the top of the screen, the icons were created specifically for ease of point-and-click mouse operations. The format of the icon bar may vary slightly depending on which window you are working with. The toolbar shown applies to the Data Editor window. Also, note that some of the icons are bright and clear and others "grayed". Grayed icons are those that are not currently available. For example, note that the Print File icon is grayed because there is no data to print. When data are entered into the Data Editor, then these icons become clear because they are now available. The best way to learn how the icons work is to gently click on them and see what happens.

The Menu bar

The Menu bar (just above the toolbar) displays the commands that perform most of the operations that SPSS provides. You will become well acquainted with these commands as you spend time in the guide provided on the above link. Whenever you click on a particular command, a series of options appears below and you will select the one that fits your particular need. The commands are now listed and briefly described:

- **File**: deals with different functions associated with files including opening, reading, and saving, as well as existing SPSS
- Edit: a number of editing functions including copying, pasting, finding, and replacing.
- **View**: several options that affect the way the screen appears; the option most frequently used is ***Value Labels***.
- **Data**: operations related to defining, configuring, and entering data; also deals with sorting cases, merging or aggregating files, and selecting or weighing cases.
- **Transform**: transformation of previously entered data including recoding, computing new variables, reordering, and dealing with missing values.
- **Analyze**: all forms of data analysis begin with a click on Analyze command.

- **Graphs**: creation of graphs or charts can begin either with a click on the Graphs command or (often) as option while other statistics are being performed.
- **Utilities**: utilities deal largely with fairly sophisticated ways of making complex data operations easier. Most of these commands are for advanced users, and I will refer you to details provided in the link above.
- **Windows**: deals with the position, status, and format of open windows. This menu may be used instead of the taskbar to change SPSS windows.
- **Help**: a truly useful aid with search capabilities, tutorials, and a statistics coach that can help you decide what type of SPSS procedure to use to analyze your data.

The Output Window

The output is the term used to identify the results of previously conducted analyses. It is the objective of all data analysis. SPSS has a long history of efforts to create clear and comprehensive output. When utilizing options described below, the SPSS version 23.0 is somewhat of an improvement, but output can still be awkward and occupy many pages. It is hoped that the information that follows will maximize your ability to identify, select, edit, and print out the most relevant output. An output is shown in the output screen.

In dealing with this screen, the objective is to edit output so that, when printed, it will be reproduced in a format most useful to you. Of course, you do not have to reorganize output before you print, but there are often advantages to doing so:

(a) Extensive outputs will often use/waste many pages of paper
(b) At times a large table will be clearer if it is reorganized
(c) Most outputs will include some information that is not necessary
(d) You may wish to type in comments or titles for ease or clarity of interpretation.

The key element of the Output window shown above, as well as detailed description of each toolbar item follows. You will notice that several of the tool bar icons are identical to those in the SPSS Data Editor window, these buttons do the same thing that they do in the Data Editor, but with the Output instead of the Data. For instance, clicking on the print icon prints the output instead of the Data.

A key feature of the SPSS Output window that you need to learn is the use of the outline view on the left of the screen. On the right side of the window is the output from the SPSS procedures that were run, and on the left is the outline (like a table of contents without page numbers) of the output. The SPSS output is actually composed of a series of output

objects: these objects may have the titles (e.g. Frequencies), tables of numbers, or charts, among other things. Each of these objects is listed in the outline view. You will notice that there is no "notes" section in the output window to correspond with the "notes" title in the outline view. This is because the notes are (by default) hidden. If you want to see the notes, just double click on the closed book icon to the right of the **notes** title. The closed book icon will then become an open-book icon and the notes will materialize in the window to the right.

The Outline view makes navigating the output easier. Consider the following:

- If you want to move to the Crosstabs output, for instance, you merely need to click on the word "Crosstabs" in the outline view, and the Crosstabs will appear in the output window.
- If you want to delete the Descriptives section (perhaps because you selected an incorrect variable), simply click on the "Descriptives" and select menu item **Edit** then click **Delete**.
- If you want to move some output from section to another (to re-arrange the order), you can select an output object (or a group of output objects), and select **Edit** then click **Cut**. After that, select another output object below which you want to place the output object(s) you have cut, select **Edit** and then click **Paste After**.
- If you have been working with the same data for a while, you may produce a lot of output. So much output may be produced, in fact, that is becomes difficult to navigate throughout the output even with the outline view. To help with this problem, you can "collapse" a group of output objects underneath a heading. To do this, click on the minus sign to the left of the heading.

