

EDU 704: STATISTICAL METHODS II



NATIONAL OPEN UNIVERSITY OF NIGERIA

Course Code	EDU 704
Course Title	Statistical Methods II
Course Developer	Dr. Osuji, U.S.A. School of Education National Open University of Nigeria Victoria Island Lagos
Course Writer	Dr. Osuji, U.S.A. School of Education National Open University of Nigeria Victoria Island Lagos
Course Editor	Dr. Femi A. Adeoye School of Education National Open University of Nigeria Victoria Island Lagos
Programme Leader	Dr. J.O. Aiyedun School of Education National Open University of Nigeria Victoria Island Lagos
Course Co-ordinator	Dr. U.S.A. Osuji School of Education National Open University of Nigeria Lagos



NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos

Abuja Annex
245 Samuel Adesujo Ademulegun Street
Central Business District
Opposite Arewa Suites
Abuja

e-mail: centralinfo@nou.edu.ng

URL: www.nou.edu.ng

National Open University of Nigeria 2006

First Printed 2006

ISBN: 978-058-230-4

All Rights Reserved

Printed by GOLD PRINT (SS) LTD
For
National Open University of Nigeria

Table of Content	Page
 Module 1	
Fundamentals of Statistics.....	1
Unit 1 Review of fundamentals of statistics and their applications.....	1 – 9
Unit 2 Basic concepts of statistical methods.....	10 – 16
 Module 2	
Probabilities.....	17
Unit 1 Probabilities – Introduction.....	17 – 24
Unit 2 Probabilities – Interpretations.....	25 – 30
Unit 3 Types of Probability of Events.....	31 – 38
Unit 4 Probabilities – continuation.....	39 – 47
Unit 5 Statistical Decision Theory.....	48 – 54
 Module 3	
Introduction to Inferential Statistics.....	55
Unit 1 Randomization and Sampling techniques.....	55 – 62
Unit 2 Sampling Errors.....	63 – 68
Unit 3 Hypothesis Testing.....	69 – 77
 Module 4	
Z and T – tests	78
Unit 1 z-test I.....	78 – 91
Unit 2 z-test II.....	92 – 100
Unit 3 t-test.....	101 – 112
 Module 5	
Analysis of Variance.....	113
Unit 1 Analysis of Variance: - Introduction.....	113 – 120
Unit 2 Applications and Uses of Analysis of Variance.....	121 – 129
Unit 3 Two way Analysis of Variance.....	130 – 140
Unit 4 Analysis of Co-variance (ANCOVA).....	141 – 150
Unit 5 Prediction and Regression.....	151 – 160

Module 6

Chi – Square tests.....	161
Unit 1 The chi-square test.....	161 – 168
Unit 2 The chi-square – a continuation.....	169 – 176
Unit 3 .Chi – Square: Conclusion	177 – 180

MODULE 1 FUNDAMENTALS OF STATISTICS

Unit 1 Review of Fundamentals of Statistics and their applications.

Unit 2 Basics concepts of statistical methods.

UNIT 1 REVIEW OF FUNDAMENTALS OF STATISTICS AND THEIR APPLICATIONS

Table of Contents

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Students need for statistics
 - 3.2 Importance of statistics to research
 - 3.3 Students' aims in the study of statistics
 - 3.4 A review of work done so far
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

You have gone and worked through EDU 701, (Statistics Methods I), you have come across different terms, concepts, formulas, arithmetical calculations and interpretations, you have also noted the applications of these basics especially in the descriptive statistics. You therefore, no longer react to statistics as a frightful course whose mysteries loom forbiddingly before you. In this unit, we will take a review of the fundamentals importance of statistics and the applications of statistics.

2.0 OBJECTIVES

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

- i. Give reasons why students need statistics
- ii. State why statistics are important in research

- iii. State students aims in the study of statistics
- iv. Precisely review or summarise the content of statistical methods.

3.0 MAIN CONTENT

3.1 Students' need for Statistics

You have seen that in everything you do you must strive to inject as much meaning and significance in your own way, as you can. You have also seen that the proper and optimal use of statistical methods and statistical thinking requires a certain minimal achievement of understanding. You have also noted that to inject adequate meaning and significance to your descriptive ability, places and actions you must use statistics. Let us therefore take note of why students need statistics.

There are four simple but undeniable reasons why students who take a required course in statistics must develop some mastery of the subject. These are:

- i. they must be able to read professional literature. Students should never finish the extension of their skills in the art of reading as this increase their vocabulary. But one cannot read much of the literature in any specialized field in the social sciences, particularly in the behavioural sciences and education without encountering statistical symbols, concepts and ideas on every interval. Therefore, persons who cannot read the average research paper in their field with intelligence and with some appreciation as to whether sound conclusions have been reached are severely limited. This appreciation will require some level of familiarity with basic statistical ideas.
- ii. they must master techniques needed in advanced courses. In every advanced courses, whether it is a laboratory course or a practicum, there are usually certain incidental techniques needed in the operations involved. In any laboratory course or experimental analysis, results cannot be treated or reports written without at least minimal statistical operations, even a field survey or the checking of a report involves inevitable statistical steps.
- iii. statistics are essential part of professional training. Trained psychologists and/or educators, as professionals they are, need statistical logic, statistical thinking operation. Since their practice requires the common technical instruments such as tests, scales,

questionnaires and inventions, psychologists and educators depend upon statistical background in their administration and in the interpretation of their results. You should note that using these tests, scales etc without knowledge of the statistical reasoning upon which they depend is like the medical diagnostician using clinical tests without the knowledge of physiology and pathology.

- iv. statistics are basic to research activities. According to Guilford and Fruchter (1981) the extent that the psychologist or educator intends to keep alive his research interests and research activities; he must as a matter of necessity lean upon his knowledge and skills in statistical methods. Let it therefore be emphasized that in any professional fields where there are still so many unknowns as in the behavioural sciences, the advancement of those professionals and of the competence of their members depends to a high degree upon the continued research attitude and research efforts of those members.

3.2 Importance of Statistics to Research

Now that you have noticed that statistics are very important to research and as well aware that research efforts and research results cannot be published without statistical thinking and operations, let us briefly summarize the advantages of statistics in research.

- i. They permit the most exact kind of description: The goal of science is the description of phenomena. But description can be complete and accurate or useful to anybody who can understand it, when he reads the symbols in terms of which those phenomena are described. Mathematics and/or statistics are a part of the descriptive language and an outgrowth of our verbal symbols particularly adapted to the efficient kind of description which the scientist demands.
- ii. They make us to be definite and exact in our procedure and in our thinking: Statistical operations direct our methods to be definite, statistical logic makes it imperative to be right.
- iii. They enable us to summarize our results in a meaningful and convenient form: Most observations taken are bewildering and meaningless, but statistics provide an unrivaled device for bringing order out of chaos and for seeing the general picture in one's result.
- iv. They enable us to draw general conclusions: There is a process of extracting conclusions from data based researches. This process is

carried out according to acceptable rules. Therefore, by means of statistical steps we can say about how much faith should be placed to any conclusion and how far we may extend our generalization.

- v. They enable us to predict: Statistics are used to predict how much of a thing will happen under such conditions we know and have measured. Statistical methods will also tell us about how much margin of error to allow in making predictions. It is not only making predictions, but we can also know how much faith to place in them.
- vi. They enable us to analyze some of the causal factors underlying complex and otherwise bewildering events: It is generally true in social sciences, psychology and education that any event or outcome is a resultant of numerous causal factors. Since it is not easy to manage people and their affairs sufficiently in experiments the best thing to do is to make a statistical study on the basis of the findings we can predict.

3.3 Students aims in the study of Statistics

Having worked through Statistical Methods 1 (EDU. 701) and having got a general idea of the advantages of statistics to everyday life in general and to research in particular, what do you think will be your aims in your study of statistics? We may summarize the aims as follows:

- i. To master the vocabulary of statistics: Statistics shares the same ordinary symbols for numerical operations with General Mathematics. From your interaction and knowledge of Mathematics, much of the statistical vocabulary is already known to you. This consists of concepts that are symbolized by words and by letter symbols which are substitute for them. As you use them, both the old and new concepts and their meanings will continue to grow.
- ii. To acquire, or to revive, and to extend skills in computation: Computation is very important, and understanding of the statistical concepts comes largely through applying them in computational operations. Computational skills include application of formulas as well as planning efficient operations. These grows with practice.
- iii. To learn to interpret statistical results correctly: When statistical results are correctly interpreted, they can be very useful. They can be as most powerful source of meaning and significance if full and proper interpretations are extracted from data. But when

inadequately interpreted, they may represent something worse than wasted efforts.

- iv. To grasp the logic of statistics: Statistics provides a way of thinking as well as a vocabulary and a language. It is a logical system and like Mathematics, it is peculiarly adaptable to the handling of scientific problems. The mastery of the logical aspects of a research problem before taking to experiment or field trip, proper formulation of a research problem and including a clear consideration of the specific statistical operations to be employed are necessarily the way out of research headaches.
- v. To learn where to apply statistics and where not to: All statistical devices can illuminate data, but each has its own limitations. Every statistic is developed as a purely mathematical idea which rests upon certain assumptions. If the assumptions are true of the particular data with which we have to deal, the statistic may be appropriately applied. Therefore, wherever a statistic is to be applied, there are likely to be mentioned certain assumptions or properties of the situation in which that statistic may be utilized.
- vi. To understand the underlying Mathematics of statistics: In your previous Mathematical training, you have been introduced to analytic geometry or calculus where you have an idea or have grasped many of the mathematical relationships underlying statistics. If you have not, well, never mind, this course will give you a more than commonsense understanding of what goes on in the use of formulas.

Activity 1

1. Give a brief description of the units treated in Statistical Methods 1 (EDU. 701)
2. What are the students' aims in studying statistics?

3.4 A Review of work done so far

Here, we want or have a review of the work you have done so far in Statistical Methods 1. In other words, we want to look at the whole forest which you have passed through. You have gone through the nature of data and you have discovered that different types of data. You are aware that numerical data generally fall into two major categories. Things are counted, and this yields frequencies; or things are measured, and this yields

metric values or scale values. Data of the first kind are often called enumeration data. While data of the second kind are called measurements or metric data. Statistical methods deal with both kinds of data. The word statistics itself has several meaning. It refers to a branch of Mathematics which specializes in enumeration data and their relation to metric data. Statistics constitute a body of scientific methods that are used for the quantitative analysis of data. Statistical procedures help us reduce vast quantities of information to manageable form and to reach reasonable decisions with limited information. You have noted that there are mainly two types of statistics: descriptive statistics which help the researcher in organizing, summarizing, interpreting and communicating quantitative information obtained from observations; and inferential statistics which permit the scientist or researcher to go beyond that data gathered from a small number of subjects to reach tentative conclusions about the large group from which the smaller group was derived.

A fundamental step in the conduct of research is measurement. This is the process through which observations are translated into numbers. In other words and according to Stevens (1951) measurement in its broad sense, is the assignment of numerals to objects, or events according to rules. The nature of the measurements process that produces the numbers determines the interpretation that can be made from them and the statistical procedures that can be meaningfully used with them. The Stevens' scale of measurement which is the most widely quoted taxonomy of measurement procedures classifies measurements into nominal, ordinal, interval and ratio.

You have also noted that the interest of the researcher is to discover the relationships between variables, which are characteristics that show variation from individual to individual or object to object. Variables are classified as (a) quantitative or qualitative (b) discrete or continuous and (c) independent or dependent.

Frequency distributions and graphs are useful for ordering data and presenting it in an easily interpreted form. Frequency distributions can be grouped or ungrouped. Some times a frequency table is more readily interpreted if the data are grouped into class intervals. You will recall that a frequency distribution table which is composite in nature is made up of the real limits, cumulative frequencies, cumulative percentages and relative cumulative frequencies.

Graphs are often useful for communicating a frequency distribution. Under the graphs we have the histogram, the bar chart, frequency polygon,

cumulative frequency polygon, cumulative percentage polygon. Curves resulting from such graphs can be symmetrical or asymmetrical, positively skewed or negatively skewed. You can also measure the skewness and kurtosis.

To communicate characteristics of the frequency distribution in a still more compact way, you calculated and discussed certain descriptive statistics such as; the measures of central tendency; which are useful indices for summarizing a whole set of measures, these are the mean, the median and the mode, the measures of variability, which are necessary for adequately describing quantitative distributions. The three most frequently used measures are the range, the quartile deviation and standard deviations.

You have also seen the methods used to indicate the relative position of scores within a given group. These measures include percentile ranks and standard scores. You have also looked at bivariate distributions and the question of correlation which refers to the extent to which two variables are related in a population.

With these, you are now set for Statistical Methods II, which is mainly inferential statistics.

4.0 CONCLUSION

You have seen that for you to strive to inject as much meaning and significance to everything you do, you must make proper and optimal use of statistical methods. Now that you have been through with Statistical Methods I and have been able to see how to use the descriptive statistics, you are on the right step to continue with Statistical Methods II, which will be mostly on inferential statistics.

5.0 SUMMARY

In this unit, you have learnt that the proper and optimal use of statistical methods and statistical thinking requires a certain minimal achievement of some mastery of the subject, Statistics. You have seen that students need statistics for four major reasons. These are:

1. they must be able to read professional literature
2. they must master techniques needed in advanced courses
3. statistics are essential part of research training and
4. statistics are basic to research activities.

You have also noted the importance of statistics to research. These are:

- i. they permit the most exact kind of description
- ii. they make us to be definite and exact in our procedures and in our thinking
- iii. they enable us to summarize our results in a meaningful and convenient form
- iv. they enable us to draw general conclusions
- v. they enable us to predict
- vi. they enable us to analyze some of the causal factors underlying complex and otherwise bewildering events.

You have known the students aims in the study of statistics. You have also reviewed the contents of the Statistical Methods I in this unit. You can now proceed in your study of statistical methods.

6.0 TUTOR-MARKED ASSIGNMENT

- i. Give reasons why students need statistics.
- ii. Why do you think statistics are important to research?
- iii. What are the students' aims in studying statistics?

7.0 REFERENCES/FURTHER READINGS

Ary, O and Jacobs, L.C. (1976) "Introduction to statistics: Purposes and Procedures". U.S.A. Hold, Rinehart and Winston.

Guilford, J.P. and Fruchter, B (1978) "Fundamental Statistics in Psychology and Education." Auckland, Bogota... Sydne, Tokyo. McGraw-Hill

Stevens, S.S. (1951) "Mathematics, Measurement and Psychophysics" in Handbook of Experimental Psychology. New York, Wiley.

UNIT 2 BASIC CONCEPTS OF STATISTICAL METHODS

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Variables
3.1.1	Quantitative Variables
3.1.2	Qualitative Variables
3.1.3	Independent Variables
3.1.4	Dependent Variables
3.1.5	Discrete Variables
3.1.6	Continuous Variables
3.1.7	Suppressor Variables
3.2	Mathematical Models
3.3	Prediction and Statistics
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

You have just finished the course Statistical Methods I and in the second unit, you learnt about basic concepts in statistics. Those concepts included population, sample, parameter, estimates of statistics, measurement, and errors of measurement.

In this unit, we shall see other concepts such as variables, mathematical models, prediction and statistics. These will help you in the preceding units.

2.0 OBJECTIVES

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

1. list and define the various types of variables used in statistics.
2. explain the importance of mathematical models.
3. explain the relationship between prediction and statistics.

3.0 MAIN CONTENT

3.1 Variables

If you refer to unit IV of the Statistical Methods I you will recall that you have treated types of variables – discrete and continuous. In this unit, we shall look at the variables in details.

Variables are those characteristic which can take on more than one value and which shows variation from person to person or from case to case or from event to event. If for instance, a researcher observes a group of individuals and measures their age, height, weight, intelligence, achievement and attitudes, such characteristics are called variables because one would expect to find variations from person to person of these attributes. The types of variables are: Quantitative, Qualitative, Independent, Dependent, Discrete, Continuous and Suppressor Variables.

3.1.1 Quantitative Variables

These are variables which vary in quantity and are therefore recorded in numerical form, examples are age, test scores, time required to solve problems, height, weight etc.

3.1.2 Qualitative Variables

These are variables which vary in quality and may be recorded by means of a verbal label or through the use of code numbers. Examples are sex, colour of hair or eyes, handedness, shape of face etc.

3.1.3 Independent Variables

These are the variables which are manipulated in an experimental study by the experimenter in order to see what effect changes in those variables have on the other variable which is hypothesized to be dependent upon it. Variations in an independent variable are presumed to result in variations in the dependent variable.

3.1.4 Dependent Variables

These are variables so called because their values are thought to depend on, or vary with the values of the independent variables. Note that in non-experimental studies the independent variable is not manipulated by the

experimenter, but is a pre-existing variable which is hypothesized to influence a dependent variable.

Note also that those characteristics which serve as variables in one study may be kept constant in another study by selecting members of a sample on the basis of similarity in these characteristics. The term constant can be used in reference to characteristics which do not vary for the members of a particular group.

3.1.5 Discrete Variables

These are variables which can take on only a specific set of values. These variables can only yield whole number values, and no fractions. Take for instance, the number of students in a class. This can be 35, 40, 42, 45 etc, but cannot be $35\frac{1}{2}$ or $40\frac{1}{4}$ etc. Family size is another discrete variables. Take your own family for example, your family may be composed of 4, 5, 7 or more people, but values between the numbers or fractions would not be possible, as you cannot have $4\frac{1}{2}$ or $5\frac{1}{2}$ etc. people. Discrete variables may be either qualitative – sex, marital status, handedness, state of origin, nationality, or quantitative – number of books in the library, number of goats in the farm, number of graduate teachers in the school, number of boys doing chemistry in a class etc.

3.1.6 Continuous Variables

These are variables which can assume any values, including fractional values, within a range of values. In other words, continuous variables are measured in both whole and fractional units. Age, height, weight, intelligent test, achievement test scores, daily caloric intake, time, length etc. are all examples of continuous variables. An individual can be described as $12\frac{1}{2}$ years old, 1.7m tall, and weight of 70.3kg etc. These variables are always quantitative in nature and exist along a continuum from the smallest amount of the variable at one extreme to the largest amount possible at the other end. For instance, to go from 25 to 26, one must pass through a large number of fractional parts such as 25.0, 25.1, 25.2 etc. Measurement here is always an approximation of the true value. This is because no matter how accurately you try to measure, it is not possible to measure and record all the possible values of a continuous variable. These are measured therefore, to the nearest convenient unit. Take 25.0 to 25.1 as an example, you will note that from 25.0 you have 25.001, 25.002, 25.003 etc. 25.01, 25.02, 25.03 etc. It can be endless to count from 25.0 to 26.

3.1.7 Suppressor Variable:

A suppressor variable is one whose function in a regression equation is to suppress in other independent variables that variance which is not represented in the criterion but which may be in some variable that does otherwise correlate with the criterion.

Activity 1

Match the definitions below with the type of variables.

- i. A value that is the same for all members of a population.
- ii. A characteristic that classifies a subject into a labeled category.
- iii. A value that is known to or believed to change as another value changes.
- iv. A characteristic that is best described by a number rather than by a verbal label.
- v. A characteristic that can take certain specific values but not intermediate values.
- vi. A value that is known or believed to influence another value.
- vii. A characteristic that can assume any value within a range of values.

Match the above with the following:

- a. quantitative variable
- b. constant
- c. qualitative variable
- d. independent variable
- e. discrete variable
- f. continuous variable
- g. dependent variable.

3.2 Mathematical Models

You have been studying Mathematics up till now. You would have believed that the universe, including man and his behaviour, is constructed along Mathematics lines and that the application of Mathematical ideas and forms in describing it is an undeniably profitable practice. Equations are Mathematical models which are expressions in symbolic form of structural ideas that describe whole range of physical phenomena.

Mathematics exists entirely in the realm of ideas. It is a logic-based system of elements and relationships, all of which are precisely defined. It can be

applied to the description of nature as a completely logical language. This is because the events and objects of nature have properties that provide a sufficient parallel to mathematical ideas. In other words, isomorphism or similarity of forms exists between Mathematical ideas and phenomena of nature. Where this is not completely exact, there will be enough agreement between the forms of nature and the forms of mathematical expression to make the description acceptable. Most of the time, there is an approximation which is often so close that once you have applied the mathematical description, you can follow where the mathematical logic and you come out with deductions that also apply to nature. Statistics as a branch of Mathematics and also a science cannot do without equations or Mathematical models.

Activity 2

List 10 Mathematical models you have come across in Statistical Methods I.

3.3 Prediction and Statistics

Most of the times, you are involved in the operation called prediction even when you do not realize it. As a teacher or a counselor or as a parent, when you tell a child to study certain subjects that will lead him/her to study certain courses in the University or the subjects that will lead him/her to take certain vocations, you are tacitly predicting relative success in one group and relative failure in the others. Take for instance, a medical doctor who diagnosis a patient as having malaria and prescribes certain anti-malaria drugs is invariably predicting improvement under that treatment as opposed to lack of improvement if the treatment is not applied.

Now, let us take another familiar example. After taking a promotion examination and using the results to promote some students to new classes, you are again predicting that these students promoted will adjust and do well in their new classes. We, therefore note that all therapies and administrative decisions imply predictions.

Predictions in education and psychology are often called actuarial. This is because they are made on statistical basis and with the knowledge that only in the long run will the practice represented by any prediction be better than other practices, based upon other predictions in single cases are recognized as involving many chance elements, and therefore the prediction is either correct or incorrect. With large numbers, there are certain probabilities of being right or wrong which can be determined by statistical methods which

provide the basis for choosing what prediction to make and the basis for knowing what the odds are of being right or wrong.

Activity 2

Make ten predictive statements and explain the implications.

4.0 CONCLUSION

Now that you have added the details of such concepts like variables, mathematical models and prediction into your vocabulary and now that you have learnt where, how and when to use them including their importance to both activities in life, statistical methods and research in general, you can now apply them with confidence and appropriately in your research and statistics.

5.0 SUMMARY

In this unit, you have learnt that variables are those characteristics which the researchers observe and measure. Variable is used to refer to a characteristic which can take on more than one value and which shows variation from event to event or from case to case. The variables are:

- i. **Quantitative Variables:** which vary in quantity and are recorded in numerical form.
- ii. **Qualitative Variables:** which vary in quality and may be recorded by means of verbal label or through the use of code numbers.
- iii. **Independent Variables:** which are manipulated in an experimental study by the experimenter in order to see what effect changes in these variables have on the other variables which are hypothesized to be dependent upon it.
- iv. **Dependent Variables:** which depend on or vary with the values of the independent variables.
- v. **Discrete variables:** which can only yield whole numbers and no fractions.

- vi. **Continuous variables:** which can assume any value including fractional values, within a range of values.
- vii. **Suppressor variables:** which are not very common but whose function in a regression equation is to suppress in other independent variables that variance which is not represented in the criterion but which may be in some variable that does otherwise correlate with the criterion.

You have also noted that Mathematics exists entirely in the realm of ideas and that it is a logic-based system of elements and relationship, all of which are precisely defined. It is applied to all the descriptions of nature because events and objects of nature have properties parallel to Mathematics. Isomorphism exists between mathematical ideas and natural phenomena. You have also seen that most activities or operations you do in Education and Psychology involve prediction which is called actuarial because they are made on a statistical basis. With statistical methods you have the basis for choosing what prediction to make.

6.0 TUTOR-MARKED ASSIGNMENT (TMA)

1. Which of these variables below are discrete and which ones are continuous?
 - a. length of hair
 - b. incidents of disruptive behaviour
 - c. amount of anxiety a person manifests
 - d. number of tables in a college library
 - e. intelligence
 - f. number of students who failed an examination
 - g. reading ability
 - h. items passed on an achievement test
 - i. number of medical doctors in a hospital
 - j. number of towns in a state
 - k. money spent on a journey
 - l. attitude to watch home movies.
2. Distinguish between:
 - a. Independent and dependent variables
 - b. Continuous and discrete variables

7.0 REFERENCE/FURTHER READINGS

Ary, D and Jacobs, L.C. (1976) Introduction to statistics: purposes and procedures. New York Chicago... London, Sydney. Holt, Rinehart and Winston

Guilford, J.P. and Fruchter, B. (1981) Fundamental Statistics in Psychology and Education. Auckland, Bogota...Sydney, Tokyo. McGraw-Hill International Book Company.

MODULE TWO PROBABILITIES

Unit 1	Probabilities: Introduction
Unit 2	Probabilities: Interpretations
Unit 3	Types of probability events
Unit 4	Probabilities continued
Unit 5	Statistical Decision Theory

UNIT 1 PROBABILITIES: INTRODUCTION

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
	3.1 Elementary probability and decision making
	3.2 Historical background
	3.3 Definition of terms in probability
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

You are aware that statistics is a very useful tool for mathematical calculations required for decision-making. At one time or the other, man is faced with some types of problems which may require him to take certain decisions. Before you started this programme, you were confronted with a problem requiring answers to the following questions:

- a. What type of programme do I want to undertake?
- b. What qualifications do I have to undertake this programme?
- c. What are the problems I will encounter in the programme?
- d. What preparations do I need to enable me start and finish this programme without a hitch?
- e. What are the advantages of undertaking this programme instead of other programmes?
- f. What are the benefits I will get at the end of the programme?
- g. What are the long run effects of this programme on my economic life?