One particularly useful command when you are working with output is the insert text command. When you click on this button, an SPSS Text object is inserted. In this box you can type comments to remind yourself what is interesting about the SPSS output. Once you have typed your comments, click on another SPSS to de-select the SPSS Text object.

Obtaining Output in SPSS

(a) To obtain a Frequency Output in SPSS

1. Once your data is entered, checked and saved click on the word Analyze at the top of the screen.
2. Select (click on) Descriptive Statistics.
3. Select Frequencies: You will be presented with the frequencies dialog box. This dialog box contains two boxes. The left -hand box lists all the variables in the data file. The right-hand box (which will be empty when you first use this command) lists the

names of the variables which will be analyzed (i.e. for which a frequencies printout will be produced).

4. Select the first variable you want to include in the analysis by clicking on the variable in the left-hand box.
5. The arrow button between the two boxes will now be highlighted and will be pointing to the right-hand box. Click on this arrow button. The selected variable will be moved to the right -hand box. Repeat this procedure until the right-hand box contains the names of all the variables you want included in the frequencies analysis.
6. When you have selected all the variables you are interested in, click on the statistics button. This will reveal the **Frequencies; Statistics** dialog box which lists all the descriptive statistics available in the **Frequencies** command.
7. In the Frequencies: Statistics dialog box select all the descriptive statistics you require by clicking in the boxes so that a tick appears.
8. When you have selected all the statistics you require, click on (the Continue button) to return to the Frequencies dialog box.
9. Finally, click on the button to execute the frequencies command.

(b) To obtain a Tables output in SPSS

1. On the Menu bar, click on the **Analyze**.
2. Click on Custom Tables
3. Click on **Basic Tables**. This will display the **Basic Tables** dialog box.
4. Click on the name of the variable you require summary descriptive statistics, then click on the arrow button next to the **Summaries** box to move the variable into the **Summaries** box.
5. Next click on the name of the grouping variable. The grouping variable will be used to create the two or more groups for which the descriptive statistics will be calculated.
6. Now click on the arrow next to either the **Down**, the **Across** or the **Separate Tables** boxes. Which of these you choose determines how the table will appear in the output. The **Down** option produces a separate row for each level of the grouping variable, whereas the Across options produces a separate column for each level of the grouping variable. The Separate Tables option produces a separate table for each level of the grouping variable. Experiment with this setting to see which suit you best.
7. Now click on the **Statistics** button. **The Basic Tables: Statistics** dialog box will appear.
8. Select the descriptive statistics you require by picking them from the list in the left of the dialog box. Click on the Add button. To add the selected statistics to the box marked **Cell Statistics**. You may need to scroll down through the list of statistics available to find all of those you require.

9. Once the required statistics have been selected, click on the continue button. This will return you to the Basic Tables Dialog box. Now click on the OK button. The tables of statistics requested will now appear in the viewer window.

(c) **To obtain Descriptives output in SPSS**

1. Once your data is entered, checked and saved click on the word **Analyze** at the top of the screen.
2. Select (click on) Descriptive Statistics.
3. Select **Descriptives.** You will now be presented with the **Descriptives** dialog box.
4. Click the desired variable -name in the box to the left and then pasting it into the **Variable(s)** box to the right by clicking the right arrow in the middle of the screen. To deselect a variable (i.e. move it from the **Variable(s)** box back to the original list), click on the variable in the active box and the front arrow in the centre will become a back arrow. Click on the left arrow to move the variable back. To clear all the variables from the active box, click the **Reset** button.
5. If you wish to calculate more than the four default statistics, after selecting the desired variables, before clicking **OK**, it is necessary to click the **Options** button. To select the desired descriptive statistics, the procedure is simply to click the box behind the desired value of the descriptive statistics you wish. This is followed by a click of Continue and **OK**.

When you have finished working with SPSS you must exit the program. Do this in the following way:
A. Click on the word File at the bottom of the Screen.
B. Click on the word Exit from the pull-down menu presented.
C. If you have made any changes to either the Data Editor window or the output Viewer window since you last saved these files, then SPSS will display a dialog box asking you if you want to save these files before you exit from SPSS. Click on YES button to resave the file and exit SPSS. If you do not want to save your changes, click on the NO button to exit without saving. If you want to abort the Exit, perhaps to allow you to save the file in under a different name, click on the Cancel button.