As you tried to provide answers to these questions, you were operating in the area of probability. This is because there may be more than one answer to each of the questions. You collected information facts and figures before you took the decision to go for this programme in this University. This is a typical example of the theory of probability and statistics at work.

In this unit, you will be looking at elementary probability and decision making.

2.0 OBJECTIVES

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

- i. Explain the term probability
- ii. Trace the historical background of probability
- iii. Define the terms used in probability

3.0 MAIN CONTENT

3.1 Elementary Probability & Decision-making

Sometimes in your life, you are faced with a situation in which you have to make decisions. This decision may be an attempt to propose an answer to solve a problem. This process of decision making may look simple and easy. It is so if all the necessary information needed to take the decision is easily available. But most of the times, it is not always the case. Some of the times, decisions are made on precise set of appropriate mathematical and physical data. These types of decisions pose less difficult in terms of the prospects for failure outcomes. If for instance, you want to buy a car and you consider the shape in relation to motion, you will find that streamlined cars are less resistant to wind during motion than cars that are not streamlined. Therefore, to make a decision on the type or shape of a car to buy, the information above is quite clear to enable you choose a car on the basis of the effect of wind on motion. However, this criterion is not enough to enable you decide on what car to buy. Buying a car requires more certain and uncertain information such as: model, brand, price, size, colour, mechanical efficiency, taste, maintenance cost, type of default from the factory, need, availability of parts, durability among others.

In situations where mathematical data cannot be used adequately to make certain decisions, the subject probability will pay a part in helping in the

attempt to make a correct or better than wrong decisions. The subject probability can be referred to as Mathematics of chance, while probability theory is a branch of Mathematics concerned with the concept and measurement of uncertainty.

Your having a proper knowledge of probability theory is very important for your good understanding of statistical inference or inductive statistics. This is because statistical inference is a technique which enables a decision maker to make probability statements about a population based on representative sample.

Activity 1

Your friend who is a business man is thinking of building a house at Abuja. He is relying on your advice for the initial planning. List necessary questions, the answers of which would help you to advice him.

3.2 Historical background of Probability

You have seen pool stackers in a pool office using certain information to predict teams that would play ‘draw’ in the English league, Australian league etc. You have also seen people in hotels playing the game of chance. Mathematics of chance began with questions raised in connection with these so called ‘games of chance’. Take these probability statements for instance:

- a. Enyimba International F.C. will defeat Sharks FC today
- b. The marriage between Tunde and Jumai will succeed
- c. It may rain today
- d. Holder will get admission into the university this year
- e. A tossed-up coin will fall “heads” or “tails”
- f. Which raffle ticket will be drawn?
- g. Which teams will play the FA Cup finals? etc.

You will note that these are all situations in which the outcome is not certain. The first major attempt at an organized treatment of probability as a subject was made by an Italian Algebraist, Girolamo Cardano (1501 – 1576). His attempt was in the form of what we refer to as a ‘gambler’s manual.’ Some other people also contributed to make the subject popular. These are Christian Huggens (1629-1695) the Dutch Scientist, the great Swiss Mathematician, Jacques Bernoulli (1654-1705) and the French Scientist Pierre S. Laplace (1749-1827).

3.3 Definition of terms in Probability

- i. A die or dice: This is a cube with the six faces marked from 1 through 6, respectively. We shall take it that one of these faces will be uppermost when the die is tossed. The number appearing on the uppermost face is said to be the number thrown or the number scored. When a coin is tossed, it will land either heads or tails.
- ii. Heads: This means the front side or top of the coin usually having the inscription of a human head.
- iii. Tails: This means the back side or bottom of the coin without the inscription of a head.
- iv. Experiment: This means an act performed with uncertain result or performing an act with uncertain result e.g. tossing a coin or dice or drawing a card from many cards.
- v. A Trial: This is one act performed e.g. a toss of a coin, dice or a draw of a card.
- vi. An Outcome: This means one of the possible results of a trial of an experiment.
- vii. An Event: This is a set of possible results or outcomes. It may be described as simple or compound event.
- viii. A Simple Event: This is an outcome of an experiment and it cannot be decomposed or partitioned any further. In set theory, an event is represented by a sample point. So, a simple event is an indecomposable result of an experiment or observation which is represented by one and only one sample point.
- ix. A Compound Event: This is one made up of two or more simple events. In this case, it could be decomposed or partitioned into simple events.
- x. Sample Space: This is the collection of all simple event which make up all the possible outcomes or results of our experiments.

Example 1

Consider the experimental case of three coins tossed at once. We could list the possible combinations of outcomes for the three successive coins to represent a sample space S using H for heads and T for tail we have:

	SCORE	f
HHH	3 heads	1
HHT	2 heads	3
HTH		
THH		
HTT	1 head	3
THT		
TTH		
TTT	0 head	1
	Total	8

Thus $S = (HHH, HHT, HTH, THH, HTT, THT, TTH, TTT)$

You will note that each of these possible results is a simple event which can be denoted by E_1 where $1 = 1, 2, 3, \dots, 8$. Therefore, $E_1 = (HHH)$, $E_2 = (HHT)$, $E_3 = (HTH)$, $E_4 = (THH)$, $E_5 = (HTT)$, $E_6 = (THT)$, $E_7 = (TTH)$, $E_8 = (TTT)$,

Now let us take E_1 to represent a sample point or a particular outcome which corresponds to the event E_1 , we will have that $S = (e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8)$.

Now let us come back to the experiment above. If we take all events or set of events having only two heads in that experiment as A , then the set will be represented by $A = (e_2, e_3, e_4)$. We shall regard this event A as a compound event. This is because A can be partitioned further into some indecomposable events such as (e_2) , (e_3) and (e_4) . These indecomposable events, otherwise called simple events can also be described as subset of event A . Thus, a subset which has 2 or more sample points is a compound event and it is made up of 2^n simple events.

Example 2

If a coin is tossed two times, (i) what will be the maximum sample points
(ii) what will be the sample space?

Solution: The maximum sample points will be 2^4 . This means that 4 sample points are generated. (ii) the sample space $S = (HH, HT, TH, TT) = (E_1, E_2, E_3 \text{ \& } E_4)$.

Activity 2

Given that a coin was tossed 4 times:

- i. Use a table to show the outcomes and
- ii Find the sample space
- iii What is the maximum sample?

Answers to Activity 2

If a coin was tossed 4 times then:

- i. the outcomes will be

SCORE	f	
HHHH	4 Heads	1
HHHT		
THHH		
HTHH	3 Heads	4
HHTH		
HTTH		
HTTH		
THTH	2 Heads	6
TTHH		
HTHT		
THHT		
HTTT	1 Head	4
THTT		
TTHT		
TTTH		
TTTT	0 Head	1

- ii. Sample space $S = (e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}, e_{11}, e_{12}, e_{13}, e_{14}, e_{15}, e_{16})$
- iii. The maximum sample point is $2^4 = 16$

4.0 CONCLUSION

You have seen that probability which started with the game of chance is a basic mathematical idea which is basic to the models used in testing hypothesis and drawing statistical inference which you will see as we go on in this course. The interest in gambling makes wagering on outcomes of events a popular past time. Mathematical descriptions of the known chances in a situation permit much better-informed bases for making bets. Much of the guess work is taking out of gambling of a scientists' evaluation of experimental results. As a researcher or scientist, you should bear in mind that the application of an appropriate probability model may help to state accurately the odds for drawing right or wrong conclusions.

5.0 SUMMARY

The subject probability is referred to as Mathematics of chance designed to help in attempting to reach a correct or better than wrong decisions. So, probability theory is a branch of Mathematics concerned with the concept and measurement of uncertainty. Probability is very important for good understanding of statistical inference which enables decision to be taken about a population based on representative sample. Probability started with the 'games of chance'.

In this unit, you have seen that experiment means performing an act with uncertain result. A trial is one act performed, while an outcome is one of the possible results of a trial of an experiment. An event is a set of possible results or outcomes. It can be simple or compound events. A sample space is the collection of all simple events which make up all the possible outcomes or results of our experiment.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define the following terms as is used in probability.
 - a. Experiment
 - b. Event,
 - c. Simple event
 - d. Compound event
 - e. Sample
 - d. compound event
 - e. sample space

2. Given that in an experiment, a coin is tossed 4 times
 - a. What is the maximum sample points?
 - b. Show the outcomes of the sample space in a table
 - c. Give the sample space S.

7.0 REFERENCES

- Ary, D. and Jacobs, L.C. (1976) Introduction to Statistics Purposes and Procedures. New York, Chicago...Toronto, Sydney. Hld Rinehart and Winston.
- Denga, I.d. and Ali, A (1983) An Introduction to Research Methods and Statistics in Education and Social Sciences. Jos. Savannah Publishers Ltd.
- Guilford, J.P. and Fruchter, B Fundamental Statistics in Psychology (1981) and Education. Auckland, Bogota... Sydney, Tokyo. McGraw Hill Int. Book Company.

UNIT 2 PROBABILITIES: INTERPRETATIONS

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Interpretations of Probability
3.2	Objective Probability
3.3	Subjective Probability
3.4	Axioms of Probability
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References

1.0 INTRODUCTION

In the last unit, you were introduced to probabilities where you learnt that probability which started with the games of chance and which is a Mathematics of chance is very important for decision making. You have learnt some of the terms used in probabilities. We shall now go ahead in this unit to look at the interpretations and the axioms of probability.

2.0 OBJECTIVES

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

- i. Explain the objective interpretation of probability
- ii. Give the subjective interpretation of probability
- iii. Explain the axioms of probability

3.0 MAIN CONTENT

3.1 Interpretations of Probability

3.2 Objective Probability

You are now familiar with the terms experiment, events, sample space, simple event, compound event etc. Let us go a step forward to apply these terms in probability. But first let us define objective probability as the

frequency interpretation of probability based on the symmetry argument. This implies that all the simple events of the sample space are likely to occur on equal bases. If we take it that the simple events are equally likely to occur, then the probability of the occurrence of an event A can be determined by the relative –frequency of the occurrence of A in the total possible occurrence making up the sample space S. We can now define the probability of the occurrence of an event, A thus:-

$$\Pr(A) = \frac{\text{No. of simple events comprising event A}}{\text{Total number of simple events in the sample space S}}$$

Example 3.3

Using the sample space S given above, find the probability of solution:

Sample space given = $S = e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8 = 8$

Event A given = $A = e_2, e_3, \& e_4 = 3$

$$\square \quad \Pr(A) = \frac{\text{No of simple events comprising events A}}{\text{Total No of simple events in the sample space}}$$

$$\square \quad \Pr(A) = \frac{3}{8}$$

You will have to note that this type of interpretation of probability that deals simply with the relative frequency of occurrence of an event in a sample space instead of using the long run can otherwise be called classical interpretation. You will also note that for this type of objective probability to exist the following conditions must be satisfied:

- i. The experiment generating the particular event must be capable of being repeated a number of times or so many times.
- ii. Each trial of the experiment must assume statistical regularity. In order words, the relative frequency of obtaining a particular event at the long run shows stability towards a particular value.
- iii. The probability of the occurrence of a particular event is approximately equal to the relative frequency of observing the event in the total experiment.
- iv. The probability is based on experience. That is to say, after a long period, a particular value may be assigned to the event.

You will also note that mathematically, a probability is symbolized by P or Pr. This may range from zero when there is no chance whatsoever of the favoured event to 1.0 or Unity, where there is absolute certainty i.e. nothing

else could happen. Other uncertain events will lie between a probability of zero to less than unity. Probability range is from 0 to 1.

Activity 1

Given that sample space has a maximum sample point of 24. Event A has e_2, e_3, e_4, e_5, e_6 , and B has e_{15}, e_{14}, e_{13} .

Give the sample space and find the probability of A and B.

Answer to Activity 1

The maximum sample points = $2^4 = 16$

Sample space = $(e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}, e_{11}, e_{12}, e_{13}, e_{14}, e_{15}, e_{16})$

Event A = $e_2, e_3, e_4, e_5, e_6 = 5$

Event B = $e_{15}, e_{14}, e_{13} = 3$

$$\Pr(A) = \frac{\text{No. of simple events in event A}}{\text{Total No. of simple events in the sample space}}$$

$$\square \quad \Pr(A) = \frac{5}{16}$$

$$\Pr(B) = \frac{3}{16}$$

3.2 Subjective Probability

You have looked at objective probability in terms of its frequency interpretation and based on symmetry argument. But in the case of subjective probability, we talk about the measure of the degree of confidence which an individual may place in the occurrence of an uncertain event. It is not dependent on the repeatability of any process, it can be applied to such events which can occur only once and once only. As a result of the non-repeatability of the events, the subjective interpretation statement of probability cannot be explained in terms of the long run relative frequency.

The interpretation of subjective probability can be as a qualified judgement of a particular individual. In other words, the individual concerned places his personal judgement on the event in question as regards the occurrence.

It is perfectly reasonable to assign a probability to an event reflecting the individual's assessment of the non-repetitive situation. When the individual is certain of the occurrence of the event, he can then assign one to the probability. But if he is sure that the event would not occur he assigns zero as the probability. In other words, all other uncertain events are assigned probabilities lying between zero and one (0-1).

3.3 Axioms of Probability

Let us look at some of the rules in probability:

(a) The probability of an event A is taken to be a real number which is non-negative. Let us use, for instance some experiments with n possible events, A_1, A_2, A_3, \dots . And then the probability of each event such as A_1 , where $1 = 1, 2, 3, \dots, n$ is non-negative real number which is given by $0 \leq \Pr(A_1) \leq 1$.

(b) If S is the sample space, then the probability of S , $\Pr(S)$ is given by $\Pr(A_1) + \Pr(A_2) + \Pr(A_3) \dots + \Pr(A_n) = 1$ thus $\Pr(A_1 \cup A_2 \cup A_3 \dots \cup A_n) = \Pr(A_1) + \Pr(A_2) + \Pr(A_3) + \dots + \Pr(A_n)$

$$\square \quad \Pr(S) = 1$$

Provided $A_i \cap A_j = \phi$ for all values of $i \neq j$

(c) If A_i and A_j are two mutually exclusive events, then the probability that at least one of the two events will occur is equal to the sum of their probabilities.

$$\square \quad \Pr(A_i \text{ or } A_j) = \Pr(A_i \cup A_j) = \Pr(A_i) + \Pr(A_j), \text{ since } A_i \cap A_j = \phi$$

where \cup = union and \cap = intersect.

When you look at these three axioms of probability, you can conclude or say that the probability of an event is a real number which lies between zero and one; inclusive. Thus $0 \leq \Pr(A) \leq 1$.

Again, if A_1 is a subset of event A_2 , then the probability of A_2 must be greater than or equal to the probability of A_1 . That is to say that $\Pr(A_1) \leq \Pr(A_2)$

Activity 2

What do these symbols mean:-

- i. \leq ii. \cup iii. \cap iv. \neq v. $\Pr(A)$ vii. S

Answer to Activity 2

- i. \leq = Less than or equal to
 ii. \cup = Union
 iii. \cap = Intersect
 iv. \neq = Not equal
 v. $\Pr(A)$ = Probability of A
 vi. S = Sample space

4.0 CONCLUSION

In this unit, you have seen some of the interpretation of the rules of probability. You have seen the objective and subjective interpretations. We shall now apply them as we go along to the next unit.

5.0 SUMMARY

In this unit, we have tried to give the interpretations of probabilities. You learnt that probability can be interpreted using:

- (i) the objective probability which is referred to as the frequency interpretation that is based on the symmetry argument and which implies that all the simple events of the sample space are likely to occur on equal bases.
- (ii) Subjective probability which is a measure of the degree of confidence which an individual has in the occurrence of an uncertain event. You have also seen the axioms of probability that (i) the probability of an event A is taken to be a real number which is non-negative (ii) if S is a sample space, then the probability of S $\Pr(S) = 1$ and
- (iii) If A_i and A_j are two mutually exclusive events, then the probability that at least one of the two events will occur is equal to the sum of their probabilities.

6.0 TUTOR-MARKED ASSIGNMENT

- i. What is the objective interpretation of probability?
- ii. If the maximum sample point of a sample space is 25 and Event A has e_3, e_4, e_5, e_6, e_7 , and e_8 . Event C has e_{32}, e_{31}, e_{30} and e_{29}
 - a. Give the sample space
 - b. Find the probability of A and C.

7.0 REFERENCES/FURTHER READINGS

Ary, D and Jacobs, L.C. (1976) Introduction to Statistics purposes and Procedures. New York, Chicago...Toronto, Sydney. Holt Rinehart and Winston.

Guilford, J.P. and Fruchter, B. (1981) Fundamental Statistics in Psychology and Education. Auckland, Bogota...Sydney, Tokyo. McGraw Hill Int. Book Company.

UNIT 3 TYPES OF PROBABILITY OF EVENTS

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Types of Probability Events
3.1.1.	Mutually Exclusive Events
3.1.2	Not Mutually Exclusive Events
3.1.3	The Complement of Events
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References

1.0 INTRODUCTION

You have been introduced to probabilities where you learnt that probability which started with the games of chance and which is a Mathematics of chance is very important for decision making. You have learnt some of the terms used in probabilities. In this unit, you will be learning about the types of probability events and how to use them.

2.0 OBJECTIVES

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

- i. Explain the types of probability events
- ii. Do some elementary calculations in probability
- iii. Apply the rules of probability.

3.0 MAIN CONTENT

3.1 Types of Probability Events

3.1.1 Mutually Exclusive Events

When two events do not have any common sample points, the events are said to be mutually exclusive. That is to say, that if the occurrence of A_1 excludes the occurrence of A_2 , we then say that events A_1 and A_2 are

mutually exclusive. In other words, both events A_1 and A_2 cannot occur at the same time.

You have noted from axiom in unit 4, that $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$ since $A_1 \cap A_2 = \phi$. Therefore $A_1 \cup A_2$ can be interpreted as “at least one of the events occurs” or that event A_1 or A_2 or both events occur. This is called the ADDITION LAW or THEOREM. Now let us look at the practical explanation of this law. We can ask some questions regarding events in games of chance.

Example 1

In tossing a die, what is the probability of either 1 or 2 coming up?

Solution:-

The chance or probability of 1 to come in one throw is 1 out of 6 chances. i.e. $\frac{1}{6}$. For 2 to come, it is also $\frac{1}{6}$. Therefore, there are two ways in which the favoured event can occur out of a total of 6 ways. \therefore by definition, the probability is $\frac{2}{6}$.

Note that this probability is the sum of the two separate probabilities. In other words, the probability is additive.

Example 2

What is the probability of drawing i.e. King or Queen from a set of playing cards numbering 13?

Solution:

Probability of drawing a King = $\frac{1}{13}$.

Probability of drawing a Queen = $\frac{1}{13}$. But the probability of alternative outcomes is the sum of the probabilities of the outcomes taken separately.

So the probability of either a King or Queen is $\frac{1}{13} + \frac{1}{13} = \frac{2}{13}$.

Example 3

In tossing 4 coins, what is the probability of obtaining
(a) HTHT? (b) 4 Heads? (c) 3 Tails? (d) 4 Heads or 3 Tails?

(e) no Heads in two successive tosses?

Solution:

The maximum sample points = 2^4 . Total number of sample events in the sample space = $2^4 = 2 \times 2 \times 2 \times 2 = 16$

- (a) HTHT is one simple event has a probability of $\frac{1}{16}$
- (b) 4 Heads = HHHH = 1 out of 16 times = $\frac{1}{16}$
- (c) 3 Tails will appear 4 times $\frac{4}{16} = \frac{1}{4}$
- (d) 4 Heads or 3 Tails = $\frac{1}{16} + \frac{4}{16} = \frac{5}{16}$
- (e) For 2 tosses, total number of events in the sample space is given by $16^2 = 16 \times 16 = 256$.

Number of no heads will be given by $\frac{1}{16} \times \frac{1}{16} = \frac{1}{256}$

Example 4

In tossing two dice, what is the probability of obtaining

- (a) A_2 and A_3 (b) A_2 and then A_3 ?
- (c) A total of 5 spots?
- (d) On repeating a throw of two dice (i.e. throwing two dice twice in succession) what is the probability of repeating exactly seven spots?

Solution:

Maximum sample points = 6^2

Total number of simple events in the sample space = $6 \times 6 = 36$

(a) To obtain A_2 and $A_3 = \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right) = \frac{1}{36} + \frac{1}{36} = \frac{2}{36} = \frac{1}{18}$

(b) To obtain $A_2 = \frac{1}{36}$ and then $A_3 = \frac{1}{36} = \frac{1}{36}$

(c) To obtain a total of 5 spots = 4.1, 3.2, 2.3, 1.4 i.e. 4 times out of 36
 $= \frac{4}{36} = \frac{1}{9}$

(d) Total number of events in one throw = 36

□ total number of events in two throws = $36 = 1296$

For 7 spots we have 6.1, 5.2, 4.3, 3.4, 2.5, 1.6 i.e. a total of six

events in one throw.

For two throws, it now becomes $6^2 = 36$

The probability of having 7 spots in two throws will then be

$$\frac{36}{36} \times \frac{1}{36} = \frac{1}{36}$$

Example 5

Find the probability of obtaining 5 or 7 as the sum of two dice.

Solution:

The total number of simple events in the sample space for a throw of 2 dice will be $6^2 = 36$. If we take A_1 to give 7 as sum of two dice, then $A_1 = (6.1, 5.2, 4.3, 3.4, 2.5, 1.6) = 6$ events

$$\square \quad \text{Probability of } A_1, \text{ given by } \Pr(A_1) = \frac{6}{36}$$

Again, if we take A_2 to give 5 as sum of two dice, we have $A_2 = (4.1, 3.2, 2.3 \text{ and } 1.4)$ it means that the probability of A_2 given by $\Pr(A_2) = \frac{4}{36}$. But

$$\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) = \frac{6}{36} + \frac{4}{36} = \frac{10}{36} = \frac{5}{18}$$

Activity 1

- (1) A teacher made a set of 26 cards corresponding to the 26 letters of the English alphabets. He asks his pupils to draw and name the alphabet drawn. What is the probability of a pupil drawing Q or P?
- (2) If two dice are thrown, show the probability of obtaining (a) 8 spots (b) 6 spots (c) if the two dice are thrown two times, what is the probability of repeating exactly 6 spots?
- (3) What is the probability of obtaining 6 or 8 as the sum of two dice?

3.1.2 Not Mutually Exclusive Events

You have seen that when two events do not have any common sample points or if the occurrence of one event excludes the occurrence of the other event, we say they are mutually exclusive events. It means therefore that two events A_1 and A_2 are said not to be mutually exclusive if the events A_1

and A_2 contain one or more common sample points. You will recall that in the mutually exclusive events we have that $\Pr(A_1 \cup A_2) = \Pr A_1 + \Pr(A_2)$ since $A_1 \cap A_2 = \phi$. But in the not mutually exclusive events, we have that $\Pr(A_1 \cup A_2) = \Pr A_1 + \Pr(A_2) - \Pr(A_1 \cap A_2)$ where $A_1 \cap A_2 \neq \phi$. This means that both A_1 and A_2 can occur simultaneously. This equation is called the General Addition Law of Probability.

Example 6

Assuming that we have 2 dice and we throw the 2 dice at once. What is the probability that 5 shows on the dice.

Solution:-

Let $A_1 = 5$ shows on the first dice

$A_2 = 5$ shows on the second dice.

Then $A_1 = [(5.1), (5.2), (5.3), (5.4), (5.5), (5.6)]$

$A_2 = [(1.5), (2.5), (3.5), (4.5), (5.5), (6.5)]$

□ $A_1 A_2 = (5.5)$

$\Rightarrow \Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)$

Note that 5 shows in 6 events out of 36 events in A_1 and A_2 that means

$$\frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$$

Activity 2

Repeat the experiment with 2 dice and find the probability that (a) 4 (b) 6. Show on the dice.

3.1.3 The Complement of Event

Now that you can explain the difference between mutually exclusive events and not mutually exclusive events we can move to the third event called the complement of event. If we take A as an event and it does NOT occur in the experiment, it is said to be the complement of an event. It is denoted by \bar{A} , i.e. 'bar A' or A^1 , i.e. 'A prime', while the probability is written thus: $\Pr(\bar{A}) = 1 - \Pr(A)$ or $\Pr(A^1) = 1 - \Pr(A)$

Example 7

If 3 coins are tossed once, find the probability that at least a head does not appear in the toss.

Solution:

A toss of 3 coins will generate a sample space: $S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$. If we take all events where at least a head shows to be A i.e. $A = \{HHH, HHT, HTH, THH, TTH, THT, HTT\}$ i.e. 7 events out of 8 then $\Pr(A) = \frac{7}{8}$.

But the complement of $A = \bar{A}$

$$\square \quad \Pr(\bar{A}) = 1 - \Pr(A) = 1 - \frac{7}{8} = \frac{1}{8}$$

Example 8

If 3 coins are tossed once, what is the probability that 2 heads do not show up?

Solution:

The sample space generated when 3 coins are tossed once $S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$. Now let A be all other events except where 2 heads show. $\bar{A} = \{HHT, HTH, THH, TTH, THT, HTT, TTT\}$. i.e. 7 events out of 8. But $\Pr(\bar{A}) = 1 - \Pr(A)$.

$$\square \quad \Pr(\bar{A}) = 1 - \frac{3}{8}$$

4.0 CONCLUSION

You have gone through the rules and regulations in probability called axioms of probability. You have also seen some of the probability events and the calculations involved. You have seen that it is a very interesting aspect of statistics. You can now make use of this mathematics of chance as it applies to the real life situation.

5.0 SUMMARY

In this unit, you saw the axioms of probability such as

- (a) the probability of an event A is taken to be a real number which is not negative.
- (b) the probability of a sample space S is the sum of the probabilities of all the events in the sample space.
- (c) when there are two mutually exclusive events A_1 and A_2 the probability that at least one of the two events will occur is equal to the sum of their probabilities.

In the types of probability events, you learnt that when two events do not have any common sample points, the events are said to be mutually exclusive. But if the two events contain one or more common sample points, they are said to be not mutually exclusive, while an event which does not occur in an experiment is referred to as the complement of an event, denoted by \bar{A} or A^c .

6.0 TUTOR-MARKED ASSIGNMENT

1. In teaching a class of pupils geometric forms and shapes, a teacher made 17 cards containing the shapes and forms. If a pupil is asked to draw from the cards, what is the probability of this pupil drawing a triangle or a pentagon?
2. In tossing 3 coins, what is the probability of obtaining
 - (a) HHT
 - (b) THT
 - (c) 2 Heads,
 - (d) 3 Heads or 2 Tails
 - (e) No Tail in two successive tosses?
3. What is the probability of getting 6 or 8 as the sum of two dice?
4. If 3 coins are tossed once, what is the probability that two tails do not show up?

7.0 REFERENCES/FURTHER READINGS

Ary, D and Jacobs, L.C. (1976) Introduction to Statistics purposes and Procedures. New York, Chicago...Toronto, Sydney. Hld Rinehart and Winston.

Guilford, J.P. and Fruchter, B Fundamental Statistics in Psychology (1981) and Education. Auckland, Bogota...Sydney, Tokyo. McGraw Hill Int. Book Company.

UNIT 4 PROBABILITIES CONTINUED

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Conditioned Probability
3.2	Multiplication Rule
3.3	Independent Events
3.4	Dependent Events
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

In the last unit, you learnt that when there are two events A_1 and A_2 occurring in an experiment, and the events do not possess any sample in common in the sample space, the events are said to be mutually exclusive. You learnt also that if the events have one or more sample points in common they are said to be non-mutually exclusive, and where our event A does not occur, it is called the compliment of an event \bar{A} or A^c . In this unit, we shall look at conditional probability and the multiplication rule, the independent and dependent events.

2.0 OBJECTIVES

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

- i. Explain conditional probabilities
- ii. Do calculations involving the multiplication rule
- iii. Find the probabilities of independent events
- iv. Explain dependent events in relation to independent events

3.0 MAIN CONTENT

3.1 Conditional Probability

This refers to the probability that an event will occur because another particular event has occurred or will occur. In other words, event A_2 on the condition that event A_1 has occurred or will occur. Let us use the political situation during an election in a state as an example. If there are 3 parties contesting elections in state X. Two elections were slated to be held. If in the first election which is the House of Assembly election, party A won majority of the seats in the house. You can confidently say that party A will win the next election which is the governorship election. Again if there is a continuous advertisement of a product A. Then there is the probability that the price/demand will be high.

The conditional probability of an event B, given that A has occurred may be denoted by $\Pr(B/A)$ which is defined as $\Pr(B/A) =$

$$\frac{\Pr(A \cap B)}{\Pr(A)} \text{ where } \Pr(A \cap B) \neq 0$$

You will take note that the conditional probability of the event B when event A has occurred equals the joint probability of the two events A and B divided by the probability of the given event A. The joint probability of A and B is simply the probability of the intersection of A and B. In order, to determine the conditional probability let us look at the multiplication rule.

3.2 Multiplication Rule

Do you remember the addition theorem or rule which you treated in the last unit! Look at it again. $\Pr(A_1 \cup A_2) = \Pr(A_1 \text{ or } A_2) = \Pr(A_1) + \Pr(A_2)$, since $A_1 \cap A_2 = \phi$.

Now, compare that addition rule with this one:

$\Pr(A \cap B) = \Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B/A)$. This is the same as the joint probability which gives the simultaneous occurrence of two events A and B. Therefore, the joint probability of event A and B is equal to the conditional probability of B given A multiplied by the marginal probability of A. In other words, if given the two events A and B, the probability that the two events will occur together is equal to the probability that the first will occur multiplied by the conditional probability that the second will

occur. The probability of the single occurrence of a particular event A is referred to as the individual marginal probability of event B.

Example 1

120 people went to a hospital for tuberculosis examination. They are classified by the presence of the disease and smoking habit as in the table below. If one of them is selected at random, the probability that he suffers from tuberculosis is 0.625 using the data in the table. Calculate

- The probability that the person examined suffers from tuberculosis given that he/she is a smoker.
- The probability that the person examined suffers from tuberculosis given that he/she is a non-smoker.

Smoking Habit

Disease	Smoker	Non-Smoker	Total
Tuberculosis	70	5	75
Non-tuberculosis	15	30	45
Total	85	35	120

Solution

Let S represent smoker. \bar{S} represents non-smoker. T represent tuberculosis, \bar{T} represent non-tuberculoid. Now, let us change the table above into a probability schedule using probability of x as $\frac{x}{120}$ (i.e. the total of all events).

Smoking Habit

Disease	S	\bar{S}	Marginal Probability
T	$\Pr(TS) = 0.583$	$\Pr(T\bar{S}) = 0.042$	$\Pr(T) = 0.625$
	$\Pr(\bar{T}) = 0.125$	$\Pr(\bar{T}\bar{S}) = 0.250$	$\Pr(\bar{T}) = 0.371$
Marginal Probability	$\Pr(\bar{S}) = 0.708$	$\Pr(\bar{S}) = 0.292$	1

$$(a) \quad \Pr(T/S) = \Pr(TS)/\Pr(S) = \frac{0.583}{0.708} = \underline{\underline{0.8235}}$$

$$(b) \quad \Pr(T/\bar{S}) = \Pr(T\bar{S})/\Pr(\bar{S}) = \frac{0.042}{0.292} = \underline{\underline{0.1438}}$$

Example 2

A private school situated in Enugu admitted 420 pupils based on their performances in their entrance examination results. After a period of three years, the students took the JSCE. If a student is picked up at random for an interview, use the table below to calculate: (a) the probability that the student passed the JSCE, given that he/she passed the Entrance Examination (b) the probability that the student passed the JSCE given that he/she did not pass the Entrance Examination.

Entrance	About cut-off marks	Below cut-off marks	Total
JSCE Passed	230	55	285
Failed	30	105	135
Total	260	160	420

Solution:

If P represents passed in JSCE, \bar{P} represents failed in JSCE. A represents passed above cut-off, \bar{A} represents below cut-off.

Then the table above becomes:-

Entrance	A	\bar{A}	Marginal Probability
JSCE	$\Pr(PA) = 0.5476$	$\Pr(P\bar{A}) = 0.1310$	$\Pr(P) = 0.6786$
\bar{P}	$\Pr(\bar{P}A) = 0.714$	$\Pr(\bar{P}\bar{A}) = 0.2500$	$\Pr(\bar{P}) = 0.3214$
Marginal Probability	$\Pr(A) = 0.6290$	$\Pr(\bar{A}) = 0.3810$	1

- $\Pr(P/A) = \Pr(PA) / \Pr(A) = \frac{0.5476}{0.6290} = 0.8705882 = \underline{\underline{0.8706}}$
- $\Pr(P\bar{A}) = \Pr(P\bar{A}) / \Pr(\bar{A}) = \frac{0.1310}{0.3810} = 0.343832 = \underline{\underline{0.3438}}$

Example 3

One day, at 1.00pm, a branch of a bank in Owerri had 30 customers waiting for service inside the bank. 20 of the customers were men and the rest were women. If a customer was selected at random to have a chat with the

branch manager, without replacement, what is the probability that the second customer selected was a woman?

Let M represent a man and W represent a woman. These are two possible ways of selecting a woman as the second customer. These are (W_1W_2) or (M_1M_2) . The probability of selecting 2 women is given by $\Pr(W_1W_2)$. But $\Pr(W_1W_2) = \Pr(W_1) \cdot \Pr(W_2/W_1)$

$$\Rightarrow \frac{10}{30} \times \frac{9}{29} = \frac{90}{870} = 0.1035$$

Secondly, if the first person is a man followed by the second person that is, a woman i.e. (M_1W_2)

$$\text{Then } \Pr(M_1W_2) = \Pr(M_1) \cdot \Pr(W_2/M_1) = \frac{20}{30} \times \frac{10}{29} = \frac{200}{870} = 0.2299$$

□ The probability that the 2nd selection is a woman will be given by $\Pr(W_1W_2 \cup M_1W_2) = \Pr(W_1W_2) + \Pr(M_1W_2)$
 $= \Pr(W_1) \cdot \Pr(W_2/W_1) + \Pr(M_1) \cdot \Pr(M_1W_2)$

$$\text{But } \Pr(W_1W_2) = 0.1035$$

$$\Pr(M_1W_2) = 0.2299$$

$$\square \Pr(W_1W_2 \cup M_1W_2) = 0.1035 + 0.2299 = \underline{\underline{0.3334}}$$

Example 4

In a college, 60 students are brought to the principal's office for questioning. If 40 of them are boys and the remaining 20 are girls. If the students are taken randomly without replacement for the interrogation, what is the probability that the second person called is a girl?

Solution:

Let B = Boy and G = Girl. The two possible ways of getting a girl as the second person are:- (G_1G_2) or (B_1B_2) . The probability of getting two girls is given by $\Pr(G_1G_2)$.

$$\text{But } \Pr(G_1G_2) = \Pr(G_1) \cdot \Pr(G_2/G_1)$$

$$= \frac{20}{60} \times \frac{19}{59} = \frac{380}{3440} = 0.1104651 = \underline{\underline{0.1105}}$$

If the first person is a boy and the second person is a girl, then we have (B_1G_2) given by $\Pr(B_1G_2) = \Pr(B_1) \cdot \Pr(G_2/B_1)$

$$= \frac{40}{60} \cdot \frac{20}{59} = \frac{800}{3540} = 0.2259887 = \underline{\underline{0.2260}}$$

□ The probability that the second person is a girl will be given by $\Pr(G_1G_2 \cup B_1G_2) = \Pr(G_1G_2) + \Pr(B_1G_2)$
 $= \Pr(G_1) \cdot \Pr(G_2/G_1) + \Pr(B_1) \cdot \Pr(G_2/B_1)$ $\Pr(G_1G_2) = 0.1105$

$$\Pr(B_1G_2) = 0.2260 \quad \Pr(G_1G_2 \cup B_1G_2) = 0.1105 + 0.2260 = \underline{\underline{0.3365}}$$

Activity 1

i) The Okigwe Zonal Education Board invited 600 applicants for interview to fill vacant teaching positions in the different schools in the zone. The teachers are classified according to their disciplines which are Science and Technology only. If the teachers are invited at random for the interview, use the table below to calculate:

1. the probability that the 1st teacher called is Technical teacher given that he/she has a degree.
2. the probability that the 1st teacher called is a Technical teacher given that he/she has no degree.

Class of Teacher	Have Degrees	Have no Degrees	Total
Technical	250	80	330
Science	150	120	270
Total	400	200	600

ii) 40 members of a youth association went for a general meeting of the association. If 25 of them are females and 15 are males, select members randomly to give the closing prayers. What is the probability that the second person selected is a male?

3.3 Independent Event

Now that you have seen some of the probability events such as mutually exclusive events, not mutually exclusive events and the complements of events, let us look at the next event called independent event.

Two events A and B are said to be independent if the occurrence of their joint probability is equal to the product of the individual or marginal probabilities. Thus $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$ so that $\Pr(B) = \Pr(B/A)$

meaning that the knowledge that A has occurred has no effect on the probability of event B.

Example 5

A farmer hired the services of two labourers to help him in his farm. What is the probability that the first labourer is a male and the second is a female?

Solution:-

Let us take that A = (1st labourer hired is a male)
B = (2nd labourer hired is a female)

Take note that the hiring of a male or female in the first or second case does not influence the other hiring.

$$\square \Pr(A \cap B) = \Pr(A) \cdot \Pr(B). \text{ But } \Pr(A) = \frac{1}{2} \cdot \Pr(B) = \frac{1}{2}$$

$$\text{So } \Pr(A \cap B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Example 6

In throwing two dice, what is the probability that 12 spots (i.e. a pair of 6s) will be thrown?

Solution:

If we take the throw of first dice = A and the throw of second dice = B.

The probability of A = $\frac{1}{6}$ and B = $\frac{1}{6}$

$$\square \Pr(A \cap B) = \Pr(A) \cdot \Pr(B) \text{ will be given by } \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Activity 2

- (i) The principal sent for two teachers for an important assignment. What is the probability that the first teacher is a man and the second teacher is a woman?
- (ii) Assuming that you are throwing two dice, what is the probability that a pair of 5 (i.e. ten spots) will appear?

3.4 Dependent Events

Two events A and B are said to be dependent if the probability of B is not equal to the probability of B/A i.e. $\Pr(B) \neq \Pr(B/A)$. Thus, the joint probabilities of A and B will be given by $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B/A)$.

Example 7

What is the probability of obtaining the sum of 5 from throwing 2 dice given that one of them must show I?

Solution:

If we take A = (at least one shows up), then
B = (the sum is 5).

□ $A \cap B$ = at least I and the sum 5 occur simultaneously so

$$A = [(1.1), (1.2), (1.3), (1.4), (1.5), (1.6), (2.1), (3.1), (4.1), (5.1), (6.1)]$$

$$\Pr(A) = \frac{11}{36}$$

$$B = \text{sum is 5} = (1.4) (2.3) (3.2) (4.1) = 4 \text{ events.}$$

$$\text{But } A \cap B = (1.4) (4.1) = 2 \text{ events } \Pr(B/A) = \frac{2}{11}$$

$$\square \Pr A \cap B = \Pr(A) \cdot \Pr(B/A) = \frac{11}{36} \times \frac{2}{11} = \frac{2}{36} \text{ so if } \Pr(B) = \frac{4}{36}$$

Then $\Pr(B) \neq \Pr(B/A)$

Activity 3

What is the probability of getting the sum 6 from throwing 2 dice, given that one of them must show I.

4.0 CONCLUSION

In this unit, you have been introduced to probability which is one aspect of techniques used in scientific inquiry. Since you have noted that the outcomes of scientific inquiry are usually probability statements rather than facts, you will also note that in the units immediately ahead, we shall be concerned almost exclusively with the estimations of values and with the question of how much confidence we should have in those estimates.

5.0 SUMMARY

In this unit, you have learnt that the probability that an event will occur because another particular event has occurred or will occur is known as conditional probability. In the multiplication rule, you have seen that if

given two events A and B, the probability that the two events will occur together is equal to the probability that the first will occur, multiplied by the conditional probability that the second will occur. Again, you have seen that two events A and B are said to be independent if the occurrence of their joint probability is equal to the product of the individual or marginal probabilities, while they are said to be dependent if the probability of B is not equal to the probability of B/A.

6.0 TUTOR-MARKED ASSIGNMENT

- i) A private manufacturing company invited 500 candidates for an oral interview for employment. The invitation was based on their performance during an aptitude test held earlier. If a candidate is picked up for the interview at random, use the table below to calculate.

Aptitude Test Result	Above 50%	Below 50%	Total
SSCE Result			
Passed	250	60	310
Failed	100	90	190
Total	350	150	500

- i. The probability that the candidate passed SSCE given that he/she passed above 50% in the aptitude test
- ii. The probability that the candidate passed SSCE given that he/she failed below 50% in the aptitude test.
- ii) 45 people are arraigned as witnesses before a police detective. If 30 of them are males and 15 are females, if the people are taken at random without replacement for questioning. What is the probability that the second person called is a female?

7.0 REFERENCES/FURTHER READINGS

Ary, D and Jacobs (1976) Introduction to Statistics Purposes and Procedures. New York, Chicago Atlanta...Sydney, Toronto. Holt, Rinehart and Winston.

Guildford, J.P. and Fruscher, B (1981) Fundamental Statistics in Psychology and Education. Auckland,.. Sydney, Tokyo.McGraw-Hill International Book Company.

UNIT 5 STATISTICAL DECISION THEORY

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Statistical Decision Theory
3.2	Decision Diagrams
3.3	The Binomial Expansion
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References

1.0 INTRODUCTION

Decision diagrams are used as a means of using deductive logic to arrive at probabilities based on a priori assumption. You will recall that probability is regarded as Mathematics of chance which is designed to help you to arrive at a better decision. In this unit, you will be going through the decision diagrams and the binomial expansion to conclude the series of units in these module-probabilities.

2.0 OBJECTIVES

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

- i. Explain the decision diagram
- ii. Draw a decision diagram
- iii. Explain the binomial expansion

3.0 MAIN CONTENT**3.1 Statistical Decision Theory****3.2 Decision Diagrams**

You will recall that we have said that probability can be referred to as Mathematics of chance which is designed to help in attempting to reach a correct or better than wrong decisions. The decision diagram is one of the various ways of using deductive logic to arrive at probabilities based on a

priori assumptions. It is a convenient method to represent conditional probabilities which is the same as tree diagram and consists of Nodes and Branches. The branches are called “Chance for”. This is because the chance indicates that there is uncertainty of outcomes. Let us consider as an example three couples producing children. We assume that males and females are equally likely to be delivered. In other words, there is a 50% chance of a son and a 50% chance of a daughter. If we apply our knowledge of deductive logic, we will note that if the first couple have a 50% chance of having a son and 50% chance for the second couple to have a son, then the probability of both having sons is $(0.5 \times 0.5) = 0.25$ etc. A tree diagram showing the likelihood of the three couples having various combinations of sons and daughters is shown below:

1 st Couple	2 nd Couple	3 rd Couple	Outcomes	Pr
Son 50%	Son 25%	Son 12.5%	3 Sons 0 Daughter	12.5%
		Daughter 12.5%	2 Sons 1 Daughter	12.5%
	Daughter 25%	Son 12.5%	2 Sons 1 Daughter	12.5%
		Daughter 12.5%	1 Son 2 Daughters	12.5%
Daughter 50%	Son 25%	Son 12.5%	2 Sons 1 Daughter	12.5%
		Daughter 12.5%	1 Son 2 Daughter	12.5%
	Daughter 25%	Son 12.5%	1 Son	12.5%
		Daughter 12.5%	0 Son 3 Daughters	12.5%

Probabilities of sons and daughters for three couples

3.2 The Binomial Expansion

Look at the tree diagram again. What do you think is the major disadvantage? Using this type of method to determine probabilities can become terribly unwieldy when we consider more than a very few cases. We will therefore look at mathematical models or procedures available that can enable us to calculate probabilities without constructing a diagram. One of these procedures is the use of a formula to determine the frequency of various combinations.

This is given by $C = \frac{n!}{r!(n-r)!}$

Where C = the number of times an outcome will occur

n = number of trials

r = the number of 'desired' outcomes in a series of trials.

Example if we require the probability of 5 sons, $r = 5 ! =$ factorial. It means multiply the number by the next smaller whole number, then multiply the product by the next smaller by 1.

e.g. $(4!) = 4.3.2.1 = 4 \times 3 \times 2 \times 1 = 24$.

Example 8

What are the various combinations of given birth to sons and daughters in ten births or trials?

Solution:

Note that if we have to use the tree diagram we would need to have 2^{10} or 1024 branches of the tree using the formula $C = \frac{n!}{r!(n-r)!}$ we have that C = 10 times, 9 times, 8 times etc.
n = 10, r = 10!

$$\begin{array}{l} 10\text{Sons} \\ 0\text{Daughters} \end{array} = \frac{10!}{10!(10-10)!} = \frac{10!}{10!} = \frac{10!}{10!} = 1 \text{ combinations}$$

$$\begin{array}{l} 9\text{Sons} \\ 1\text{Daughter} \end{array} = \frac{10!}{9!(10-9)!} = \frac{10!}{9!.1!} = \frac{10.9!}{9!.1!} = \frac{10}{1} = 10 \text{ Combinations}$$

$$\begin{array}{l} 8\text{Sons} \\ 2\text{Daughters} \end{array} = \frac{10!}{8!(10-8)!} = \frac{10!}{8!.2!} = \frac{10.9.8!}{8!.2.1} = \frac{10 \times 9}{2} = 45 \text{ Combinations}$$

$$\begin{array}{l} 7\text{Sons} \\ 3\text{Daughters} \end{array} = \frac{10!}{7!(10-7)!} = \frac{10!}{7!.3!} = \frac{10.9.8.7!}{7!.3.2.1} = \frac{720}{6} = 120 \text{ combinations}$$

$$\begin{array}{l} 6\text{Sons} \\ 4\text{Daughters} \end{array} = \frac{10!}{6!(10-6)!} = \frac{10!}{6!.4!} = \frac{10.9.8.7.6!}{6!.4.3.2.1} = \frac{5040}{24} = 210 \text{ Combinations}$$

$$\begin{array}{l} 5\text{Sons} \\ 5\text{Daughters} \end{array} = \frac{10!}{5!(10-5)!} = \frac{10!}{5!5!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5!}{5! \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{30240}{120} = 252 \text{ Combinations}$$

$$\begin{array}{l} 4\text{Sons} \\ 6\text{Daughters} \end{array} = \frac{10!}{4!(10-4)!} = \frac{10!}{4!6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4!}{4! \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{151200}{720} = 210 \quad "$$

$$\begin{array}{l} 3\text{Sons} \\ 7\text{Daughters} \end{array} = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3!}{3! \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 120 \text{ Combinations}$$

$$\begin{array}{l} 2\text{Sons} \\ 8\text{Daughters} \end{array} = \frac{10!}{2!(10-2)!} = \frac{10!}{2!8!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 2!}{2! \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 45 \text{ Combinations}$$

$$\begin{array}{l} 1\text{Son} \\ 9\text{Daughters} \end{array} = \frac{10!}{1!(10-1)!} = \frac{10!}{1!9!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 2 \cdot 1!}{1! \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 2 \cdot 1} = 10 \text{ Combinations}$$

$$\begin{array}{l} 0\text{Son} \\ 10\text{Daughters} \end{array} = \frac{10!}{0!10!} = \frac{10!}{10!} = 1 \text{ Combination}$$

If you want to get the probabilities, remember that probability is given by:

$$\frac{\text{No. of simple events comprising event A}}{\text{Total No. of simple events in the sample space}}$$

In this example, total number of events in the sample space = $2^{10} = 1024$

$$\square \text{ten sons} = \frac{1}{1024}, \text{ nine sons} = \frac{10}{1024}, \text{ 8 sons} = \frac{45}{1024}, \text{ 7 sons} = \frac{120}{1024} \text{ etc.}$$

When the probabilities are added together, it will sum up to 1. Again, if we add up all the combinations, it will sum up to 1024.

$$\text{So the probability} = \frac{1024}{1024} = 1.$$

Note also that you can apply your knowledge of basic algebra in this type of binomial expansion. For instance, you have known that $(a+b)^2 = a^2 + 2ab + b^2$.

This can be applied in the case of throwing two coins the equation will be given by $(\frac{1}{2} + \frac{1}{2})^2 = (\frac{1}{2})^2 + 2(\frac{1}{2} + \frac{1}{2}) + (\frac{1}{2})^2 = \frac{1}{4} + \frac{1}{2} + \frac{1}{4}$.

$$\text{In the case of } (\frac{1}{2} + \frac{1}{2})^3 = (\frac{1}{2})^3 + 3(\frac{1}{2})^2(\frac{1}{2}) + 3(\frac{1}{2})(\frac{1}{2})^2 + (\frac{1}{2})^3$$

$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8}.$$

Here the numerators expressed the frequencies with which each kind of outcome occurs, while the denominators express the total number of occurrence and each ratio expresses a probability. The distribution is always symmetrical.

4.0 CONCLUSION

In this unit, we looked at the decision diagrams and the binomial expansion. You have noted that using the decision diagrams may not be the best in some circumstances and therefore a mathematical mode becomes imperative. This is why as you have seen the binomial equation or the binomial expansion is used. It implies that you can calculate probabilities without using diagrams. This unit is just to show you that it is possible to do that. You are not required to dwell much of your time in it.

5.0 SUMMARY

You have learnt that the decision diagram is one of the various ways of using deductive logic to get to probabilities based on assumptions made earlier. It uses a tree diagram having nodes and branches. You were told that the branches are called 'Chance Fork.' You can also use a mathematical model to find probabilities.

One such models is given by $C = \frac{n!}{r!(n-r)!}$. You also learnt that you can apply your knowledge of basic algebra in the binomial expansion, using as an example $(a+b)^2 = a^2 + 2ab + b^2$.