Self-Assessment Exercise

Click on the link below again and study further on how to use SPSS for data analysis

https://med.und.edu/daccota/_files/pdfs/berdc_resource_pdfs/data_analysis_using_spss.pdf

5.5 Summary

You were told in this unit that computer statistical analysis is very useful in handling a large volume of data. It is faster and much more accurate than the manual calculators. A good knowledge of computer operating system and the kind of analysis you want to run is important. The correct instruction must be given to the computer because your input determines the output you get from the computer. In this last unit, we explained how you are can use the Excel Sheet and SPSS as computer applications. SPSS is an acronym of Statistical Package for Social Science but now it can also be referred to as statistical product and service solutions. Apart from Excel Sheet, there are other computer applications to carry out your data analysis.

SPSS is an application software developed by international business machines (IBM) since 2009; though originally developed by SPSS incorporated. SPSS was originally designed to be used in the social science, however, its usage and capabilities has extended to other areas of learning. The most recent release or version of SPSS is version 25.0 which was released in August 2017, which similar to other versions. SPSS is a tool and it only does what it is "told" to do. SPSS does not do the thinking for you. To use SPSS, you must have some basic knowledge of statistics. At first look the SPSS screen resembles a typical spreadsheet; but there is a lot more to SPSS. In short, you have been shown how to get into and out of SPSS.

5.6 References/Further Reading

Akintunde Elijah, (2017) Introduction to Statistics and SPSS, a manual prepared for Political Science students.

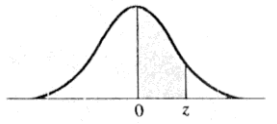
Olayinka Adeyemi Atoyebi (2003), *Statistical Methods in Political Science*, UI: Distance Learning Centre

Kenneth J. Meier, Jeffrey L. Brudney and John Bohte, (2006), *Applied Statistics for Public and Nonprofit Administration*, Canada: Thomson Wadsworth. Sixth edition

Adigun Agbaje and A. Ismail Alarape, 2005, Introductory Lectures on Research Methodology

Appendix I

Table I: Normal Curve Areas



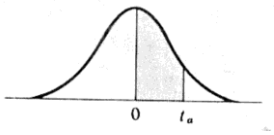
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.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	.2517	.2549
0.7	.2580	.2611	.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	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.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	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.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	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.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	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

** Adapted from OTT, Lyman, R.F. Larson and W. Mendenhall (1983) STATISTICS: A tool for the Social Sciences. (3rd ed) BOSTON: Duxbury press.*

Appendix II

Table II: Percentage Points of the *t* Distribution

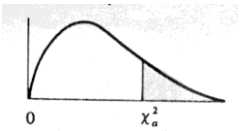


d.f	*a* = .10	*a* = .05	*a* = .025	*a* = .010	*a* = .005
1.	3.078	6.314	12.706	31.821	63.657
2.	1.886	2.920	4.303	6.965	9.925
3.	1.638	2.353	3.182	4.541	5.8411
4.	1.533	2.132	2.776	3.747	4.604
5.	1.476	2.015	5.571	3.365	4.032
6.	1.440	1.943	2.447	3.143	3.707
7.	1.415	1.895	2.365	2.998	3.499
8.	1.397	1.860	2.306	2.896	3.355
9.	1.383	1.833	2.262	2.821	3.250
10.	1.372	1.812	2.228	2.764	3.169
11.	1.363	1.796	2.201	2.718	3.106
12.	1.356	1.782	2.179	2.681	3.055
13.	1.350	1.771	2.160	2.650	3.012
14.	1.345	1.761	2.145	2.624	2.977
15.	1.341	1.753	2.131	2.602	2.947
16.	1.337	1.746	2.120	2.583	2.921
17.	1.333	1.740	2.110	2.567	2.898
18.	1.330	1.734	2.101	2.552	2.878
19.	1.328	1.729	2.093	2.539	2.861
20.	1.325	1.725	2.086	2.528	2.845
21.	1.323	1.721	2.080	2.518	2.831
22.	1.321	1.717	2.074	2.508	2.819
23.	1.319	1.714	2.069	2.500	2.807
24.	1.318	1.711	2.064	2.492	2.797
25.	1.316	1.708	2.060	2.485	2.787
26.	1.315	1.706	2.056	2.479	2.779
27.	1.34	1.703	2.052	2.473	2.771
28.	1.313	1.701	2.048	2.467	2.763
29.	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