6.0 TUTOR-ARKED ASSIGNMENT

- i. What is a decision diagram?
- ii. Explain the formula $C = \frac{n!}{r!(n-r)!}$

7.0 REFERENCES/FURTHER READING

Ary, D and Jacobs, L.C. (1976) Introduction to Statistics purposes and Procedures. New York, Chicago...Toronto, Sydney. Holt Rinehart and Winston.

Guilford, J.P. and Fruchter, B Fundamental Statistics in Psychology (1981) and Education. Auckland, Bogota...Sydney, Tokyo. McGraw Hill Int. Book Company.

MODULE 3 INTRODUCTION TO INFERENTIAL STATISTICS

Unit 1 Randomization and Sampling Techniques

Unit 2 Sampling Errors

Unit 3 Hypothesis Testing

UNIT 1 RANDOMIZATION AND SAMPLING TECHNIQUES

Table of Contents

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Population and Sample
 - 3.2 The concept of Randomization
 - 3.2.1 Features of Randomization
 - 3.3 Sampling Techniques
 - 3.3.1 Random Sampling
 - 3.3.2 Systematic Sampling
 - 3.3.3 Stratified Random Sampling
 - 3.3.4 Clusters Sampling
 - 3.3.5 Purposive Samples
 - 3.3.6 Incidental Samples
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

You have just finished reading through the units on probabilities and you can remember that probability has been defined as the Mathematics of chance. The Laws of probability will apply, as you will see presently, in randomization, which is a prominent factor of inferential statistics. This is because the basic rule of randomization is that chance and chance alone determines which members of the population are included in the sample,

since the purpose of most educational research is to investigate the characteristics of population. In this unit therefore, we shall look at population and samples, the concept of randomization, sampling techniques.

2.0 OBJECTIVES

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

- i. Define population of a study
- ii. Distinguish between population and sample
- iii. Explain the concept of randomization
- iv. Describe the different sampling techniques

3.0 MAIN CONTENT

3.1 Population and Sample

The concepts of population and sample are of fundamental importance to research. A central and sustainable premise upon which scientific research is based on is the investigation of a problem using small representational and proportionate groups that is observed and from whom findings are made, and to discover new knowledge that can be generalized to a defined group. Generalizability of research findings depends on the extent to which the population of the study is defined as well as on the adequacy of the sampling procedure adopted in the study.

Talking about the population, you will note that the population in a statistical investigation is always arbitrarily defined by naming its unique properties. Some statisticians call it universe because their idea of population is quite different from the population idea. We can define population of a study to include all sets of individuals, objects, events or reactions that can be described as having a unique combination of qualities. It could be a group of human individuals, families, species, or orders of animals or of plants. In other words a population is all the cases (persons, objects, events) that constitute an identified group.

Examples are:- all SSS students offering Chemistry in the SSCE in the Imo State school system; all students of the National Open University of Nigeria; all registered voters in the last elections in Nigeria; all Junior Secondary School Students in Ihitte/Uboma L.G.A. of Imo State; all Science teachers in the Nigerian school system; all smokers in Lagos State;

all candidates who passed the National Common Entrance Examination to the Unity Schools in 2004; etc.

Activity 1

Give 40 examples of population that can be used for statistical investigations.

The parts or portions or sub-sets of the population studied on the bases of which conclusions are drawn on entire population are called samples. Samples are meant to represent population and are of great importance especially in survey research. We can now say that the small group which a researcher purposefully pulls out for an investigation is known as sample while the totality of the group from which the sample is drawn is called the population. The process of getting this representative proportion out from the population is called sampling.

3.2 The concept of Randomization

You have learnt that the sample is a sub-set of the population of study. The question that will border you now is how to draw the samples from the target population. This leads to the concept of randomization which implies the effort made by a researcher to ensure that every member of the population has an equal and independent chance of being selected in drawing the samples. It means that no special group or group of the population is particularly favoured in the sample selection. In other words, each individual in the population has equal chances of being selected. It also implies that each member of the target population may be selected on his or her own merit, not as determined by another member. Therefore, any sampling method or procedure which fails to give every member of the population the chance of being selected by merit is described as a biased sample. This is not acceptable as a basis for decision making or for generalization and drawing conclusion in research.

3.2.1 Features of Randomization

Now that you know that randomization is a very important key to representative sampling, what would you say are the important features? As the best and most reliable way to compose a sample whose attributes are guaranteed to exist in the population from which it was drawn, randomization has the following features:-

- (a) It ensures that every member of the population has an equal and independent chance of being drawn. This is in compliance with the first law of probability which talks about events having fair and equal chance of occurrence rather than being contrived and shrouded in fraud and deceit.
- (b) Drawing one event, object, subject or individual is independent of drawing another event, object, subject or individual because each of these is mutually an exclusive or independent unit or entity.
- (c) It satisfies all the requirements of making inferences about the population.
- (d) It makes the findings of a study more reliable and valid. This is because randomization eliminates all forms of bias from the sampling procedure and makes it certain that the findings reflect the true situation in the study.
- (e) It is very essential if the findings of a study must be generalizable to the entire population. If we use a sample for a study and the results permit us to draw inference which applies to an entire population then such inference is authentic to the extent that randomization was used in drawing the sample from the population.

3.3 Sampling Techniques

You have learnt that inferential statistics involve generalizing or making inferences from sample statistics to population parameters. You have also noted that randomization is the basis for inferential statistical reasoning. Let us therefore look at the different methods of composing samples in research.

3.3.1 Random Sampling

This is the best technique of sampling which refers to the selection procedure whereby all the cases in the defined population have an equal probability or chance of being selected and the selection of each case from the pool of cases is independent of the selection of another case. If a study involves a finite population that is relatively small, readily accessible and homogeneous, a simple random sample is feasible. In other words, for you to meet the criterion of a simple random sample, each member of the population must have the same probability of inclusion in the sample. The two major methods is otherwise and simply called balloting, and the other

one which is more systematic, refined and scientific method is called table of Random Numbers. You will get the details of these in your EDU 703 titled 'Research Methods'.

3.3.2 Systematic Sampling

This is used to obtain a random sample from a defined population by taking every K^{th} cases from a list of the population. That is to say, you have to number all the subjects in the population, and shuffling the numbers in a bag to achieve randomization. Let us take that the sample size = n , and the population size $N =$, then the sampling interval K^{th} will be given by $K^{\text{th}} = N/n$. For instance, if $N = 1000$, $n = 100$ then $K = 10$. we can randomly pick any number from 1 to 10. In this case, the selection of any number determines the entire sample. Example: if we pick 2, then 2, 12, 22, 32, 42 etc automatically become members of the sample.

What do you think will be the advantages and the disadvantages of this method of sampling?

You would have noticed that the main advantage here is that it requires less work. The disadvantage can be from the fact that if the listing of the population is not randomly done, periodicity can be introduced. Periodicity means a situation where every K^{th} member of the population has some characteristics peculiar or unique to only those members. For instance, in a study involving a population of policemen, in which every K^{th} member is a traffic police, this type of sample will definitely have effect on the dependent variable because the sample is already biased. Therefore, if periodicity cannot be avoided, it is better to use another sampling method.

3.3.3 Stratified Random Sampling

If a type of research is conducted with the population composed of sub-groups which tend to make the population heterogeneous, stratified random sampling is the most appropriate. You will note that most studies in education and social sciences have their populations stratified by nature. This is because a whole range of differences exist even within a particular characteristic. With regards to height, for instance, you will get giants, tall people, average height people, short people, dwarfs and midgets. Also, research subjects in schools can comprise male and female, old and young people, experienced and inexperienced administrators, occupational differences, income differences, urban and rural schools, socio-economic status, senior and junior students, academic and non-academic staff, graduate and non-graduate staff etc.

Random selection can be carried out within each sub-group. Then, the randomly selected representatives of the sub-groups together form the stratified sample. The random selection can be done in proportion, according to the size or number in the population of each sub-group. This is called proportional allocation. This requires information about the relative sizes of the strata in the population. That is to say that the exact population numbers or good estimates of these numbers should be available.

There is another method of allocation called Optimum Allocation. In addition to the size information, optimum allocation requires good estimates or exact values of the variances of the strata or sub-groups in the population.

3.3.4 Cluster Sampling

This is a method of sampling involving a naturally occurring group of individuals rather than an individual. In other words, a cluster sample is one in which the research interest characteristics have been identified, the areas in which these characteristics exist have also been identified and zoned reflecting these characteristics and samples from each of the identified zones randomly constituted. The justifications include cost reduction, time saving among others. It is most appropriate for sociological studies and education researches. It is also used where there is an attempt to study characteristics in their natural settings or to ensure geographic representation of intact groups whose distinct characteristics are of interest in a research.

Activity 2

Give 2 examples each of population suitable for

- i. Cluster sampling
- ii. Stratified random sampling
- iii. Systematic random sampling
- iv. Simple random sampling

3.3.5 Purposive Samples

These are samples that are arbitrarily selected maybe because there is good evidence that they are good representatives of the target population. For instance, if in national issues, it has been observed that a particular zone or

state has shown consistently in public opinion policy close reflections of national opinion time after time, any researcher willing to depend upon this experience may decide to use the limited population as source of sample to employ as a 'barometer' for the total population. But there are some disadvantages. These include that the conditions may have changed and the opinions may no longer be identical to that of the nation, or that prior information used to be obtained was not obtained this time or some new issues have introduced new changes etc.

3.3.6 Incidental Samples

When samples are used because they are the most readily available, they are called incidental samples. Results yielded by such samples can be generalized beyond the samples with some risks.

Other samples like volunteer and non-probability sample, convenience samples, judgement samples, quota samples, etc. you are going to read them in your Research Methods in Education EDU 703.

4.0 CONCLUSION

In this unit, you have learnt that in conducting experimental studies in education, your samples must be randomly composed. This is because when non-randomly selected and biased samples are used, the laws of probability no longer hold. Again results derived from using non-random samples in experimental purposes or even other research designs can be misleading because of sampling errors and non-representativeness of the population. Therefore, for educational studies, you must insist on large samples that are representative of the target population and/or samples that are randomly composed.

5.0 SUMMARY

Going through this unit, you have noted that inferential statistics are used to make reasonable decisions on the basis of incomplete data. You have learnt that samples are used to make inferences about a population when the samples are randomly drawn or when subjects from available population are randomly assigned to treatments. You can now define a given target population using its characteristics and the different types of sampling methods such as random sampling, systematic, stratified and cluster samplings. These are called probability samples which are biased and cannot be used for generalization.

6.0 TUTOR-MARKED ASSIGNMENT

1. What is random sample?
2. Match the following definitions with the terms given below:
 - a. all cases that constitute an identified group
 - b. a sample characteristic
 - c. a population characteristic
 - d. using a chance procedure to select a sample from population
 - e. using a chance procedure to assign subjects to groups
 - f. a sub-group of a population – statistic, population, random sampling, random assignment, sample, parameter.
3. List 4 types of probability sampling and 2 types of non-probability sampling

7.0 REFERENCES/FURTHER READINGS

- Ali, A. (1996) Fundamentals of Research in Education, Awka. Mekis Publications (Nig)
- Ary, D. and Jacobs, L.C. (1976) Introduction to Statistics Purposes and Procedures: New York, Chicago, San Francisco, Atlanta, Dallas, Toronto, London, Sydney, Holt, Rinehart and Winston.
- Denga, I.D. and Ali, A An Introduction to Research Methods and (1983) Statistics in Education and Social Sciences. Jos. Savannah Publishers.
- Guilford, J.P. and Fundamental Statistics in Psychology and Fruchter, B ((1981) Education (ISE) Auckland, Bogota, Guatemala, Hamburg, Johannesburg, Lisbon London... McGraw-Hill International BookCompany.
- Olaitan, S.O. and Practical Research Methods in Education. Nwoke, G.I. (1988) Onitsha, Summer Educational Publishers Limited.
- Zehna, P.W. (1974) Introductory Statistics. Boston, Massachusetts. Pondle, Weber and Schmidt, Inc.

UNIT 2 SAMPLING ERRORS

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Sampling Error
3.2	The Lawful Nature of Sampling Errors
3.2.1	The Expected Mean of Sampling Error is Zero
3.2.2	Sampling Error is an Inverse Function of Sample Size
3.2.3	Sampling Errors is a Direct Function of the Standard Deviation
3.2.4	Sampling Errors are Normally Distributed
3.3	Standard Error of the Mean
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

In the last unit, you worked through population and samples. You learnt the different sampling techniques and the advantages of randomization. But even with randomization, sampling errors will still occur. This is because no sample is purely identical to the population. This is why absolute generalization is not advisable. In this unit, you will learn the sampling error and the lawful nature of sampling error.

2.0 OBJECTIVES

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

- i. Define and explain the concept of sampling error.
- ii. Explain the lawful nature of sampling errors
- iii. Explain the standard error of the mean.

3.0 MAIN CONTENT

3.1 Sampling Error

You have noted that for a research to be meaningful, the samples must be randomly selected. But you have to note also that when an inference is made from a sample to a population a certain amount of error is involved. This is because no sample is absolutely identical in all respects with the population from which it was drawn. Even random samples from the same population are expected to vary from one to another. For instance, in a population of primary six pupils in Ihitte/Uboma Local Government Area of Imo State, the mean intelligence score of one random sample may be different from the mean intelligence score of another random sample from the same population. Such differences are called sampling errors and they result from the fact that you have observed only a sample and not the entire population.

Sampling error is therefore defined as the difference between a population parameter and a sample statistic. It explains why the results of a study based on data obtained from a sample investigated may not be absolutely (100%) generalizable to the study's population. For example, if we know the mean of the entire population, symbolized by μ , and the mean of a random sample symbolised by \bar{X} , from that population, then the difference between the two is the sampling error, symbolized by e . Thus $e = \bar{X} - \mu$.

Example 1

If the mean intelligence score of 20,000 primary six pupils from Ihitte/Uboma L.G.A. of Imo State is 120, and a particular random sample of 500 has a mean of 118. What is the sampling error?

Solution:

The sampling error e given by $e = \bar{X} - \mu$, where $\bar{X} = 118$ and $\mu = 120$
 $\square e = 118 - 120 = -2$

You would have noted that we usually depend on sample statistics to estimate population parameters. This is because; the notion of how samples are expected to vary from population is a basic element of inferential statistics. In other words, since in most cases the population is too large, we are not expected to find out their means or even knowing them. However, Inferential Statistics allows us to use sample data to generalize on

the population despite the presence of sampling error because it accounts for such errors in its mode of treating data obtained from samples. It also allows us to estimate the variability that would be expected under such a circumstance of sampling error.

3.2 The lawful nature of Sampling Errors

Sampling errors have systematic laws which guide their occurrence, magnitude and effect on research samples. They behave in a lawful and predictable manner. The laws concerning sampling error have been derived through deductive logic and have been confirmed through experience extensively and carefully, it has been found that they follow regular laws. These are:

3.2.1 The Expected Mean of Sampling Errors is Zero

This means that if we have an infinite number of random samples drawn from a target population, the positive sampling errors would be expected to balance out with the negative sampling errors so that the mean of the sampling errors for all the samples will be zero. For example, if we take our target population to be all the SS3 students in Obowo L.G.A. in the 2004/2005 session; if their mean height is 1.65, you would notice that some would be more than 1.65 and some would be less. Therefore, if several random samples are drawn from the population, in the long run the positive and the negative sampling errors will balance. This is because positive errors will equal negative errors. It implies then that a single sample mean is as likely to underestimate a population mean as to overestimate it. We can justifiably say that a sample mean is an unbiased and reasonable estimate of the population mean.

3.2.2 Sampling Error is an Inverse Function of Sample Size.

This means that as the size of a sample increases, the expected sampling error decreases. In other words, small samples are more prone to sampling errors than large ones. As the sample size increases there is the likelihood that the mean of the sample comes very close or near the population mean.

3.2.3 Sampling Error is a Direct Function of the Standard Deviation of the Population

If you think of the heights of the members in a given population, you will note that there will be more spread or variation than when you consider

within the height, tall, giants, short or dwarfs. In other words, if you take ‘tall’ as an attribute of the population, you will note that this is a close attribute and the sampling error will be smaller. Thus the more spread or variation we have among members of a population, the more spread or variation we expect in sample mean.

3.2.4 Sampling Errors are Distributed in a Normal or Near Normal Manner Around the Expected Mean of Zero

Any time you move farther and farther away from the population mean, you will find fewer and fewer sample means occurring, you will also note that the means of random samples are distributed in a normal or near normal manner around the population mean. In other words, sample means near the population mean will occur more frequently the same way the distribution of sampling errors is normal or near normal in shape because sampling error is the difference between sample mean and population mean.

3.3 Standard Error of the Mean

This is a standard deviation of the distribution of sample means. It is used when we need an estimate of the magnitude of the sampling error associated with the sample mean when it is used as an estimate of the population mean. Note that the extent and the distribution of sampling errors can be predicted. This is done as an estimate of the magnitude of the sampling error associated with the sample mean when it is used as an estimate of the population mean. You have noted that sampling error manifests itself in the variability of sample means. Therefore, if you calculate the standard deviation of a collection of means from random samples from a single population, you would have got an estimate of the amount of sampling error. We can obtain this estimate on the basis of one sample, but we need two items of information. These are the population parameter σ and the sample size N . This expected standard deviation of sampling error of the mean otherwise called standard error of the mean is represented by the symbol $\sigma_{\bar{x}}$. This has been proofed through deductive logic to be equal to the standard deviation of the population (σ), divided by the square root of the number in each sample (\sqrt{N}).

Thus we have $\sigma_{\bar{x}} = \sigma / \sqrt{N}$: where $\sigma_{\bar{x}}$ = Standard error of the mean

σ = Standard deviation of the population

N = number in each sample

Activity 1

1. What is sampling error?
2. List laws guiding sampling errors.

Answers to Activity 1

You may have given these answers

- i. Sampling error is the difference between a population parameter and a sampling statistic
- ii. The laws guiding sampling errors are.
 - a. The expected mean of sampling errors is zero
 - b. Sampling error is an inverse function of sample size
 - c. Sampling error is a direct function of the standard deviation of the population.
 - d. Sampling errors are distributed in a normal or near normal manner around the expected mean of zero.

4.0 CONCLUSION

You have seen that even with the rigours of randomization and sampling, generalization cannot be absolute to the population. But this is not to deter you from having a randomized sample in your research work. Since you can now estimate the magnitude of the sampling errors.

5.0 SUMMARY

In this unit, you have worked through the sampling error, defining it as the difference between a population parameter and a sample statistic. We have also dealt with the lawful nature of sampling errors stating the laws as:

- i. The expected mean of sampling error is zero
- ii. Sampling errors is an inverse function of sampling size
- iii. Sampling error is a direct function of the standard deviation of the population.

- iv. Sampling errors are distributed in a normal or near normal manner around the expected mean of zero.

In this unit also, you learnt about the standard error of the mean and how to predict the extent and distribution of the sampling error. You know that if you have the population parameter and the sample size, you can estimate the sample error. In this case we can apply the formula $\sigma_{\bar{x}} = \sigma / \sqrt{N}$.

6.0 TUTOR MARKED ASSIGNMENT

- i. What is the relationship between sample size and sampling error associated with sample mean?
- ii. What is the relationship between the standard deviation of a population and sampling error associated with sample means?
- iii. What is sampling error?
- iv. What is the formula for estimating standard of the mean?

7.0 REFERENCES/FURTHER READINGS

Ali, A (1996) Fundamentals of Research in Education, Awka. Mekis Publications (Nig)

Ary, D. and Jacobs, L.C. (1976) Introduction to Statistics Purposes and Procedures: New York, Chicago, San Francisco, Atlanta, Dallas, Toronto, London, Sydney, Holt, Rinehart and Winston.

Denga, I.D. and Ali, A.(1983)An Introduction to Research Methods and Statistics in Education and Social Sciences. Jos. Savannah Publishers.

Guilford, J.P. and Fundamental Statistics in Psychology and Fruchter, B ((1981) Education (ISE) Auckland, Bogotá, Guatemala, Hamburg, Johannesburg, Lisbon London... McGraw-Hill International BookCompany.

Olaitan, S.O. and Practical Research Methods in Education. Nwoke, G.I. (1988)Onitsha, Summer Educational Publishers Limited.

Zehna, P.W. (1974) Introductory Statistics. Boston, Massachusetts. Prindle, Weber and Schmidt, Inc.

UNIT 3 HYPOTHESIS TESTING

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Inferential Statistics
3.2	Testing the Hypothesis Statistically
3.2.1	Null Hypothesis H_0
3.2.2	Alternative Statistics H_1
3.3	Testing for Significance Confidence Limits
3.3.1	0.05 or 5% Level of Significance
3.3.2	0.01 or 1% Level of Significance
3.4	The Two Types of Errors
3.4.1	Type I Error
3.4.2	Type II Error
3.5	One Tailed and Two tailed Test
3.6	Concept of Degree of Freedom
3.7	Types of Inferential Statistics and how to use them
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

Most research studies involve investigating the attributes or characteristics of research samples because it is not possible, neither are it desirable to study exhaustively the attributes of all the members of the population. This is what you have learnt in the previous unit. We can say therefore that data obtained from the sample of the population can be used to make inferences and/or estimate the parameters. The statistical techniques for treatment of those data obtained are called inferential statistics. In this unit therefore, you will learn how to explain inferential statistics, testing for significance testing, the hypothesis, the two types of error, one and two tailed test, the concept of degree of freedom and types of inferential statistics and how to use them.

2.0 OBJECTIVES

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

- i. explain inferential statistics
- ii. explain the confidence limits 0.05 and 0.01
- iii. differentiate a null hypothesis from alternative hypothesis
- iv. explain the two types of errors – types I and II error
- v. differentiate between one tailed and two tailed test.
- vi. Explain the concept of degree of freedom
- vii. List and say how to use different types of inferential statistics.

3.0 MAIN CONTENT

3.1 Inferential Statistics

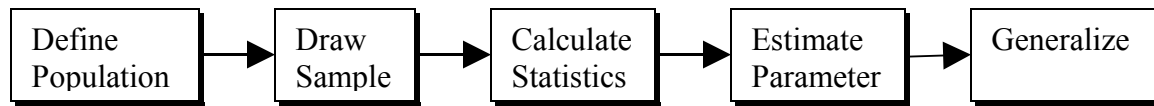
You have known that it is not possible for population parameters or characteristics to be studied exhaustively. We therefore use data collected from samples to generate the population from which the samples were drawn. This, we do using induction or inference, while the statistical test we use to make valid inferences are called inferential or parametric statistics. We can now say that statistical inference is a means of vigorously estimating parameters or population characteristics from the sample statistics, based on the laws of probability. In other words, inferential statistics are used to make reasonable decisions on the basis of incomplete data. Parametric statistics form an aspect of inferential statistics and are also the most powerful and most sensitive. Parametric data provide more reliable evidence. You have seen that errors can be introduced in studies through sampling. These errors can mitigate the strength of inferencing in the so called internal and external validity threats. But if the inferential statistics are used in line with and complying with certain underlying basic assumptions, the mitigation threats will be minimized. There are some basic assumptions underlying the use of parametric or inferential statistics about population parameters. These according to Denga and Ali (1983) are:

- i. Scores in a given population are normally distributed about the mean.
- ii. There is the equality of population variances of the comparison groups in the study.
- iii. The scores being used are derived from interval or continuous data.

You will recall that we have said that inferential statistics is defined as the science of making reasonable decisions with limited information. You will also recall that inferential statistics are means of establishing cause-effect relationships between variables and stated as probability statements. The steps the statistical inference processes take are:

- i. Define the population.
- ii. Draw the samples randomly or through randomization.
- iii. Calculate the statistics from random samples.
- iv. Estimate the parameters and
- v. Generalize to the target population.

This can be stated diagrammatically as follows:



Activity 1

1. What is inferential statistics? What is the major difference between inferential and descriptive statistics? What are the major assumptions underlying the use of inferential or parametric statistics? (Check the answers at the last part of the unit).

3.2 Testing for Significance – Confidence Limits

Now that you can explain the two types of hypotheses, let us move a step further in testing the hypothesis. You can test a hypothesis in research using parametric or inferential statistics, usually at some pre-determined level of significance or alpha level. You will note that based on this confidence limit, the means of two or more comparative groups are tested to enable you or the researcher to decide whether the significance of the differences observed for the two or more means is large enough for him to reject the null hypothesis or the difference is too small for him to accept the null hypothesis. There are two main cut-off points or confidence limits in testing research hypothesis. These are:

3.3.1 0.05 or 5% Level of Significance

This asserts that the outcome or result of an event observed will be due to chance 5% of the time or less. It means that if a significant inferential value is found as a result of comparing the significance of the difference between means, we will reject the null hypothesis at the 5% level of

significance. In other words, if we repeat a study using the same treatment conditions for 100 times, we are sure 5% of the time or 5% of our result will be due to chance; while 95% is not due to chance. ($P < 0.05$).

3.3.2 0.01 or 1% Level of Significance

This asserts that the outcome or result of an event observed will be due to chance 1% of the time. This is a more rigorous and stranger confident alpha level. At this point or level, the indication is that 1% or 1 out of 100 replications of our study treatments will be due to error or chance while 99% or in 99 out of 100 replications are the observed differences in outcome due to the effects of the experimental treatment rather than to chance or error.