** Adapted from OTT, Lyman, et al (Ibid)*

Appendix III

Table II: Percentage Points of the Chi-square Distribution



d.f	***a* = .995**	***a* = .990**	***a* = .975**	***a* = .950**	***a* = .900**
1.	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2.	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3.	0.0717212	0.114832	0.215795	0.351846	0.584375
4.	0.206990	0.297110	0.484419	0.710721	1.063623
5.	0.411740	0.554300	0.831211	1.145476	1.61031
6.	0.675727	0.872085	1.237347	1.63539	2.20413
7.	0.989265	1.239043	1.68987	2.16735	2.83311
8.	1.344419	1.646482	2.17973	2.73264	3.48954
9.	1.734926	2.087912	2.70039	3.32511	4.16816
10.	2.15585	2.55821	3.24697	3.94030	4.86518
11.	2.60321	3.05347	3.81575	4.57481	5.57779
12.	3.07382	3.57056	4.40379	5.22603	6.30380
13.	3.56503	4.10691	5.00874	5.89186	7.04150
14.	4.07468	4.66043	5.62872	6.57063	7.78953
15.	4.60094	5.22935	6.26214	7.26094	8.54675
16.	5.14224	5.81221	6.90766	7.96164	9.31223
17.	5.69724	6.40776	7.56418	8.67176	10.0852
18.	6.26481	7.01491	8.23075	9.39046	10.8649
19.	6.84398	7.63273	8.90655	10.1170	11.6509
20.	7.43386	8.26040	9.59083	10.8508	12.4426
21.	8.03366	8.89720	10.28293	11.5913	13.2396
22.	8.64272	9.54249	10.9823	12.3380	14.0415
23.	9.26042	10.19567	11.6885	13.0905	14.8479
24.	9.88623	10.8564	12.4011	13.8484	15.6587
25.	10.5197	11.5240	13.1197	14.6114	16.4734
26.	11.1603	12.1981	13.8439	15.3791	17.2919
27.	11.8076	12.8786	14.5733	16.1513	18.1138
28.	12.4613	13.5648	15.3079	16.9279	18.9392
29.	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992

40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581
***a* = .10**	***a* = .05**	***a* = .025**	***a* = .010**	***a* = .005**	**d.f.**
2.70554	3.84146	5.02389	6.63490	7.87944	1.
4.60517	5.99147	7.37776	9.21034	10.5966	2.
6.25139	7.81473	9.34840	1.3449	12.8381	3.
7.77944	9.48773	11.1433	13.2767	14.8602	4.
9.23635	11.0705	12.8325	15.0863	16.7496	5.
10.6446	12.5916	14.4494	16.8119	18.5476	6.
12.0170	14.0671	16.0128	18.4753	20.2777	7.
13.3616	15.5073	17.5346	20.0902	21.9550	8.
14.6837	16.9190	19.0228	21.6660	23.5893	9.
5.9871	18.3070	20.4831	23.2093	25.1882	10.
17.2750	19.6751	21.9200	24.7250	26.7569	11.
18.5494	21.0261	23.3367	26.2170	28.2995	12.
19.8119	22.3621	24.7356	27.6883	29.8194	13.
21.0642	23.6848	26.1190	29.1413	31.3193	14.
22.3072	24.9958	27.4884	30.5779	32.8013	15.
23.5418	26.2962	28.8454	31.9999	34.2672	16.
24.7690	27.5871	30.1910	33.4087	35.7185	17.
25.9894	28.8693	31.5264	34.8053	37.1564	18.
27.2036	30.1435	32.8523	36.1908	38.5822	19.
28.4120	31.4104	34.1696	37.5662	39.9968	20.
29.6151	32.6705	35.4789	38.9321	41.4010	21.
30.8133	33.9244	36.7807	40.2894	42.7956	22.
32.0069	35.1725	38.0757	41.6384	44.1813	23.
33.1963	36.4151	39.3641	42.9798	45.5585	24.
34.3816	37.6525	40.6465	44.3141	46.9278	25.
35.5631	38.8852	41.9232	45.6417	48.2899	26.
36.7412	40.1133	43.1944	46.9630	49.6449	27.
37.9159	41.3372	44.4607	48.2782	50.9933	28.
39.0875	42.5569	45.7222	49.5879	52.3356	29.

40.2560	43.7729	46.9792	50.8922	53.6720	30
51.8050	55.7585	59.3417	63.6907	66.7659	40
63.1671	67.5048	71.4202	76.1539	79.4900	50
74.3970	79.0819	83.2976	88.3794	91.9517	60
85.5271	90.5312	95.0231	100.425	104.215	70
96.5782	101.879	106.629	112.329	116.321	80
107.565	113.145	118.136	124.116	128.299	90
118.498	124.342	129.561	135.807	140.169	100

** Adapted from OTT, Lyman, et al (Ibid)*