You will have to note that there is no formula for determining the alpha. It is an arbitrary probability point which divides those probabilities that will lead to retention of the null hypothesis from those probabilities that will lead to its rejection. You will also note that whenever error in sampling or chance is not responsible for the observed differences in the outcomes of an experiment or study, the null hypothesis is rejected. Whenever error or chance is responsible for the observed differences in the outcomes of an experiment, the null hypothesis is accepted. Therefore, you will choose an alpha by weighing the relative seriousness of Type I and Type II errors.

3.4 The two types of Error

3.4.1 Type I Error:

This occurs when the researcher rejects the null hypothesis or declares it false when it is actually true and should be returned. The probability of making type I error or rejecting the null hypothesis is very small = α (alpha). If you set the significance level at 5% or 0.05, it means that the mistake of rejecting the null hypothesis is approximately 5% of the time. In order to avoid type I error, we set the confidence limit as low as possible. For instance, at a confidence level of 0.001, we would likely make a type I error only about once in every thousand. This is negligible and amounts to almost no error at all. Caution must be taken here because the lower the alpha level, the greater the risk or possibility of making another type of error called type II error.

3.4.2 Type II Error

This occurs when a researcher accepts a null hypothesis when in fact it is false and therefore should be rejected. The probability of making a type II error is called β . You have to note that the chances of making a type II error is far greater than the chances of making type I error. You have to note also that alpha (α) and beta (β) are inversely related, that is, as one decreases, the other increases. Alpha is under our direct control while beta is only indirectly under our direct control while beta is only indirectly under our control through its inverse relation to alpha. The probabilities of making the two errors can be shown diagrammatically as follows:

	Reject H_0	Accept H_0
H_0 True	Type I Error $P = \alpha$	Correct
H_0 False	Correct	Type II Error $P = \beta$

Activity 3

What do you understand by $P < 0.05$?

What is the difference between type I error and type II error? (Check the answers at the last part of the unit)

3.5 One-tailed and Two-tailed Tests

Consider these hypotheses:

- there is no significant difference in the results of students who do physics and those who do not, in the mathematics test.
- students who do physics will perform better in the mathematics test than those who do not do physics.

If you consider the two hypotheses, you will note that the first one has no direction while the second one has a direction. If a hypothesis is stated in such a way that two or more groups, procedures or events are compared with the comparison giving indication of direction of difference, such a hypothesis is said to be one-tailed test. In other words, in a one-tailed test of significance, the hypothesis claims that a superiority or difference exist and also indicates the direction of the superiority or difference. For instance,

- (a) Sportsmen and women who are muscularly built do better in short put and discuss than those who are not.
- (b) Left handed people are more intelligent than right handed people
- (c) Chalk and talk is a better method of teaching than the lecture method.

Now if you examine these hypotheses you will see the direction of difference immediately.

On the other hand, if a hypothesis is stated to say that there is no difference exists but does not show the direction of this difference, that hypothesis is a two-tailed hypothesis, and should be treated with two tailed test. For instance:

- (a) there is no significant difference in performance at the undergraduate level between students who attended Unity Schools and those who did not.
- (b) there is a significant difference in the performance of boys and girls in the chemistry practical test
- (c) women have equal IQ scores like men.

These hypotheses are two-tailed because the directions of superiority are not indicated. Both null hypothesis and the alternative hypothesis can be one tailed or two-tailed.

Activity 4

- i. Construct 5 null hypotheses that are one-tailed
- ii. Construct 5 alternative hypotheses that are one-tailed
- iii. Construct 5 alternative hypotheses that are two-tailed.

3.6 Concept of degree of freedom

After you have determined the type of hypothesis to test, the level of significance to test it and trying to avoid type I or type II error, and after computing the test of significance using any inferential statistics, you will want to take a decision whether to accept or reject your hypothesis. The decision can only come after you have got the degree of freedom from the appropriate critical values of the particular statistics you are using. This degree of freedom refers to the number of observations which are allowed to vary around a fixed or constant parameter. In other words, there are the numbers or quantities or values which are free to vary after placing certain restrictions on the data we are comparatively testing. For example, if you

are given the number 200 and you are asked to give five numbers including 22 that must add up to 200. You can give 50, 65, 35, 28 and 22 to make it 200. You can freely give any four numbers which must be added to 22 to make 200. But you are not free to name 22, and 22 must not vary otherwise it will not be 200. Therefore, in this case you have four degrees of freedom because you have lost one chance of naming a number out of five numbers. Degree of freedom here becomes $(N-1 \text{ or } 5-1 = 4)$. Note that each test of significance has its own method of finding its degree of freedom.

3.7 Types of Inferential Statistics and how to use them.

You have been able to go through the different aspects and components guiding the use of inferential statistics in testing hypothesis. This time we will look at the different types of inferential statistics and how and when to use them. After this, you will begin to use them. As you know, there are different types of distributions as well as different types of statistics available for treating data for inference. They include binomial, normal, poisson, gamma, beta, t, chi, F, etc. The type of test to be used depends on the type of research design and sometimes the type of data collected the number of samples per group and the number of groups. Let us use the following as examples:

S/N	CHARACTERISTICS OF SAMPLE / VARIABLE	APPROPRIATE TEST
1	Two groups comparison randomly composed and unpre-tested subjects = Independent means with number less than 20	t – test of independent samples
2	When there is pretest – post test and selection of subjects = non – independent means	t – test of non-independent samples
3	If N is more than 30 for a post test only and randomly sampled subjects	Critical ratio 2 – test of independent means
4	More than two groups comparison of independent mean, no pretest and one independent variable	ANOVA
5	If two or more independent variables are involved	Path analysis or meta analysis or MANOVA
6	In a pretest, post-test design	ANCOVA
7	If more than two groups are involved in a pretest, post-test, with two or more variables	MANCOVA

8	For significance of the difference between observed and expected frequencies	Chi – square
---	--	--------------

Activity 5

Give the full meaning of ANOVA, MANOVA, ANCOVA and MANCOVA

In what conditions do they apply? (Check the answer at the last part of the unit)

4.0 CONCLUSION

You have seen that statistics provide a means of treating data so that they are summoned and reduced to a point they are interpretable. When the data obtained from large samples that have been randomly composed are treated with appropriate statistical tests, generalizations to the target population are more accurate compared to when the data are obtained from small, selected or biased samples. You have also seen that for using the inferential statistics, certain conditions have to be met before data can be inferentially treated and interpreted. You should therefore keep to these conditions whenever you want to test your hypothesis using inferential statistics.

5.0 SUMMARY

In this unit, you have learnt that inferential statistics are the ones which enable us make valid inference and generalizations from sample data to the population. There are underlying assumptions in the use of inferential statistics. These are:

- i. the samples are randomly composed
- ii. the variables are normally distributed and
- iii. measurements are at interval or ratio levels.

You have seen that in research two types of hypotheses are tested. These are the null hypothesis and the alternative hypothesis. These hypotheses can be one-tailed or two-tailed and can be tested at 5% or 1% confidence limits or alpha level. In doing this, you will remember to be cautious to avoid type I error or type II error. In this unit also, you went through the concept of degree of freedom and the different types of inferential statistics and when to use them.

6.0 TUTOR-MARKED ASSIGNMENT

- i. What is inferential statistics
- ii. Differentiate between a null and alternative hypotheses
- iii. Explain the confidence limits 5% and 1%.
- iv. Define the two types of errors in research
- v. What is the major difference between one-tailed and two-tailed test?

7.0 REFERENCES/FURTHER READINGS

- Ali, A (1996) Fundamentals of Research in Education, Awka. Mekis Publications (Nig)
- Ary, D. and Jacobs, L.C. Introduction to Statistics Purposes and (1976) Procedures: New York, Chicago, San Francisco, Atlanta, Dallas, Toronto, London, Sydney, Holt, Rinehart and Winston.
- Denga, I.D. and Ali, A. (1983) An Introduction to Research Methods and Statistics in Education and Social Sciences. Jos. Savannah Publishers.
- Guilford, J.P. and Fundamental Statistics in Psychology and Fruchter, B ((1981) Education (ISE) Auckland, Bogotá, Guatemala, Hamburg, Johannesburg, Lisbon London...McGraw-Hill International BookCompany.
- Olaitan, S.O. and Practical Research Methods in Education. Nwoke, G.I. (1988) Onitsha, Summer Educational Publishers Limited.
- Zehna, P.W. (1974) Introductory Statistics. Boston, Massachusetts. Prindle, Weber and Schmidt, Inc.

MODULE 4 Z AND t-TESTS

Unit 1 Z –TESTS I

Unit 2 Z –TEST II

Unit 3 t –TESTS

UNIT 1 Z –TESTS I

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	The Z-Tests
3.2	Basic Assumptions underlying the Z-test
3.3	The Difference between Sample and Population Means
3.4	Difference between Means of Two Independent Samples
3.5	Difference between Means of Paired/Matched Samples
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

In the last unit, you learnt some fundamental concepts used in inferential statistics and hypothesis testing. In this unit, you will be introduced to the use of one of the inferential statistics which you may like to use in your research works. It is called the Z-test. Attempt will be made to use simple examples to illustrate this concept.

2.0 OBJECTIVES

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

- i. State the uses of the Z-tests
- ii. State the basic assumptions underlying the Z-test.
- iii. Solve some problems using the Z-test.

3.0 MAIN CONTENT

3.1 The Z-Test

The Z-test is an inferential statistical test. It is based on the characteristics and/or properties of the normal distribution and the normal curve. This is why the normal curve is sometimes called the Z-curve. You will remember that a standard normal curve has a bell-shape and symmetrical at the centre. The z-test can be used to ascertain whether some differences exists, and whether such difference is statistically significant at a given alpha-level, between:

- i. two independent sample means
- ii. two paired or matched sample means
- iii. sample proportion and population proportion
- iv. two independent sample proportions.

3.2 Basic assumptions underlying the Z –Test

These are the conditions which should be satisfied or assumed to be satisfied in using a z – test. They are:

- i. The population from which the samples are drawn should be normally distributed.
- ii. The sample size used in the test should be large enough (according to Ogomaka (2004) $n \geq 30$ is appropriate but $n \geq 60$ is better).
- iii. The variances of the population should not be significant i.e. the population should be homogeneous.
- iv. The samples are randomly or independently drawn from the population.

You will have to note that in carrying out your research work, if you use a large sample drawn from the target population through randomization, you are safe guarded by one of the central limit theories as regards the assumption of normal distribution. Now let us go to the examples.

3.3 The difference between Sample and Population Means

If we take \bar{X} as the sample means
 μ as the population mean
 n as the sample size or number
 δ as population standard deviation

$\sqrt{\quad}$ as square root

$$\text{then } Z = \frac{\bar{X} - \mu}{\delta\sqrt{n}} \quad \frac{(\bar{X} - \mu)}{\delta}$$

This is used to compare sample mean and population mean. For instance, it can be employed to test the hypothesis – “There is no significant difference between a given sample mean \bar{X} and a given population mean μ , at a specified alpha level.” But you know that most of the times population variance or standard deviation is not given or known. You have to estimate this using this

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \text{ or } \frac{\sum (x - \bar{x})^2}{n - 1}. Z \text{ will therefore be}$$

$$Z = \frac{\bar{X} - \mu}{\delta \sqrt{n}} \text{ or } \frac{\bar{X} - \mu}{\delta \sqrt{n - 1}}$$

You will also note that when n is large \sqrt{n} and $\sqrt{n - 1}$ are approximately equal ($\sqrt{n} \cong \sqrt{n - 1}$).

Example 1

An examination board developed, validated and standardized an aptitude test for Junior Secondary School (JSS) students in Lagos State. After administering this test for so many years in the state, the board claimed that the mean performance of Junior Secondary School (JSS) students in Lagos State is 65%. To this effect, the state government set up a study team to investigate and verify the claim. This team used randomization to select 1,500 JSS students from all the zones of the state and administered the same test on the samples. The result shows a mean of 68% and a standard deviation.

Solution

To test whether the performance of the sample is significantly different from that of the population, we have to use a hypothesis for the testing. Let us use a null hypothesis thus: This is no significant difference between the mean performance of the sample and that of the population at 0.05 alpha

level $z = \frac{\bar{X} - \mu}{\delta \sqrt{n}}$. But since δ is not given we estimate using $\hat{\delta}$. Therefore,

$$z = \frac{\bar{X} - \mu}{\hat{\delta} \sqrt{n}}. \bar{X} = 68\%, \mu = 65\%, \hat{\delta} = 28 \text{ and } n = 1,500$$

$$\square z = \frac{68 - 65}{28 / \sqrt{1500}} = \frac{68 - 65}{28} \times \sqrt{1500} = \frac{3}{28} \times \sqrt{1500} = 4.149625 = \underline{\underline{4.150}}$$

The value of z referred to as $z_{cal} = 4.150$. This is called the absolute value of z_{cal} which can be compared with the value on the table referred to as z_{tab} and called critical value of z_{tab} . If z_{cal} is greater than z_{tab} , the null hypothesis is rejected. If otherwise, the null hypothesis is accepted. Again, you will have to note that this null hypothesis is a two-tailed hypothesis. You remember the one-tailed and two-tailed test which you did in the last unit? So for a two-tailed test divide the alpha level by $2x$ and look at it in the z -table i.e. $z_{0.05/2} = z_{0.025}$. This is 1.960. So $z_{cal} = 4.150$, $z_{tab} = 1.960$.

Decision: Since z_{cal} is greater than z_{tab} , we reject that there is no significant difference and accept that there is a significant difference between the means. With the evidence in this test, we say that the claim of the examination board is false, or may be the sample used is not a representative of the population described by the board.

Example 2

An instructor give his class of 25 an examination which, based on years of use, has been shown to have mean = 80. His class obtains an $\bar{X} = 84$ and $\hat{\delta} = 10$ on the test. Find out, using a two-tailed test whether the difference between this class mean and the original is statistically significant at the 0.05 level.

Solution

Recall that $z = \frac{\bar{X} - \mu}{\hat{\delta} \sqrt{n}}$ Given: $\bar{X} = 84, \dots$

$$\square z = \frac{84 - 80}{10 / \sqrt{25}} = \frac{4}{10/5} = \frac{4}{2} = 2.000$$

$$z_{cal} = 2.000$$

Degree of Freedom (df) = $n-1 = 25-1 = 24$

ztab at 24d.f at 0.05 = 2.064

Decision

Since z_{cal} is less than z_{tab} , we accept that there is no significant difference between the class mean. Retain null hypothesis. Class mean of 84 is not significantly different from $\mu = 80$.

Activity 1

An investigation wishes to test the hypothesis that the mean of a certain population is 80. He used a random sample of ten drawn from the population. On testing the sample, the scores were 15, 45, 60, 75, 75, 90, 105, 105, 120 & 60. Using a null hypothesis and a two-tailed test verify that $\mu = 80$ is significant at 0.05 level.

3.4 Difference between Means of two independent Samples

Independent samples are those in which the selection of cases in one sample is not influenced by the selection of cases in the second sample. They are completely independent of each other. There is no logical pairing of subjects in the two groups or any reason to connect any given measure in one sample with measures in the other sample. There are basically two types of independent samples used in research. These are:

- i. When samples are randomly selected from two different populations where a researcher wishes to determine whether the two populations differ significantly on some criterion variable. For instance, drawing samples from male and female population to determine which group has a higher reasoning ability, using their means.
- ii. Two randomly sampled groups selected from the same population, but may be exposed to two different treatments or experimental conditions. The test to be employed is:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}}$$

$$\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}$$

$$\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}$$

where the denominator is the standard error of measurement of the sampling distribution of the difference between means of two independent samples of large sizes. This formula $z = \frac{X_1 - X_2}{\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}}$.

$$\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}$$

can be shown like this: $Z = \frac{\overline{X} - \overline{X}}{\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}}$

This can be made longer in this way:

$$Z = \frac{X_1 - X_2}{\sqrt{[\sum X_1^2 - ((\sum X_1)^2/n_1)] + [\sum X_2^2 - ((\sum X_2)^2/n_2/n_2]}}$$

Example 3

An agricultural science teacher in one of the Junior Secondary Schools in Kano State conducted a study to find out the attitude of his students towards practical agriculture. He grouped his students into those residences in the urban areas and those residences in the rural areas. He then developed an attitudinal scale which he administered on the students. He obtained the following results:

Residential Area	Mean	Standard deviation	Sample size
Rural	73.08	16	75
Urban	70.25	21	75

Verify whether the observed difference in the result is statistically significant at ($P < 0.05$).

Solution

First propose a null hypothesis thus: there is no significant difference between the means of the two groups of student at ($P < 0.05$). This implies

that the alternative hypothesis H_1 will be there is a significant difference between the means of the two groups of students.

$$\text{Given: } \bar{X}_1 = 73.08$$

$$\bar{X}_2 = 70.25$$

$$S_1 = 16$$

$$S_2 = 21$$

$$\text{Recall that the formula } z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\delta}{n_1} + \frac{\delta}{n_2}}}$$

NB n_1 can vary from n_2 .

$$\text{Substituting, we have } \frac{73.08 - 70.25}{\sqrt{1 \frac{16^2}{75} + \frac{21^2}{75}}}$$

$$\Rightarrow \frac{2.83}{\sqrt{\frac{256 + 441}{75}}} = \frac{2.83}{\sqrt{\frac{697}{75}}} = \frac{2.83}{\sqrt{9.293}}$$

$$= \frac{2.83}{3.0484969} = 0.9283263$$

$$\cong \underline{\underline{0.928}}$$

So, $z_{cal} = 0.928$; z_{tab} at $(0.05 \div 2 = 0.025) = 1.96$

It means that z_{tab} is greater than z_{cal} .

Decision:

Since $z_{cal} < z_{tab}$, we accept the null hypothesis H_0 . It means that the two means are not significantly different.

Example 4

In a research study, the researcher investigated the effect of two types of blood capsules on the weight of some children. He used two groups of children in a school and administered the capsules, one on one group and

the other on another group. After a period of eight weeks, he collected their weights for comparison on their weight increase. The results are:

Average increase for group I = 5.50 units, number 80

Average increase for group II = 8.30 units, number 75

Standard deviation for group I = 6 units for group II = 8 units

Verify whether the capsules are significantly different.

Given: $X_1 = 5.50$; $X_2 = 8.30$; $S_1 = 6$; $S_2 = 8$; $n_1 = 80$; $n_2 = 75$

$$\begin{aligned} \text{Solution:- } Z &= \Rightarrow \frac{5.50 - 8.30}{\sqrt{\frac{6^2}{80} + \frac{8^2}{75}}} \\ &= \frac{2.80}{\sqrt{\frac{36}{80} + \frac{64}{75}}} = \frac{2.80}{\sqrt{0.45 + 0.853}} = \frac{2.80}{\sqrt{1.303}} \\ &= \frac{2.80}{1.1416363} = 2.45262 = 2.453 \end{aligned}$$

If the H_0 reads: There is no significant difference between the effect of the two capsules on weight increase of the children. Then we note that $z_{cal} = 2.453$ and z_{tab} at $0.05/2 = 1.96$. It means that $z_{cal} > z_{tab}$.

Decision

We reject the null hypothesis H_0 and accept that there is a significant difference between the effect of the two capsules on the weight increase of the children.

Activity 2

A researcher wanted to investigate the effect of individualized instructions on learning outcomes. He used two groups of students for the experiment. Group A is the experimental group, while group B is the control group. He used individualized instruction on group A and group instruction on group B. At the end of the experiment, he came out with the following scores:-

Mean score for group A = 50

Mean score for group B = 40

Number of students in group A = 72, standard deviation = 11
 Number of students in group B = 80, standard deviation = 13

- i. State the null hypothesis H_0
- ii. State the alternative hypothesis H_1
- iii. Calculate the z-value
- iv. Draw the appropriate conclusions at 0.05 and 0.01 alpha levels.

3.5 Difference between means of paired/matched Samples

In this unit, you have noticed that we considered research designs in which the two samples observed are drawn independently from their respective populations. Sometimes, some studies may be interested in two samples that are not independent but are related to each other. In this case, a researcher may wish to find out if there is any significant difference between two large sets of scores that indicate measures of two variables, as exhibited by the same set of objects. Such variables like performance in two tests, reaction times, resistance to heat, adjustment level or time etc can be used for the same set of persons, animals, materials, chemicals, objects etc. Such samples are called correlated or non-independent samples. The research designs involved here can be the repeated measurement design where two measurements are made on each of the subjects in a sample, or the matched pairs design where the individuals in the two samples (experimental and control groups) are matched on one or more variables that are known to be correlated with the dependent variable or the criterion variable of the study. The test to be used is

$$z = \frac{\bar{d} - 0}{S/\sqrt{n}} = Z = \frac{\bar{d} - 0}{S/\sqrt{n-1}} = \frac{\bar{d}\sqrt{n-1}}{S}$$

$$z = \frac{\bar{d}(n-1)\sqrt{n}}{\sqrt{nEd^2 - (Ed)^2}} = \frac{Ed\sqrt{n-1}}{\sqrt{nEd^2 - (Ed)^2}}$$

where \bar{d} is the mean of the difference in scores between pairs of matched observations.

n = the sample size of one sample

S/\sqrt{n} = the standard error of measurement for the sampling distribution of the difference between means of matched/paired samples.

Example 5

Two tests – pre-test, post-test were given to a group of 40 students from Government Secondary School, Minna, Niger State in an experimental situation. The scores collected are as follows:

S/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X ₁	48	25	36	17	43	45	34	33	29	22	26	10	17	09	24	24	20	45	38	43
X ₂	40	30	28	21	40	46	30	29	27	25	18	08	19	07	18	23	23	43	38	36

S/N	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
X ₁	48	25	36	17	43	45	34	33	29	22	26	10	17	09	24	24	20	45	38	43
X ₂	40	30	28	21	40	46	30	29	27	25	18	08	19	07	18	23	23	43	38	36

Propose hypothesis and use it to verify whether there is a significant difference between the pre-test and post-test scores.

Solution:

- i. complete the composite table by introducing two more columns for d and d^2 as shown overleaf
- ii. find d = difference (linear) between X_1 and X_2
- iii. find \bar{d} = mean of $d = \frac{Ed}{n} = \frac{24}{40} = \underline{0.6}$
- iv. find Ed_2 = the sum of $d^2 = 896$

$$\text{v. find } S = \frac{\text{Standard deviation}}{\sqrt{\frac{40 \times 869 - 24^2}{40 \times 40}}} = \frac{\sqrt{\frac{nEd - (Ed)^2}{n^2}}}{\sqrt{\frac{34760 - 576}{1600}}} = \frac{\sqrt{\frac{34184}{1600}}}{21.365}$$

$$\text{vi. use the formula } Z = \frac{\bar{d}}{S/\sqrt{n-1}} \text{ or } \frac{d\sqrt{n-1}}{S} \text{ and substitute}$$

$$= \frac{0.6 \times \sqrt{39}}{21.365} = \frac{0.6 \times 6.244998}{21.365}$$

$$= \frac{3.7469988}{21.365} = 0.1753802 = \underline{0.175}$$

S/N	X ₁	X ₂	D	d ²
1	48	40	8	64
2	25	30	-5	25
3	36	28	8	64
4	17	21	-4	16
5	43	40	3	9
6	45	46	-1	1
7	34	30	4	16
8	33	29	4	16
9	29	27	2	4
10	22	25	-3	9
11	26	18	8	64
12	10	08	-2	04
13	17	19	-2	4
14	09	07	2	4
15	24	18	6	36
16	24	25	-1	1
17	20	23	-3	9
18	45	43	2	4
19	38	38	--	--
20	43	36	7	49
21	16	20	-4	16
22	17	27	-10	100
23	29	32	-3	9
24	33	31	2	4
25	14	08	6	36
26	36	34	2	4
27	41	38	3	6
28	25	31	-6	36
29	28	27	1	1
30	41	45	-4	16
31	26	28	-2	4
32	31	39	-8	64
33	47	38	9	81
34	20	18	2	4
35	28	32	-4	16
36	40	41	-1	1
37	46	38	8	64
38	25	25	--	--
39	18	16	2	4
40	35	37	-2	4

$z_{cal} < z_{tab}$. Accept H_0 .

$z_{cal} = z_{tab}$ without decision and try again.

In the example above $z_{cal} = 0.175$ while z_{tab} at $Z(n-1, 0.05) = z(40-1, 0.05) = (39 : 0.05)$

vii. If the null hypothesis is shown there is no significant difference between the pre-test scores and post-test scores of the 40 students from Government Secondary School, Minna in the experiment and the alternative hypothesis is there is a significant difference between the pre-test scores and the post-test scores; then we will use the available evidence to take a decision.

Can you remember the decision rule? The decision rule is as follows:

$z_{cal} > z_{tab}$. Reject H_0 .

But this is a two-tailed test. Therefore, the confidence level will be divided by two.

Thus we have $z(39 : 0.025) = 2.042$.

Since z_{tab} is greater than z_{cal} , we accept H_0 .

Activity:- 11.3

SN	23	24	25	26	27	28	29	30	31	32
X	36	39	39	42	26	27	28	38	20	26
Y	35	36	30	41	32	29	26	34	28	29

SN	12	13	14	15	16	17	18	19	20	21	22
X	33	27	24	22	21	31	34	22	30	34	36
Y	26	34	28	30	26	21	24	28	28	30	26

SN	1	2	3	4	5	6	7	8	9	10	11
X	33	35	28	27	30	26	26	38	29	32	39
Y	29	29	37	24	35	27	30	38	29	26	29

Using a null hypothesis, verify if there is significant difference between the pair of scores given below:

4.0 CONCLUSION

In this unit, you have learnt the Z-test, its uses and the basic assumptions underlying its use. You have also done some basic calculations in the application of the Z-test. You can use it in your research studies, since it is one of the prominent and powerful tests for researches which can be applied in a natural setting or normal distribution.

5.0 SUMMARY

In this unit, we discussed that the Z-test, as an inferential statistical test is based on the properties of the normal distribution. It is used to test the significant differences at a given alpha-level between: We also concluded that:

$$\text{Sample mean and population mean, } z = \frac{(X - \mu)\sqrt{n}}{\delta}$$

Two independent sample means, $z = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{S1}{n1} + \frac{S2}{n2}}}$

Two paired or matched sample mean, $Z = \frac{d\sqrt{n-1}}{S}$ etc.

6.0 TUTOR-MARKED ASSIGNMENT

- The mean performance score on an intelligence test of a population of primary six pupils preparing to take the National Common Entrance Examination from Lantano LGA of Plateau State is 90. A researcher sampled 50 schools from the Local Government Area and collected four of the best pupils from each school. A total of 200 pupils took the intelligence test. The results are:

Mean Score = 96, Standard deviation = 15

Formulate a suitable hypothesis and verify whether the selected group of pupils is actually above average.

- Given the scores of a group of students in a Technical Drawing (TD) test and Mathematics test as follows:

S/N	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
TD	14	18	12	14	19	12	11	10	20	14	15	19	11	20	15
MATHS	10	20	10	16	20	18	14	15	17	12	18	13	14	20	13

S/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
TD	8	12	14	18	14	16	19	14	12	13	14	11	15	10	12
MATHS	7	15	13	20	15	17	12	18	15	16	12	19	17	10	16

S/N	31	32	33	34	35	36	37	38	39	40	41	42
TD	16	10	18	15	19	14	13	19	14	17	14	19
MATHS	19	15	17	16	13	16	18	16	12	15	18	15

Propose a hypothesis to see whether the two sets of scores are significantly different.

7.0 REFERENCES/FURTHER READINGS

Ary, D. and Jacobs, L.C.(1976) Introduction to Statistics Purposes and Procedures. New York, Chicago, San Fransisco, Atlanta, Dallas, Montreal, Toronto, London, Sydney.

Ogomaka, P.M.C. (2004) Inferential Statistics for Research in Education and Social Sciences. Owerri. Peacewise Systems and Prints.

UNIT 2 Z –TESTS II

Table of Contents

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Difference between Population and Sample Proportions
 - 3.2 Difference between Two Independent Sample Proportions
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor – Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In the last unit, you learnt the uses of the Z-test and the basic assumptions underlying the use of the Z-test. You also did some calculations involving some aspects of the Z-test. In this unit, you will learn the remaining aspects of Z-test. As in the last unit, you will need to have your calculator and Z-table by your side. The test will be presented as natural as you will find and use it in the field or research.

2.0 OBJECTIVES

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

- i. Calculate Z for the difference between population and sample proportion.
- ii. Calculate Z for the difference between two independent samples' proportions.

3.0 MAIN CONTENT

3.1 Difference between Population and Sample Proportions

You have noticed that in the last unit, none of the variables treated presented proportional data. So in this sub-unit and the next, we will be handling scores or data that are related to proportions.

Have you noticed that in research, there are certain variables, attributes or characteristics which are exhibited by some objects of research that are naturally dichotomized or that take on only two values? There are also certain variables, attributes or characteristics which are possessed by some objects of research that are artificially dichotomized or can be assigned to only two values. Now, think of some examples of such variables, attributes or characteristics which can be naturally dichotomized or artificially dichotomized. For instance, when you talk about gender, your mind will go straight to male and female. This is one of the attributes that naturally dichotomized. Again, when you play a lottery game, you either win or loose; in a game of shooting, you either hit the target or you do not hit the target. These are some examples of artificially dichotomized.

If in a population designated for a study we have a proportion of the population on one side and the remaining part of the proportion on the other side, then we can use this type of Z-test. In this situation, the proportion of the population of the objects which possess or exhibit one value or attribute of interest should be known. Therefore, the known proportion becomes the population proportion. This is usually denoted by the small letter p. Hence, the proportion of the population which possesses the other value of the variable under consideration is denoted by the letter q. For example, in a study where the variable is gender. If a proportion of the target population which is male is known, this becomes p, while the proportion of the female will become q.

Therefore, to get q we use, $1 - p = q$. In such a study, the researcher may decide to verify whether the sample drawn 'belong to the populations', is typical of the population or is atypical of the population. The test to use is $\frac{P - p}{\sqrt{Pq/n}}$ or $\frac{nP - np}{\sqrt{npq}}$ where P = sample proportion, p = known population proportion, q = derived population proportion, n = number.

Example 1

An educational consultancy firm developed a standardized test which they use for the entrance examination into the Junior Secondary Schools in a state. The average number of pupils that take this examination from all the local government areas of the state is 24,000. In one year, one local government decided to find out if their students' performance in the entrance is significantly greater or otherwise than the performance of the entire group of pupils from the whole state. If the proportion of the entire students who passed from at least the 250 marks out of 300 is 0.54, and the

proportion of the pupils from the local government numbering 500 is 0.60 (who passed with at least 250), use a hypothesis to verify the local government claim.

Solution:

i. Proposed a hypothesis

$H_0 =$ The proportion of the pupils from the local government who scored at least 250 is not significantly greater than the proportion of pupils from the entire state who scored at least 250. ($P < 0.05$).

ii. $z = \frac{P - p}{\sqrt{Pq/n}}$ $P = 0.60$, $p = 0.54$, $q = 1 - p$
 $q = 1 - 0.54 = 0.46$, $n = 500$

$$\begin{aligned} \square z &= \frac{0.60 - 0.54}{\sqrt{\frac{0.54 \times 0.46}{500}}} = \frac{0.06}{\sqrt{\frac{0.2484}{500}}} \\ &= \frac{0.06}{0.022289} = 2.6919108 \\ &= \underline{\underline{2.692}} \end{aligned}$$

Decision:

z_{tab} at 0.05 = 1.96

$z_{cal} = 2.692$

Since z_{cal} is greater than z_{tab} , we reject the H_0 and accept the H_1 . it implies that the proportion of the pupils from the local government who scored at least 250 is significantly greater than the proportion of pupils from the entire state who scored at least 250 out of 300.

Example 2

A publishing house has the reputation of having 0.45 of its products labeled standard. In the last few years, the chief editor went on retirement and a new chief editor employed. This new editor claims he is better than the retired one. The publishing house therefore selected 80 titles from the ones

published during this year's production. These titles were subjected to test by standard organization. The result shows that 25 out of the 80 titles were labeled standard. Is the new chief editor better as he claimed?

Solution:

- i. Propose a null hypothesis as follows: the proportion of book titles labeled standard, produced during this year's production not significantly different from the proportion when the retired chief editor was on seat. ($P < 0.05$).

$$\text{ii. } z = \frac{P - p}{\sqrt{pq/n}} \cdot P = 0.45, p = \frac{25}{80} = 0.313$$

$$q = 1 - p = 1 - 0.45 = 0.55$$

$$n = 80$$

$$= \frac{0.313 - 0.45}{\sqrt{\frac{0.45 \times 0.55}{80}}} = \frac{-0.137}{\sqrt{\frac{0.2475}{80}}}$$

$$= \frac{-0.137}{0.05562214} = 2.4630808 = \underline{\underline{2.463}}$$

Decision:

$$z_{cal} = 2.463 \cdot \text{The absolute value} = \underline{\underline{2.463}}$$

$$z_{tab} = 2.000$$

$z_{cal} > z_{tab} \therefore$ We reject H_0 . It means that the new chief editor is significantly better than the former editor.

Activity 1

It is on record that the proportion of students passing Technical Drawing with credits and above in a school called Umunne-Umuihi High School, Etiti during the years is 0.56. Last year, a new Technical Teacher was posted to this school, while the former Technical teacher resigned and joined politics. The new Technical teacher boasted that he was better than the former. The school wanted to verify this claim by subjecting the results of his students to verification statistically. 65 students took technical

drawing in SSCE this year and 30 of them passed with credits and above. Is the new Technical teacher significantly better than the former one? Use the evidence to verify. (Check your answer at the end of the unit).

3.2 Difference between two independent Samples' Proportions

You have noticed that in the last sub-unit, the interest of the researcher was on variables that have values which are dichotomized either naturally or artificially. Can you list some of such objects or characteristics that can yield values which are dichotomized?

Activity 2

List some variables that have values which are dichotomized. Say whether they are naturally dichotomized or artificially dichotomized.

In this sub-unit, the focus is also on variables which have values that can be dichotomized. But in this case, the investigation is on the difference between proportions of two independent samples. If you have worked through the last sub-unit very well, you will have no difficulty working through this.

Example 3

Two randomly selected troops of youth corps members were drilled in physical exercises by the instructors during their orientation camping. Troop A = 50, Troop B = 48. By the end of the orientation exercises, 28 members of troop A and 32 members of troop B were declared unfit to undertake strenuous exercises. Verify whether the proportions of the two troops declared unfit for further exercises are significantly different.

Solution:

H_0 : The proportions of the two troops declared unfit for further exercises are not significantly different. ($P < 0.05$).

$$\text{The test to be used} = z = \frac{p_1 - p_2}{\sqrt{\frac{pq(n_1 + n_2)}{n_1 n_2}}}$$

$$\text{where } P = \frac{n_1 P_1 + P_2 P_2}{n_1 + n_2} \cdot q = 1 - P$$

$$\begin{aligned} P_A &= \frac{28}{50} = 0.56. \quad P_B = \frac{32}{48} = 0.67, \quad n_1 = 50, \quad n_2 = 48 \\ &= P = \frac{(50 \times 0.56) + (48 \times 0.67)}{50 + 48} = \frac{28 + 32.16}{98} = \frac{60.16}{98} \\ &= 0.6138775 = 0.614 \end{aligned}$$

$$\begin{aligned} \square q &= 1 - P = 1 - 0.6138775 = 0.386225 \\ &= 0.386 \end{aligned}$$

$$\begin{aligned} \square z &= \frac{0.56 - 0.67}{\sqrt{\frac{0.614 \times 0.386(50 + 48)}{50 \times 48}}} \\ &= \frac{-0.11}{\sqrt{\frac{0.237004 \times 98}{2400}}} = \frac{-0.11}{\sqrt{\frac{23.226392}{2400}}} \\ &= \frac{-0.11}{0.0983751} = -1.1181691 = \underline{\underline{-1.118}} \end{aligned}$$

$$\text{Absolute value} = \underline{\underline{1.118}}$$

Decision:- $z_{cal} = 1.118$; $z_{tab} = 1.96$; $z_{cal} < z_{tab}$. We accept H_0 . It implies that the proportions of the two troops declared unfit for further exercises are not significantly different. It also means that the observed difference, i.e. apparent difference, may be due to sampling error.

Example 4

An educational research organization developed an instrument called Standardized Students Laboratory Interest Scale (SSLIS), for use in senior secondary schools. In the manual, it was recommended that for effective use of the instrument, it should be revised every five years. In the second year of its use, a school administered the instrument on a randomly selected students offering sciences and numbering 150. The students were asked to indicate their opinion about the test. 0.65 of the students indicated the test was very good. After five years i.e. in the seventh year of the use of the

instrument, the school again selected 100 science students randomly and administered the instrument on them. The students' opinion showed 0.70 indicating that the test was very good. Do you think this instrument needs revision?

Solution:

H_0 = There is no significant difference between the proportions of the students who indicated that the instrument was very good in the second year and in the seventh year.

$$P_B = 0.70. P_A = 0.65. n_B = 100, n_A = 150$$

$$z = \frac{P_B - P_A}{\sqrt{\frac{pq(n_A + n_B)}{n_A + n_B}}} = \frac{P_B - P_a}{\sqrt{\frac{pq(n_A + n_B)}{n_A + n_B}}} = \frac{(150 \times 0.65) + 100 \times 0.70}{150 \times 100}$$

$$= \frac{97.5 + 70}{15000} = \frac{6825}{15000}$$

$$= 0.455$$

$$\square q = 1 - P = 1 - 0.455 = 0.545$$

$$\square z = \frac{0.70 - 0.65}{\sqrt{\frac{0.455 \times 0.545 (150 + 100)}{150 + 100}}} = \frac{0.05}{\sqrt{\frac{0.2479 \times 250}{250}}}$$

$$= \frac{0.05}{\sqrt{\frac{61.99375}{250}}} = \frac{0.05}{\sqrt{0.247975}}$$

$$= \frac{0.05}{0.4979708} = 0.1004074$$

$$= \underline{\underline{0.101}}$$

Decision: $z_{cal} = 0.101$ $z_{tab} = 1.96$

Since z_{cal} is less than z_{tab} , we accept that there is no significant difference between the two proportions. It implies that the instrument does not need any revision yet.

Activity 3

Given the following as the result of a research carried out by Post graduate student of a University:

Sample A, number 120, proportion 0.63.

Sample B number – 80, proportion = 0.75.

Are the two sample proportions significantly different? (compare your answer with the one at the end of the unit).

4.0 CONCLUSION

Having successfully learnt this unit, you can now handle any type of research which needs to employ the use of the Z-test. You can be sure you are ready to learn the next unit which is going to be on t-test. They are very powerful tests in research and statistics.

5.0 SUMMARY

In this unit, you have learnt the remaining and concluding part of the Z-test. In this, we have gone through how to use the Z-test when we have to find the difference between population and sample proportions. We have also discussed how to find the difference between samples' proportions. You will remember that we said that variables in this unit are those that have values that can be dichotomized either naturally or artificially.

6.0 TUTOR-MARKED ASSIGNMENT

The proportions of male and female learners from samples' sizes of 75 and 66 respectively from a distance education institution who are addicted to the use of coffee as a stimulant for reading are given as 0.35 and 0.24 respectively. Are these proportions significantly different? ($P < 0.05$).

7.0 REFERENCES/FURTHER REFERENCES

Ary, D and Jacobs, L.C (1976) Introduction to Statistics: Purposes and Procedures. New York, Chicago...Sydney, Holt, Rinehart & Winston.

Ogomaka, P.M.C. (2004) Inferential Statistics for Research in Education & Social Sciences. Owerri, Peacewise Systems & Prints.

UNIT 3 t – TEST

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Origin of the t-test
3.2	Assumptions of the t-test
3.3	Difference between population and sample means
3.4	The t-test for the difference between two means
3.5	Difference between two matched sample means
4.0	Conclusion
5.0	Summary
6.0	Tutor Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

In the last two units we discussed the z-test and said it is a parametric test which can be used when the number is large. In this unit, we shall look at the t-test which is a more versatile parametric test. This is because, like the z-test, it can be used for:

- (a) the significant difference between population mean and sample mean
- (b) the difference between means of two matched samples
- (c) the significant difference between means of two independent samples etc.

The t-test applies in all situations where z-test can be applied; like in large numbers as well as when the number is not large (less than 30). However, if the number is too small (less than 10) t-test is not reliable. Before we go to the applications of the t-test, let us specify the objectives of this unit and look at the origin of the t-test.

2.0 OBJECTIVES

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

- i. List the assumptions underlying the use of t-test
- ii. Calculate the t-test for the significant difference between population mean and sample mean.
- iii. Test the difference between means of two matched samples
- iv. Verify the significant difference between means of two independent samples

3.0 MAIN CONTENT

3.1 Origin of the t –Test

The t-test was developed to ensure minimum error in determining the trend of distribution if small samples are involved in experiments or studies. Some of the times small samples may be wrongly considered to be normally distributed as the population they were drawn from. In this case, the t-test would be introduced to avoid the error which would have occurred.

The t-test was developed in 1915 by a consulting statistician. William S. Gosset who was working for Guinness Breweries in Dublin, Irish Republic. This time, he published an article in which he worked out the equation of a distribution statistics called t-distribution or student's t to determine the nature and scope of distribution whose population variance is not known and therefore a non-normal distribution. Gosset published his work under the pseudonym 'student' because the brewery he worked for prohibited its staff from publishing any works. Today, the statistic is referred to as student's t-test or simply t-test.

3.2 Assumptions of the t-Test

Now that you have known the origin of the t-test, let us look at the assumptions. The basic assumptions here are those conditions which should prevail or be guaranteed if conclusions are to be valid. The assumptions are the same with those which you have seen in the z-test. These assumptions are:

- i. The population from which the sample is drawn is normally distributed.

- ii. The population variances as estimated by the samples variances are homogeneous
- iii. The samples are independently and randomly drawn from the target population.
- iv. The variables yield continuous values.
- v. The sample size may not be large but not less than ten.

3.4 Difference between Population and Sample Means

This is given by the model t-test = $\frac{\bar{X} - \mu}{S/\sqrt{n-1}}$ or $\frac{(\bar{X} - \mu)(n-1)}{\sqrt{\sum X^2 + \frac{(\sum X)^2}{n}}}$

Can you compare this formula with that of z-test?

Example 1

A psychometrician made a standardized test for Junior Secondary School students in Lokoja Educational Zone. He came out with a result saying that the mean for Lokoja zone is 52%. Using a sample $n = 200$, $\bar{X} = 49.5$ and $S = 12$, verify his claim.

Solution

1. Propose a null hypothesis that there is no significant difference between the population mean and the sample mean at $p < 0.05$.

Given that: $n = 200$, $\bar{X} = 49.5$, $S = 12$, $\mu = 52\%$, $p < 0.05$.

$$\begin{aligned} \frac{\bar{X} - \mu}{S/\sqrt{n-1}} &= \frac{\bar{X} - \mu \sqrt{n-1}}{S} \\ t &= \frac{(49.5 - 52)\sqrt{199}}{12} = \frac{2.5\sqrt{199}}{12} \\ &= \frac{2.5 \times 14.106736}{12} = \frac{35.26684}{12} = 2.9389033 \\ &= \underline{\underline{2.94}} \end{aligned}$$

You will recall that to take your decision, you will have to compare the t-calculated (tcal) with the t on the table (ttab), using the degree of freedom n-1 and the confidence level or alpha level of 0.05. Therefore, ttab at (199.0.05) = 1.96, tcal = 2.94. Since tcal is greater than ttab, we reject the null hypothesis and accept the alternative hypothesis H₁ that there is a significant difference between the population mean and the sample mean.

Activity 1

A Technical Drawing class in a Comprehensive College located in Jalingo zone of Taraba State had 25 students. The teacher gave them an examination which, based on years of use, had been shown to have a mean of 80. The result showed a mean of 84 and standard deviation of 10 on the test. Is the difference between this class mean and the population mean so significant at 0.05?

3.5 The t-test for the difference between two means

When you draw up two samples randomly and/or independently from a normal population; if the variances of the sample as estimates of the population variance are homogeneous or do not differ significantly, then a sample of the differences between means of pairs of such samples has a t-distribution. In other words, the estimated variability of the differences between the means of two random samples that can be expected due to chance factors can be determined using t-test to see whether the observed difference between the two means is likely to be a function of chance or not t-statistics can be given by

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\delta X_1 - X_2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where \bar{X}_1 and \bar{X}_2 are means of the two samples.

n_1 and n_2 are the sample sizes or numbers

$\delta X_1 - X_2$ is the standard error of the difference between means.

But it is not easy to get the δ . Therefore, we convert the above

$$\text{formula as follows: } t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2](n_1 + n_2)}{(n_1 + n_2 - 2)(n_1)(n_2)}}}$$

$$\text{where } S = \sqrt{\frac{\sum (X - \bar{X})^2}{n_1 - 1}}$$

$$\text{but if } n_1 = n_2 = n \text{ then } t \text{ is given by } t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The most widely used formula is where the t-test is calculated directly from the raw scores and it is given by

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{[\sum X_1^2 - (\sum X_1)^2/n_1] + [\sum X_2^2 - (\sum X_2)^2/n_2]}{n_1 + n_2 - 2}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\text{Where } S = \sqrt{\frac{\sum (X - \bar{X})^2}{n_1 - 1}}$$

$$\text{But if } n_1 = n_2 = n \text{ then } t \text{ is given by } t = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The most widely used formula is where the t-test is calculated directly from the raw scores and it is given by $t =$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{[\sum X_1^2 - (\sum X_1)^2/n_1] + [\sum X_2^2 - (\sum X_2)^2/n_2]}{n_1 + n_2 - 2}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Where $\bar{X}_1 - \bar{X}_2$ = the difference between the two means

$\sum X_1^2$ = the sum of the square of each X-score in group I

$\sum X_2^2$ = the sum of the squares of each X-score in group 2.

$(\sum X_1)^2$ = the sum of the raw scores in group 1 squared.

$(\sum X_2)^2$ = the sum of the raw scores in group 2 squared.

n_1 = number of cases in group 1

n_2 = number of cases in group 2.

Example 2

Some researchers are interested in finding out the effect of two methods of motivation on the learning outcome of students. They decided to use two randomly sampled groups of students from a population of SSI students. One sample group received motivation method A and the other group received motivation method B. After a period of time, the two groups were compared using a set of achievement tests. The results are as indicated below:

S/N	X_1	X_1^2	X_2	X_2^2
1	18	324	14	196
2	16	256	12	144
3	15	225	10	100
4	14	196	9	81
5	12	144	8	64
6	11	121	6	36
7	10	100	5	25
8	9	81	4	16
9	7	49	3	9
10	5	25		
Σ	117 $n_1=10$	1521	71 $n_2=9$	671

Solution:

- Find the squares of X_1 and X_2 scores.
- Find the sum of squares = ΣX_1^2 and ΣX_2^2
- Find their means $\overline{X_1}$ and $\overline{X_2}$
- Using the formula for t-test, we have

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

$$\begin{aligned}
& \frac{11.7 - 7.9}{\sqrt{\frac{1521 - \left(\frac{117^2}{10} + 671 - \left(\frac{71^2}{9}\right)\left(\frac{1}{10} + \frac{1}{9}\right)\right)}{10 + 9 - 2}}} \\
&= \frac{3.8}{\sqrt{\frac{(152.1 + 1368.9) + (671 - 560.1)\left(\frac{19}{90}\right)}{17}}} \\
&= \frac{3.8}{\sqrt{\frac{(152.1 + 110.9)\left(\frac{19}{90}\right)}{17}}} \\
&= \frac{3.8}{\sqrt{\frac{263}{17}(0.21111)}} = \frac{3.8}{\sqrt{15.470588 \times 2111}} \\
&= \frac{3.8}{\sqrt{3.2659959}} = \frac{3.8}{1.0072067} \\
&= t = 3.7728105 = \underline{\underline{3.77}}
\end{aligned}$$

If the null hypothesis H_0 is that there is no significant difference between the means of the two samples that received the two types of motivation; degree of freedom (df) = $n_1 + n_2 - 2 = 10 + 9 - 2 = 17$

alpha level = 0.05

$t_{cal} = 3.77$ and $t_{tab} = t(14, 0.05) = 2.110$

therefore the decision will be:

since $t_{cal} > t_{tab}$ we reject that there is no significant difference between the two samples.

Activity: 1

Two groups of pupils in a primary school took an aptitude test and the results are as follows:

Group 1	40, 35, 45, 25, 30, 20, 15, 18, 38, 29, 30, 41
Group 2	32, 40, 22, 38, 25, 31, 37, 41, 29, 15, 19, 14

1. Find the t-value
2. Determine if the two groups are significantly different.

3.5 Difference between two matched sample means

In the previous section, we considered research studies in which the two samples are independently drawn from their respective populations. But sometimes research studies are concerned with two samples which are related to each other. These can be as a result of repeated measurements design, in which case, it could be pre-test, post-test design or using the same subjects in a sample under different treatments or conditions. It can also be a case of matched pairs design in which case the individuals in the two samples are matched on one or more variables, under consideration.

The formula to be used is $t = \frac{\bar{d}}{S} \sqrt{\frac{n-1}{S}}$ where \bar{d} = the mean of the difference in scores between pairs of matched observations.

$$S = \text{standard deviation} = \sqrt{\frac{n \sum d^2 - (\sum d)^2}{n}}$$

Example 3

15 students were given two tests, one in Mathematics and the other in Technical Drawing. The results are as shown. Find whether the two scores are significantly different.

S/N	X ₁	X ₂	d	d ²
1	48	40	8	64
2	25	30	-5	25
3	36	28	8	64
4	17	21	-4	16
5	43	40	3	09
6	45	46	-1	01
7	34	30	4	16
8	33	29	4	16
9	29	27	2	04
10	22	25	-3	09
11	26	18	8	64
12	10	08	2	04
13	17	19	-2	04
14	09	07	2	04
15	24	18	6	36
Σ			32	336
$\bar{d} = \frac{32}{15} = 2.1$				

Solution:

- i. Propose a null hypothesis like: there is no significant difference between the two sets of scores from the students.
- ii. Find the d = difference (linear) between X₁ and X₂ in the Composite table.
- iii. Find \bar{d} = mean of d = $\frac{\Sigma d}{n} = 2.1$
- iv. Find $\Sigma d^2 = 336$
- v. Find S = standard deviation using

$$\sqrt{\frac{n \Sigma X^2 - (\Sigma X)^2}{n^2}} = \sqrt{\frac{15 \times 336 - 1024}{15 \times 15}} = \underline{\underline{4.225}}$$

$$\begin{aligned}
 \text{vi.} \quad \text{substitute for } t &= \frac{d\sqrt{n-1}}{S} \\
 &= \frac{2.1\sqrt{14}}{4.225} = \frac{2.1 \times 3.7416574}{4.225} \\
 &= \frac{7.8574805}{4.225} = 1.859758698 = \underline{\underline{1.86}}
 \end{aligned}$$

Decision:

$t_{cal} = 1.86$. t_{tab} at $t(n-1, 0.05) = t(14, 0.05) = 2.145$

t_{cal} is less than t_{tab} . Therefore, we accept H_0 , meaning that there is no significant difference between the two sets of scores.

Activity: 2

Given the two sets of scores generated in an experiment by 20 students in a class as follows:

S/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	55	20	31	48	18	15	22	42	19	25	32	45	50	38	43	17	39	42	51	11
Y	65	42	46	60	25	20	20	40	20	30	38	50	50	36	45	20	41	45	55	10

Find out if the two set of scores are significantly different.

4.0 CONCLUSION

We have said that for small samples, some statistics exhibit sampling distributions that depart from normality in various ways. The student's t is a small-sample statistical test. It can be used in place of z-test for large samples. It is defined like the z as the ratio of a deviation from the mean or other parameter in a distribution of sample statistics, to the standard error of that distribution. You can therefore use the t-test for small samples and in place of z for large samples.

5.0 SUMMARY

In this unit, we have said that the t-test was developed to ensure minimum error in determining the trend of distribution when small samples are involved and even large samples. Recall that for your conclusions to be valid there are some assumptions that should be guaranteed. It is used in all

situations where the z-test can be used and even where z cannot be used like when the number is small.

6.0 TUTOR-MARKED ASSIGNMENT

Pretest	10	15	37	29	40	35	25	16	20	11	15	5	8
Post-test	60	42	61	37	35	58	44	37	50	15	35	10	45

Test whether the post-test is significantly different from the pre-test.

7.0 REFERENCES/FURTHER READINGS

- Ali, A. (1996) Fundamental of Research in Education. Awka, Onitsha. Meks Publishers (Nig.).
- Ary, D. and Jacobs, L.C. Introduction to Statistics Purposes and Procedures. New York, Chicago...Sydney, Toronto. (1972) Holt Rinehart and Winston.
- Ogomaka, P.M.C.(2004) Inferential Statistics for Research in Education & Social Science. Owerri. Peacewise Systems.

In this case, Z or t-tests are ruled out. Therefore, the appropriate technique or test to use is called Analysis of Variance (ANOVA). In this unit, we shall discuss some of the uses and applications of ANOVA in relation to educational researchers.

2.0 OBJECTIVES

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

- i. Explain the concept of Analysis of Variance.
- ii. Describe the sources of variances in ANOVA
- iii. Explain the sum of squares
- iv. Explain the mean squares
- v. Define the F-ratio.

3.0 MAIN CONTENT

3.1 The Analysis of Variance (ANOVA)

In the last unit, you learnt the uses and applications of the t-test and you noticed that it can only be used with only two means. The analysis of variance which is also called one way analysis of variance (one-way ANOVA), like the t-test can be used in a two-sample situation. When it is used with only two-sample means, it is mathematically equivalent to the t-test. So, you can use it in the place of t-test. But ANOVA is most commonly used when there are three or more samples. It is a more versatile statistical technique than the t-test. In fact, most statisticians prefer to use it even when there are only two samples. It is one of the most widely used statistical tests. It is used for comparing the significance of the differences between two or more independent means. The underlying assumption is that the independent means are bound to be different. Therefore, ANOVA enables us to determine whether such variances or differences are due to chance alone or sampling error or indeed as a result of the effects of the independent variable on the dependent variable. This is very effective especially if the study is a true experimental study that involves two or more groups. Now, think of a situation where you have six groups and you decide to use t-test, it means that doing it pair wise would be cumbersome as you may have to do it several times. To avoid this, we can use the ANOVA especially to test whether any observed differences between groups or error terms are significant or not.

Note that the general rationale for the use of ANOVA is that the total variance of all the scores or data in an experimental study can be separated

and attributed to two sources. These sources are ‘variance between groups or variance among groups’ and ‘variance within groups.’

- i. Variance within groups reflects the spread of scores or data within each of the number of groups. It represents differences among subjects that have nothing to do with the independent variable in the experiment. It is sometimes called error variance.
- ii. Variance between groups reflects the magnitude of the difference between the number of group means. It may be due to the effect of the independent variable or just a function of chance.

You will remember that we have said earlier that the purpose of ANOVA is to establish whether the variation between groups is likely to be a function of chance or not. There are three types of ANOVA models. These are the random model, the fixed model and the mixed models. Because the fixed model is the one which is most widely used in educational and psychological researches, our descriptions here will be based on the fixed model. Again, the mathematical considerations and derivations of the statistical technique and formula are considered beyond the scope of this course. So we shall only touch a small aspect of the possible application and uses of the ANOVA statistics.

3.2 The sum of Squares

The sum of squares is the basic ingredients of the ANOVA procedure. It is the measure of variability which is analyzed here. It is the total of the squared differences or the deviations between a set of individual scores and a mean. Look at this notation $\sum X^2$ or $\sum (X - \bar{X})^2$. What does it represent? You will recall that it represents sum of squares. In ANOVA, sum of squares are not represented like that, rather by convention the symbol SS is used to represent sum of squares. Now let us look at the various types of sum of squares as are used in ANOVA.

- i. **Total sum of squares:** This is represented by the symbol SSt. It refers to the sum of squares of the deviations of each of the observations from the grand mean. The mean of all the scores taken together as a group is called the grand mean. It is represented by the symbol \bar{X} . However, the total sum of squares is given by $SSt = \sum (X - \bar{X}_t)^2$ where SSt = total sum of squares
 X = an individual score
 \bar{X}_t = the mean of all the scores (grand mean).

- ii. **Sum of squares within groups:** This is also symbolized by SS_w . It is a component of the total sum of squares that can be calculated. It is a basic component of the error term. It is not related to any difference in treatment. It can be found by calculating the deviation of each individual score in each group from the mean of its own group and then squaring and adding up the squared deviations. It is given by $SS_w = \sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2 + \sum(X_3 - \bar{X}_3)^2 + \sum(X_4 - \bar{X}_4)^2 \dots \sum(X_n - \bar{X}_n)^2$ where SS_w = sum of square within, \bar{X}_1 = the mean of the first group, \bar{X}_2 = the mean of the second group, \bar{X}_3 = the mean of the third group etc.

X_1 = an individual score in the first group

X_2 = an individual score in the second group

X_3 = an individual score in the third group etc.

n = number of groups.

- iii. **Sum of squares between groups:** This is the variability from group to group represented by SS_b . It is also a component of the total sum of squares. It is the variation that may be due to the experimental treatment which is sometimes called treatment SS^1 . It is got by computing the sum of squares of the deviations of each separate group mean from the grand mean. It is given by the formula $SS_b = n_1(\bar{X}_1 - \bar{X}_t)^2 + n_2(\bar{X}_2 - \bar{X}_t)^2 + n_3(\bar{X}_3 - \bar{X}_t)^2 \dots$ to all the groups, where

SS_b = sum of squares between \bar{X}_1 the mean of the first group.

\bar{X}_1 = the mean of the first group

\bar{X}_2 = the mean of the second group

\bar{X}_3 = the mean of the third group etc

n_1 = the number in the first group,

n_2 = the number in the second group, etc.

\bar{X}_t = the mean of all the scores.

Activity 1

- i. What is the analysis of variance?
- ii. What are the two sources where in experimental studies are attributed to?
- iii. What do the following mean? SSt, SSw and SSn

Answers to Activity 1

- i. Analysis of variance is a versatile test which is widely used for comparing the significance of the differences between two or more independent means. It helps us to determine whether the variances or differences are due to chance alone or sampling error or as a result of the effects of the independent variable on the dependent variable.
- ii. Variance between groups and variance within groups
- iii. SSt = total sum of squares
SSw = sum of squares within groups
SSb = sum of squares between groups.

3.3 Mean Squares

The sum squares between and within can be used to describe estimates of the population variance in ANOVA. These are then used to verify the null hypothesis. If the SSb and SSw are divided by their appropriate degrees of freedom in order to obtain the estimates of variance that are needed, the two variance estimates are called mean squares. It is symbolized by MS. Just as we have SSb and SSw, in the same way we have the mean square for within groups and for between groups. What do you think will be the symbols?

- i. Mean square within groups: This is the estimate of the variance derived from within groups' data. It is given by:

$$MS_w = \frac{SS_w}{df_w} = \frac{SS_w}{N - K}$$

Where SSw = sum of squares within groups

N = total number of scores

K = the number of groups

You will have to note that N-K is given as the number of degrees of freedom associated with the within-groups sum of squares because the degrees of freedom for the separate groups are summed up. So if

the number of degrees of freedom for each group is $n-1$ and for K groups then $(n_2 - 1) + (n_3 - 1) + \dots + (n_k - 1) = \sum n - k = N - K$ degrees of freedom.

- ii. Mean square between groups. MSb: This is the estimate of the variance between groups. It is given by the formula:

$$MSb = \frac{SSb}{dfb} = \frac{SSb}{K - 1} \text{ - where}$$

SSb = the sum of squares between groups

K = the number of groups.

Again, you will have to take note that $K-1$ is the number of degrees of freedom associated with the between-groups sum of squares. This is because there are K means and 1 degree of freedom is lost by subtracting the grand mean from each group mean. It is possible to add the degrees of freedom for both the within and the between groups in order to make it significant. Thus, $N - 1 = (N - K) + (K - 1)$ i.e. Total Within + Between.

3.4 The F-Ratio

In the last sub-unit 3.3 you learnt about the mean squares = MSb and MSw.

F can be defined as the ratio of $F = \frac{MSb}{MSw}$.

The number in this F-ratio can be influenced by the observed differences between the groups, while the denominator represents the error term since it is derived from variation within groups. This F-ratio is used for the comparison of the estimate of the population variance derived from the sample between-groups data MSb, to the estimate of the population variance estimate derived from the sample within-groups data, MSw. Note that as the difference between the groups' increases, the F-ratio increases. In this test, ANOVA, the null hypothesis that is verified is that the sample means being composed with the F-ratio are not different from what is expected from random samples from the same population. It means therefore that if the null hypothesis is true, then the variance estimate based on the differences between groups and the variance estimate based on the difference within groups will tend to be about the same. This is because both of them are estimates of the same common population variance. Therefore, any difference in the two mean squares would be the result of random variation. You will thus expect the ratio of MSb/MSw to be about

I. When there is a genuine difference between the groups, the MS_b or the variance estimate derived from the variance of groups means around the grand mean is markedly greater than the mean square within groups or the variance estimate derived from the variation of scores within each group, and F will be considerably greater than I. As the difference between the mean squares increases, the F-ratio increases and the probability that the null hypothesis is correct, decreases. It is only when the values of F are greater than I that they would be considered as evidence against the null hypothesis. Before we continue, let us pause so that you can do this activity. Do not proceed until you are through with it.

Activity 2

- i. The ratio which is used to test hypothesis in ANOVA is called.....
- ii. What is the mean square?
- iii. What is the symbol of mean square?
- iv. What are the two types of mean square?
- v. What is the formula for F-ratio?

Answers to Activity 2

- i. F-ratio
- ii. Mean Square: the division of the SS_b and SS_w by their appropriate degrees of freedom in order to obtain the estimates which are called mean squares. It is symbolized by MS.
- iii. The symbol of mean square = MS
- iv. The two types of mean square are
 - a. Mean square within groups MS_w and
 - b. Mean square between groups MS_b
- v. The formula for F-ratio is $\frac{MS_b}{MS_w}$

4.0 CONCLUSION

In this unit, we have treated the analysis of variance which is widely used statistical test. This is because it is appropriate for testing hypothesis about means in a variety of experimental situations. It can be used with only two groups as well as with more than two groups. It is therefore a more versatile test than the t-test.

5.0 SUMMARY

In this unit, you have studied the analysis of variance (ANOVA) otherwise called one-way analysis of variance. You have seen that it is used to test the statistical significance of differences between means when more than two groups are involved in an experimental research. The total variance of all the scores or data in an experimental study can be attributed to two sources. These are variance between groups and variance within groups. The sum of squares involved in ANOVA are total sum of squares SS_t , sum of squares within groups, SS_w and sum of squares between groups SS_b . You also learnt about the mean squares which can be within and between groups, MS_w and MS_b . The F-ratio which is the test for hypothesis in the ANOVA is given by $F = \frac{MS_b}{MS_w}$. In the next unit, we shall be applying all these in some calculations involving ANOVA.

6.0 TUTOR-MARKED ASSIGNMENT

- i. What is the full meaning of ANOVA?
- ii. Where does ANOVA apply?
- iii. What is the sum of squares?
- iv. What are the 3 types of sum of squares?
- v. What is the mean square?
- vi. What are the 2 types of mean squares?

7.0 REFERENCE/FURTHER READINGS

Ali, A (1996) Fundamentals of Research in Education, Awka Meks, Publishers (Nig.)

Ary, D and Jacobs, L.C (1976) Introduction to Statistics: Purposes and Procedures. New York, Chicago. Sydney, Holt, Rinehart & Winston.

Ogomaka, P.M.C. (2004) Inferential Statistics for Research in Education & Social Sciences. Owerri, Peacewise Systems & Prints.

UNIT 2 ANALYSIS OF VARIANCE – USES & APPLICATIONS

Table of Contents

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Computational formula for sum of squares
 - 3.2 Calculations for illustrations
 - 3.3 Another illustration
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In the last unit, you learnt that the analysis of variance has to do with variance of population as estimated by the samples' variance. You also noted that the purpose of ANOVA is to establish whether the variation between groups is likely to be a function of chance or not. We looked at the variances – total variance, variance within groups and variance between groups. As you know variance of a population or sample is calculated based on sum of squares: total sum of squares, sum of squares within and sum of squares between groups. You also learnt about the F-ratio which is used to test hypothesis in ANOVA. In this unit, we shall try to apply and use the ANOVA test. This will enable you master the application.

2.0 OBJECTIVES

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

- i. Explain the computational formula for sum of squares
- ii. Calculate the F-ratio when given the data
- iii. Verify hypothesis involved with ANOVA

3.0 MAIN CONTENT

3.1 Computational formula for Sum of Squares

In ANOVA, calculation of F-ratio involves only getting the sum of squares total, sum of squares between groups and sum of squares within groups. From these, we can obtain the mean square between, the mean square within and finally the F-ratio. You do not need to find the deviation scores as this will waste a lot of time. Instead, we shall introduce a simpler method which involves dealing directly with the raw scores. Now let us list the steps to follow.

- i. Calculate the total sum of squares (SSt) using

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

where N = the total number of observation ($n_1+n_2+n_s+\dots n_k = N$)

- ii. Determine the sum of squares between groups given by:-

$$SSb = \frac{\sum X_1^2}{n_1} + \frac{\sum X_2^2}{n_2} + \frac{\sum X_3^2}{n_3} + \dots + \frac{\sum X_k^2}{n_k} - \frac{(\sum X)^2}{N}$$

where $\sum X_1$ = sum of the first group. $\sum X_2$ = sum of the second group etc. n_1 = number in the first group, n_2 = number in the second group etc. $\sum X$ = sum of all the scores. N = total number of scores = $n_1+n_3+\dots nk$.

- iii. Determine the sum of squares within groups given by:-

$$SSw = \sum X_1^2 - \frac{(\sum X_1)^2}{n_1} + \sum X_2^2 - \frac{(\sum X_2)^2}{n_2} + \dots + \sum X_k^2 - \frac{(\sum X_r)^2}{n_k}$$

You may not need to use this formula because it is longer, except to use it as a check or verification formula, but a very simple short cut to this formula is to use $SSw = SSt - SSb$. This means that once you compute the total sum of squares and the between sum of squares, you then subtract: $SSt - SSb$ or get the within sum of squares. Note that the null hypothesis to be tested is given by $\mu = \mu_2 = \mu_3$, while the alternative hypothesis is that not all means are equal ($\mu, \neq \mu_2 \neq \mu_3$).

3.2 Calculations for illustration

A teacher in his research study set out to find the effect of group projects, class project and individual project works on learning outcomes of students. He used Technical Drawing for his experiment; at the end he generated the data below:

BP= X_1	15	13	12	14	10	18	20	19	11	9	7	13	19	20	4	8	5	16	18	20
CP= X_2	20	16	18	11	16	18	12	14	15	19	10	11	4	6	9	2	20	14	13	19
IP= X_3	10	15	19	10	14	13	16	17	11	12	18	14	17	9	5	7	11	10	14	13

In this experiment, the teacher wanted to find out if there are any effects of these projects on the students' performance or if the three samples belong to the same population.

Solution

- i. Propose a null hypothesis. Thus $H_0 = \mu_1 = \mu_2 = \mu_3$. $H_1 = \mu \neq \mu_1 \neq \mu_2 \neq \mu_3$
- ii. Complete the composite table by adding the squares.
- iii. Find the sum of squares.

$$(i) \sum X = 271 + 267 + 255 = 793$$

$$(ii) \sum X^2 = 4185 + 4091 + 3511 = 11787$$

- iv. Find the total sum of squares. It is given by:-

$$\begin{aligned} SSt &= \sum X^2 - \frac{(\sum X)^2}{N} = 11787 - \frac{(793)^2}{60} \\ &= 11787 - \frac{628849}{60} \\ &= 11787 - 10480.81667 = \underline{\underline{1306.18333}} \end{aligned}$$

- v. Find the sum of squares between

$$\begin{aligned} SSb &= \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{(\sum X)^2}{N} \\ &= \frac{271^2}{20} + \frac{267^2}{20} + \frac{255^2}{20} - \frac{793^2}{60} \end{aligned}$$

$$= 10487.75 - 10480.81667 = \underline{\underline{6.933}}$$

$$= 3672.05 + 3564.45 + 3251.25 - 10480.81667$$

$$= \bar{X} \quad 13.55 \quad 13.35 \quad 12.76$$

	X ₁	X ₂	X ₃	X ₁ ²	X ₂ ²	X ₃ ²
1	15	20	10	225	400	100
2	13	16	15	169	256	225
3	12	18	19	144	324	361
4	14	11	10	196	121	100
5	10	16	14	100	256	196
6	18	18	13	324	324	169
7	20	12	16	400	144	256
8	19	14	17	361	196	289
9	11	15	11	121	225	121
10	09	19	12	81	361	144
11	01	10	18	49	100	324
12	13	11	14	169	121	196
13	19	04	17	361	16	289
14	20	06	09	400	36	81
15	04	09	05	16	81	25
16	08	02	07	64	04	49
17	05	20	11	25	400	121
18	16	14	10	256	196	100
19	18	13	14	324	169	196
20	20	19	13	400	361	169
Σ	271	267	255	4985	4091	3511

vi. Find the sum of squares within

$$SS_w = \text{Total SS} - \text{Between SS} = SSt - SSb = 1306.18333 - 6.933 = 1299.25003 = \underline{\underline{1299.25}}$$

vii. Find the mean square between $MS_b = \frac{SS_b}{K - 1} = \frac{6.9333}{3 - 1} = \frac{6.9333}{2} = \underline{\underline{3.46665}}$

viii. Find the mean square within groups

$$MS_w = \frac{SS_w}{N - K} = \frac{1299.25}{60 - 3} = \frac{1299.25}{57} = 22.793859 = \underline{\underline{22.794}}$$

ix. Find the F-ratio $= \frac{MS_b}{MS_w} = \frac{3.46665}{22.795} = 0.152079403 = \underline{\underline{0.152}}$

x. It is a convention to show the result of the ANOVA in a summary table. This is done below:

Source of Variation	Sum of squares	df	Mean of square	F	P
Between Groups	6.933	2	3.467		
Within Groups	12.99.250	57	22.784	0.152	0.05
Total	1306.1833	59			

You will note that

- i. The sums of squares in the 2nd column are divided by the degrees of freedom df in the 3rd column to get the mean squares in the 4th column. To get the F-ratio, divide the mean square between groups by the mean square within groups.
- ii. The final step is to take decision. Do you remember that we said about F-ratio being less than 1. Now, our F-ratio of 0.152 is less than 1, so it is not significant, we do not need to go to the table. It means therefore that whenever the F-ratio is less than 1, we accept the null hypothesis and conclude that there are no significant differences between the means of the treatment groups.

So from the composite table, you see that means are different yet the differences are not statistically significant.

3.3 Another illustration

You have worked through the illustration in sub-unit 3.2 where there are three groups. Now let us use the result of a research in which four groups were involved.

GP ₁ =X ₁	10	8	7	6	9	5	10	9	8	6	4	3	2	7	5
GP ₂ =X ₃	8	9	10	10	7	8	9	5	9	4	3	8	5	6	3
GP ₃ =X ₃	2	8	9	10	10	7	5	6	7	8	10	5	7	3	9
Gp ₄ =X ₄	5	8	7	9	2	4	10	3	2	1	5	3	8	9	10

Solution:

- i. Propose a null hypothesis $H_o = \mu_1 = \mu_2 = \mu_3 = \mu_4$
- ii. Complete the composite table as shown.
- iii. Find the sum of squares:

$$(a) \sum X = 994104 + 116 + 86 = 405$$

$$(b) \Sigma X^2 = 739 + 804 + 836 + 632 = 3011$$

iv Find the total sum of squares

$$\begin{aligned} SSt &= \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 3011 - \frac{(405)^2}{60} = 3011 - \frac{164025}{60} \\ &= 3011 - 2733.75 = \underline{277.25} \end{aligned}$$

v. Find the sum of squares between groups

$$\begin{aligned} SSb &= \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{(\Sigma X)^2}{N} \\ &= \frac{99^2}{15} + \frac{104^2}{15} + \frac{116^2}{15} + \frac{86^2}{15} - \frac{405^2}{60} \\ &= 653.4 + 721.067 + 897.067 + 493.067 - 2733.75 = \underline{30.851} \end{aligned}$$

vi. Find the sum of squares within groups

$$SSw = SSt - SSb = 277.25 - 30.851 = \underline{246.399}$$

vii. Find the mean square between group $MSb = \frac{SSb}{K - 1} =$
 $\frac{30.851}{4 - 1} = \frac{30.851}{3} = \underline{10.283}$

viii. Find the mean square within group. $MSw = \frac{SSw}{N - K} =$
 $\frac{246.399}{60 - 4} = \frac{246.399}{56} = 4.399982143 = \underline{4.400}$

ix. Find the F-ratio $= \frac{MSb}{MSw} = \frac{10.283}{4.400} = \underline{2.337}$

S/N	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
ix 1	10	100	8	64	2	04	5	25
2	8	64	9	81	8	64	8	64
3	7	49	10	100	9	81	7	49
4	6	36	10	100	10	100	9	81
5	9	81	7	49	10	100	2	04
6	5	25	8	64	7	49	4	16
7	10	100	9	81	5	25	10	100
8	9	81	5	25	6	36	3	09
9	8	64	9	81	7	49	2	04
10	6	36	4	16	8	64	1	01
11	4	16	3	09	10	100	5	25
12	3	09	8	64	5	25	3	09
13	2	04	5	25	7	49	8	64
14	7	49	6	36	3	09	9	81
15	5	25	3	09	9	81	10	100
Σ	99	739	104	804	116	836	86	632

x. Present the result in a summary table

Source of variation	Sum of squares	d-f	Mean square	F	P
Between Groups	30.851	3	10.283		
Within Groups	246.399	56	4.400	2.337	0.05
Total	277.25	59			

Decision

For F_{tab} , go to the table for F-ratio, under $d.f(4-1) = 3$, go down this column until you get to the row entry of $(60 - 4) = 56$. That is, the point of intersection of 3 and 56. The top value is for 0.05 while the value under is for 0.01:

$$\square \quad F_{tab} \text{ at } (3 \ \& \ 56; 0.05) = 2.78$$

$$F_{cal} = 2.337$$

So $F_{cal} < F_{tab}$. We accept the null hypothesis that the means are not significantly different.

Activity 1

Given the following results, complete the table of ANOVA

Source of variation	Sum of squares	d-f	Mean square	F	P
Between Groups		2			0.05
Within Groups	3129.50				
Total	3480.12	41			

- What is the between groups sum of squares?
- What is the within groups degree of freedom?
- What is the mean square between groups?
- What is the F-ratio?
- Is the F significant at the 0.05?

Answer to Activity 1

Source of variation	Sum of squares	Df	Mean square	F	P
Between Groups	35-.62	2	175.31		
Within Groups	3129.50	39	80.24	2.18	0.05
Total	3480.12	41			

- $SS_b = 3480.12 - 3129.50 = 350.62$
- Within group d-f = $41 - 2 = 39$
- $MS_b = 350.62 \div 2 = 175.31$
- $MS_a = 3129.50 \div 39 = 80.24$
- F-ratio = $\frac{MS_b}{MS_w} = \frac{175.31}{80.24} = 2.184820538 = \underline{\underline{2.18}}$
- $F_{cal} = 2.18$

Since F_{cal} is less than F_{tab} , \therefore the F is not significant at 0.05.

4.0 CONCLUSION

You have seen the computational formula for getting the F-ratio. You have also seen the illustrations and calculations of the F-test. The onus is on you to apply them in your experimental researches.

5.0 SUMMARY

In this unit, you learnt the simplest method of calculating the F-test. This involves:

i. calculations of the total sum of squares using $SSt = \sum X^2 - \frac{(\sum X)^2}{N}$

ii. determination of sum of squares between using $SSb =$

$$\frac{\sum X_1^2}{n_1} + \frac{\sum X_2^2}{n_2} + \dots + \frac{\sum X_k^2}{n_k} - \frac{(\sum X)^2}{N}$$

iii. Determination of sum of squares within using $SSw = SSt - SSb$.

These have been illustrated for you to study. You also learnt how to find the F using $\frac{MSb}{MSw}$ and how to verify the result from the F table. This is one-way analysis of variance. In the next unit you will be studying the two-way analysis of variance.

6.0 TUTOR-MARKED ASSIGNMENT

Using the data below determine whether the means are significantly different.

X_1 10, 9, 8, 7, 12, 10, 11, 12, 9, 8, 5, 4, 6, 3, 7.

X_2 4, 6, 4, 8, 6, 4, 5, 7, 10, 11, 9, 3, 2, 6, 5.

X_3 5, 2, 8, 7, 10, 2, 1, 4, 3, 4, 8, 5, 6, 9, 12.

7.0 REFERENCES/FURTHER READINGS

Ali, A (1996) Fundamentals of Research in Education. Awka, Meks Publishers (Nig).

Ary, D and Jacobs, L.C. (1976) Introduction to Statistics Purposes and Procedures. New York...Sydney Holt, Rinehart and Winston.

Ogomaka, P.M.C. (2004) Inferential Statistics for Research in Education and Social Sciences. Owerri, Peacewise Systems and Prints

UNIT 3 TWO-WAY ANALYSIS OF VARIANCE

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	The Two Way Analysis of Variance
3.2	Factorial ANOVA
3.3	Illustrations of Two-way ANOVA
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment
7.0	References/Further Readings

1.0 INTRODUCTION

You have studied the one-way analysis of variance and its applications and uses in the last two units. You learnt that analysis of variance (one-way) is a widely used statistical test used with two or more numbers of groups. It is used mainly to investigate the effect of one independent variable on a dependent variable. But you know that in nature more than one independent variables may affect the dependent variable. In such research studies, where multiple effects of variables on the dependent variable are studied may require a more complex analytical tool for the treatment of the data obtained from such studies. In this unit, therefore, you will be learning the Factorial ANOVA and the computations involved.

2.0 OBJECTIVES

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

- i. Explain the two-way analysis of variance
- ii. Identify situations in which a two-way ANOVA is used
- iii. Calculate F-ratios in a two-way ANOVA problem.

3.0 MAIN CONTENT

3.1 The two-way Analysis of Variance

This is a technique employed to test the presence or otherwise of some treatment and interactions effects in experiments where two treatment variables are involved. In other words, each subject in the study is exposed to one of the levels of one of the variances, and one level of the other. There are some true experimental designs which involve the use of more than one level of independent variable and to consequently investigate their respective effects on the dependent variables. In any research, where the investigation of the multiple effects of variables on the dependent variables are involved require a more complex analytical statistical tool for treating the resulting data from such studies. You have to note that any experimental design study where the investigation of two or more independent variables are involved is called a factorial design and the data obtained through the use of such design and the data obtained through the use of such design are treated by means of Multifactor Analysis of Variance, otherwise called MANOVA. Now let us give more explanations in the next sub-unit.

3.2 Factorial ANOVA

You have now known that one-way ANOVA is used to find out the effect of one independent variable on a dependent variable. But you know that the effect of a single independent variable alone cannot be the same with the effect when interacting with another independent variable. Let us consider one example. You know that the effectiveness of a given teaching methodology can depend on a number of other variables. These variables which can contribute to the effectiveness include motivation level, sex of the students, ability level, the size of the class, teaching materials or aids and even the personality of the teacher. This implies that two independent variables in combination may have an effect which may not be accounted for by the effect of the two independent variables taken separately. When two independent variables are combined, they produce an effect which is called INTERACTION. Some of the times two or more independent variables can be manipulated at the same time so as to determine (a) their independent effects on the dependent variable and (b) the interaction effects of the variables in combination. There are so many advantages of these combinations. These include:

- (i) In one experiment, you can investigate independent variables which would ordinarily require two or more separate studies. You would also get the measure of the interaction of these variables which would not be possible if the analysis is done in separate experiments
- (ii) In any experimental design with two independent variables, these may be manipulated in the experiment or only one may be manipulated and the other used as a control variable. Take for instance, where such variables like sex, age, home background, intelligence, family status, geographical location, experience etc. can influence the independent variable, they can be incorporated in the study as an independent variable thereby serving as a control measure. This control function is an advantage.

In an experimental design of this nature, the independent variables are called factors. The statistical method used for the analysis of the independent and interaction effects of these independent variables or factors is called Factorial Analysis of Variance or Multifactor Analysis of Variance. In more complex designs having more than two independent variables or factors where each variable has two or more possible values, such possible value of the independent variable is called a level. Take the case of three methods of teaching and two learning situations. It means that we have 3 levels for the first factor and 2 levels for the second. That means 3×2 designs. Where the subjects are further classified into male and female, then we have $3 \times 2 \times 2$. Now, let us go to the computations and illustrations.

3.3 Illustrations of two-way ANOVA

A teacher used four groups of randomly selected samples in his experimental study. Group one was taught Technical Drawing with project method and class evaluation, Group two was project method and group evaluation, Group three has talk and chalk method with class evaluation while Group four has talk and chalk method with group evaluation. The following results were obtained:

	Class Evaluation	Group Evaluation	
Project Method	I 12 X ₁ 10 11 12 9 10 8 Σ 72	2 10 11 9 X ² 7 5 12 8 Σ 62	$\Sigma X_1 = 72, \bar{X}_1 = 10.286, n_1 = 7$ $\Sigma X_2 = 62, \bar{X}_2 = 8.857, n_2 = 7$ $\Sigma X_3 = 51, \bar{X}_3 = 7.286, n_3 = 7$ $\Sigma X_4 = 60, \bar{X}_4 = 8.571, n_4 = 7$ $\Sigma Xr_1 = 72 + 62 = 134$
Talk and Chalk Method	3 4 6 5 X ₃ 7 8 10 11 Σ 51	4 10 11 9 X ₄ 10 8 7 5 Σ 60	$\bar{X}r_1 = 134 \div 14 = 9.571$ $\Sigma Xr_2 = 51 + 60 = 111$ $\bar{X}r_2 = 111 \div 14 = 7.929$ $\bar{X}c_1 = 72 + 51 = 123$ $\bar{X}c_1 = 123 \div 14 = 8.786$
			$\bar{X}c_2 = 62 + 60 = 122$ $\bar{X}c_2 = 122 \div 14 = 8.714$ $\Sigma XTotal = 72 + 62 + 51 + 60 = 245$ $\bar{X}(grandmean) = 8.75$

After getting the totals and the means shown above, the next step is to get the sum of squares as in ANOVA.

X_1	X_1^2	X_2	X_2^2	
12	144	10	100	$\Sigma X^2 = 754 + 584 + 411 + 540 = \underline{2289}$ $\frac{(\Sigma X)^2}{N} = \frac{245^2}{28} = \frac{60025}{28} = \underline{2143.75}$
10	100	11	121	
11	121	9	81	
12	144	7	49	
9	81	5	25	
10	100	12	144	
<u>8</u>	<u>64</u>	<u>8</u>	<u>64</u>	
$\Sigma X_1 = 72$	$\Sigma X_1^2 = 754$	$\Sigma X_2 = 62$	$\Sigma X_2^2 = 584$	
X_3	X_3^2	X_4	X_4^2	
4	16	10	100	
6	36	11	121	
5	25	9	81	
7	49	10	100	
8	64	8	64	
10	100	7	49	
<u>11</u>	<u>121</u>	<u>5</u>	<u>25</u>	
$\Sigma X_3 = 51$	$\Sigma X_3^2 = 411$	$\Sigma X_4 = 60$	$\Sigma X_4^2 = 540$	

$$\text{ind SS}_t = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 2289 - 2143.75 = \underline{145.25}$$

$$\begin{aligned} \text{Find SS}_b &= \frac{(\Sigma X)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{(\Sigma X)^2}{N} \\ &= \frac{72^2}{7} + \frac{62^2}{7} + \frac{51^2}{7} + \frac{60^2}{7} - \frac{245^2}{28} \\ &= 740.571 + 549.143 + 371.571 + 514.286 - 2143.75 = \underline{31.821} \end{aligned}$$

$$\text{Find SS}_w = \text{SS}_w = \text{SS}_t - \text{SS}_b = 145.25 - 31.821 = \underline{113.429}$$

Find SS_{bc} = sum of squares between columns

$$= \frac{(\Sigma X_{c1})^2}{n_{c1}} + \frac{(\Sigma X_{c2})^2}{n_{c2}} - \frac{(\Sigma X)^2}{N} = \frac{123^2}{14} + \frac{122^2}{14} - \frac{245^2}{28}$$

$$= 1080.643 + 1063.143 - 2143.75 = \underline{0.036}$$

Find SSbr = sum of squares between rows

$$\begin{aligned} &= \frac{(\sum X_{r1})^2}{n_{r1}} + \frac{(\sum X_{r2})^2}{n_{r2}} - \frac{(\sum X)^2}{N} = \frac{134^2}{14} + \frac{111^2}{14} - \frac{245^2}{28} \\ &= 1282.571 + 880.071 - 2143.75 = \underline{18.892} \end{aligned}$$

Find SSrc = Interaction sum of squares

$$\begin{aligned} &= \text{SSrc} = \text{SSb} - (\text{SSbc} + \text{SSbr}) = 31.821 - (0.036 + \\ &18.892) \\ &= 31.821 - 18.928 = \underline{12.893} \end{aligned}$$

Find the degrees of freedom

$$\text{df for between columns} = c - 1 = 2 - 1 = 1$$

$$\text{df for between rows sum of squares} = 2 - 1 = 2 - 1 = 1$$

$$\text{df for interaction} = (c - 1)(2 - 1) = (2 - 1)(2 - 1) = 1$$

$$\text{df for between groups sum of squares} = (K - 1) = 4 - 1 = 3$$

$$\text{df for within groups sum of squares} = \sum(n - 1) = (7 - 1) + (7 - 1) + (7 - 1)$$

$$\text{df for total sum of squares} = N - 1 = 28 - 1 = \underline{27}$$

Find the mean squares. Each of the three sums of squares is divided by the correspondence df to get the MS.

$$\text{i. MS}_c = \text{Mean square for columns} = \frac{\text{SSbc}}{\text{df}} = \frac{0.036}{1} = \underline{0.036}$$

$$\text{ii. MS}_r = \text{Mean square for rows} = \frac{\text{SSbr}}{\text{df}} = \frac{18.892}{1} = \underline{18.892}$$

$$\text{iii. MS}_{rc} = \text{Mean square for interaction} = \frac{\text{SSrc}}{\text{df}} = \frac{12.893}{1} = \underline{12.893}$$

$$\begin{aligned} \text{iv. MS}_w &= \text{Mean square within} \\ \frac{\text{SS}_w}{N - K} &= \frac{113.429}{28 - 4} = \frac{113.429}{24} = \underline{4.726} \end{aligned} =$$

Source of Variance	Sum of sq	df	Mean sq	F	P>0.05
Between columns	0.036	1	0.036	0.008	
Between rows	18.892	1	18.892	3.997	
Columns by rows(Inter)	12.893	1	12.893	2.728	
Within groups	113.429	24	4.726		
Total	145.25	27			

NB: For F-ratio, we use:

- i. For columns = $\frac{MSc}{MSw} = \frac{0.036}{4.726} = \underline{\underline{0.008}}$
- ii. For rows = $\frac{MSr}{MSw} = \frac{18.892}{4.726} = \underline{\underline{3.997}}$
- iii. For interaction = $\frac{MSrc}{MSw} = \frac{12.893}{4.726} = \underline{\underline{2.728}}$

For Decision: Look for the F-table under 1 and 24 Ftab at (1,24;0.05)= 4.49

It means that only at 4.49 and above can any of the F-ratios be significant. None of these is significant.

Activity 1

Given the following data, complete the analysis of variance and set up the summary table. Test the null hypothesis at 0.05.

	A ₁	A ₂	A ₁	A ₂
	10	5	15	8
	9	4	14	6
	5	5	13	7
B ₁	6	5	B ₂ 12	5
	7	4	11	9
	8	3	10	4
	10	2	9	3
	10	4	8	9
	9	3	7	8
	8	2	6	6

Answer to Activity 1

	A_1	A_1^2	A_2	A_2^2	
B₁	10	100	5	25	$\Sigma A_1 = 82, \overline{A_1} = 8.2, n_1 = 10$
	9	81	4	16	$\Sigma A_2 = 37, \overline{A_2} = 3.7, n_2 = 10$
	5	25	5	25	$\Sigma A_3 = 105, \overline{A_3} = 10.5, n_3 = 10$
	6	36	5	25	$\Sigma A_4 = 68, \overline{A_4} = 6.8, n_4 = 10$
	7	49	4	16	$\Sigma A_{r1} = 82, = 82 + 37 = 119$
	8	64	3	09	$\Sigma A_{r2} = 82, = 105 + 68 = 173$
	10	100	2	04	$\Sigma A_{c1} = 82, = 82 + 105 = 187$
	10	100	4	16	$\Sigma A_{c2} = 37 + 68 = 105$
	9	81	3	09	
	<u>8</u>	<u>64</u>	<u>2</u>	<u>04</u>	
	$\Sigma 82$	700	37	149	
A₃	A_3^3	A₄	A_4^4		
15	225	8	64	$\Sigma A_{total} = 82 + 37 + 105 + 68 = 292$	
14	196	6	36	$\Sigma A^2 = 700 + 149 + 1185 + 461 = 2495$	
13	169	7	49	$\frac{(\Sigma A)^2}{N} = \frac{292^2}{40} = \frac{85264}{40} = \underline{\underline{2131.6}}$	
12	144	5	25		
11	121	9	81		
10	100	4	16		
9	81	3	09		
8	64	9	81		
7	49	8	64		
<u>6</u>	<u>36</u>	<u>6</u>	<u>36</u>		
105	1185	68	461		

$$\text{Find SS}_t = \Sigma A^2 - \frac{(\Sigma A)^2}{N} = 2495 - 2131.6 = \underline{\underline{363.4}}$$

$$\text{Find SS}_b = \frac{(\Sigma A_1)^2}{n_1} + \frac{(\Sigma A_2)^2}{n_2} + \frac{(\Sigma A_3)^2}{n_3} + \frac{(\Sigma A_4)^2}{n_4} - \frac{(\Sigma A)^2}{N}$$

$$\frac{82^2}{10} + \frac{37^2}{10} + \frac{105^2}{10} + \frac{68^2}{10} - \frac{292^2}{40} = 672.4 + 136.9 + 1102.5 + 462.4 - 2131.6 = \underline{\underline{242.6}}$$

$$\text{Find SS}_w = \text{SS}_t - \text{SS}_b = 363.4 - 242.6 = \underline{\underline{120.8}}$$

$$\text{Find SS}_{bc} = \frac{(\Sigma A_{c1})^2}{n_{c1}} + \frac{(\Sigma A_{c2})^2}{n_{c2}} - \frac{(\Sigma A)^2}{N} = \frac{187^2}{20} + \frac{105^2}{20} - \frac{292^2}{40}$$

$$= 1748.45 + 551.25 - 2131.6 = \underline{\underline{168.1}}$$

$$\begin{aligned} \text{Find SSbr} &= \frac{(\sum A_{r1})^2}{n_{r1}} + \frac{\sum A_{r2}}{n_{r2}} - \frac{(\sum A)^2}{N} = \frac{119^2}{20} + \frac{173^2}{20} - \frac{292^2}{40} \\ &= 708.05 + 1496.45 - 2131.6 = \underline{\underline{72.9}} \end{aligned}$$

$$\text{Find SSrc} = SSb - (SSbc + SSbr) = 242.6 - (168.1 + 72.9) = 242.6 - 241 = \underline{\underline{1.6}}$$

Find the degrees of freedom

df for between columns = $2 - 1 = 1$. df for between rows = $2 - 1 = 1$

df for interaction = $(2-1)(2-1) = 1$. df for between groups = $(4-1) = 3$

df for within groups sum of squares =

$$\sum(n-1) = (10-1) + (10-1) + (10-1) + (10-1) = \underline{\underline{36}}$$

$$\text{df for total sum of squares} = N - 1 = 40 - 1 = \underline{\underline{39}}$$

Find the Mean Squares:-

$$\text{i. } MSc = \frac{SSbc}{df} = \frac{168.1}{1} = \underline{\underline{168.1}}$$

$$\text{ii. } MSr = \frac{SSbr}{df} = \frac{72.9}{1} = \underline{\underline{72.9}}$$

$$\text{iii. } MSw = \frac{SSw}{N - K} = \frac{120.8}{40 - 4} = \frac{120.8}{36} = \underline{\underline{3.36}}$$

$$\text{iv. } MSrc = \frac{SSrt}{df} = \frac{1.6}{1} = \underline{\underline{1.6}}$$

Find F-ratios

$$\text{i. } \text{For rows} = \frac{MSr}{MSw} = \frac{72.9}{3.36} = \underline{\underline{21.70}}$$

$$\text{ii. } \text{For columns} = \frac{MSc}{MSw} = \frac{168.1}{3.36} = \underline{\underline{50.03}}$$

$$\text{iii. } \text{For interactions} = \frac{MSrc}{MSw} = \frac{1.6}{3.36} = \underline{\underline{0.476}}$$

Sources of variation	Sum of squares	df	Mean Square	F	P>0.05
Between columns	168.1	1	168.1	50.03	
Between rows	72.9	1	72.9	21.70	
Columns b rows(Interaction)	1.6	1	1.6	0.476	
Within groups	120.8	36	3.36		
Total	363.4	39			

Decision

$F_{cal} = 50.03, 21.70$ and 0.476

F_{tab} at $(1.36 ; 0.05) = 4.11$

This means that the F-ratios for the columns and the rows are significant. That means that there is no significant difference in their means. But the F-ratio for interaction is not significant. That means that for the interaction we accept that there is significant difference.

4.0 CONCLUSION

The two-way analysis of variance, popularly called two-way ANOVA is an extension of the simple or one-way ANOVA which you studied in Units 14 and 15. In an experimental situation, the two-way ANOVA involves two or more independent variables manipulated simultaneously. This is a technique which enables you to determine the main effects of interaction between the variables. In a single experiment therefore, you can accomplish what would ordinarily require two separate studies, and in addition study the effect of the variables in combination. In this case, you may get some information that may be of practical and theoretical importance to humanity especially when it is concerned with research dealing with human beings. You have now seen that it is not always good to restrict studies to one variable which leads to the over simplification of a complex situation. This is because the effect of a variable may depend on the presence of another variable or variables. It is for this reason that factorial designs involve two or more factors which are very appropriate for research in education and the social sciences. This development and use of multi-factor analysis of variance has brought tremendous advances in educational research. You may be required to use it in your dissertation. If so you need to study it very well.

5.0 SUMMARY

In this unit, you have worked through the two-way analysis of variance in which you have noted can be used in an experiment which involves the

manipulation of two or more independent variables for their effect on the dependent variable. You have seen that the computation of this factorial or multi-factor ANOVA is similar to that of one-way ANOVA except for the fact that there are more sources of variations to be considered. You have seen how to compute the total sum of squares SS. This sum of squares SS is divided or partitioned into four parts. These are: SS for each of the independent variables; sum of squares for interaction and sum of squares within. You have also seen that three F-ratios are obtained by comparing the mean square for the variable X, variable Y and interaction between X and Y, to the within mean square in each case. These F-ratios will test the significance of the two main effects and the interaction effect.

6.0 TUTOR-MARKED ASSIGNMENT

The data below show the results of an experiment conducted by a researcher. Compute the analysis of variance and set up the summary table. Test the significance at 0.05.

	A	B
X ₁	10, 9, 8, 7, 12, 10, 11, 12	9, 8, 13, 10, 12, 9, 10, 11
X ₂	5, 2, 8, 7, 10, 2, 1, 4	3, 4, 10, 7, 6, 8, 5, 6
X ₃	4, 6, 4, 8, 6, 4, 5, 7	10, 11, 9, 10, 8, 7, 9, 6

7.0 REFERENCES/FURTHER READINGS

- Ali, R. (1996). *Fundamentals of Research in Education*. Awka Meks Publishers (Nig)
- Arg, D and Jacobs, L.C. (1976) *Introduction to Statistics: Purposes and Procedures*. New York Sydney. Holt, Rinehart & Winston
- Ogomaka, P.M.C. (2004) *Inferential Statistics for Research in Education and Social Sciences*. Owerri, Peacewise Systems & Prints.

UNIT 4 ANALYSIS OF COVARIANCE (ANCOVA)

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	The Analysis of Covariance (ANCOVA)
3.2	Uses of ANCOVA
3.3	Illustration of ANCOVA
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment (TMAs)
7.0	References/Further Readings

1.0 INTRODUCTION

In the last three units, you have worked through the Analysis of Variance and the Multifactor Analysis of variance (ANCOVA and MANOVA). You have learnt that these ANOVA tests are used for parametrically comparing the means of two or more groups, events, or observations investigated through experimental design procedures. You will have to note that, traditionally and in most of the studies in the social sciences and Education, research studies have restrictions imposed on them. These restrictions can be as a result of the sampling of subjects, pretesting subjects and manipulation of research conditions or no choice at all in working with the subjects in his own direction or the direction of his research design. The researcher may therefore have to depend on a number of measures which the net effect will be the imposition of a far more complex method of analysing data. This is because of the pre-test, post test design and the realities of sampling. One of the appropriate measures to be used is the Analysis of Covariance (ANCOVA). In this unit, we shall be discussing the Analysis of covariance and its uses and applications. But before we continue, let us define what you should be able to achieve at the end of the unit.

2.0 OBJECTIVES

After carefully working through this unit, you should be able to:

- i) Explain the term analysis of covariance.
- ii) Describe the uses of ANCOVA.

3.0 MAIN CONTENT

3.1 Analysis of Covariance (ANCOVA)

You have learnt that analysis of variance is mainly used for comparing the means of two or more groups which are investigated by the experimental procedures. Elsewhere, you will read about the regression analysis which is a method used for making predictions of outcomes based on survey or observational data. We can say then that analysis of covariance (ANCOVA) is a statistical tool or technique which is basically a combination of the linear regression and analysis of variance. In fact, it is a more complex extension of both. The application of ANCOVA is a very complex procedure, but we shall try to present it simplified and non-technical, to enable you understand it well. We shall now discuss it based on situations drawn from observational and experimental data. In other words, we shall use situations which are very familiar with you.

ANCOVA generally involves three sets of variables or factors. These are:

- (i) The qualitative experimental variables called the independent variables;
- (ii) The inbuilt, otherwise called inherent or incidental quantitative variables which are not usually manipulated like the independent variables, are referred to as the concomitant variables, and
- (iii) The quantitative dependent variables.

In nature, most of the times it is not easy to use treatment and control groups that are equivalent or approximately equal in terms of some concomitant variables. You remember that we have said that a concomitant variable is that which influences or determines to a large extent the values of the dependent variables whether the equalization is done by randomization or not. As a result, a researcher may decide to use the groups as they occur, naturally, situationally or circumstantially. In other words, a researcher can decide to use an intact group without any deliberate effort to equalize them. In this case, ANCOVA is the appropriate test that can be used to determine whether there is any significant difference among the means of the groups. Now let us look at the uses of ANCOVA.

3.2 Uses of ANCOVA

You have learnt that the ANCOVA is a statistical tool used for analysing differences between experimental treatment and control groups on the dependent variable based on pretest – post test design and under a situation

where subjects were selected or used as intact groups. In this situation, ANCOVA serves two broad purposes:

It is used as a technique for controlling extraneous variables and contaminations as well as a means of increasing the power of the analysis done using it. For example, if you want to conduct a study where it may not be possible for you to do any randomization of the subjects or have the research conditions and assumptions in place, and again, if the subjects are pretested, you do not need to go to find out how equivalent the subjects are. ANCOVA does that very well for you. It removes the initial differences between groups so that the selected or pretested groups can be correctly considered as equated or equivalent. ANCOVA does this by removing score differences in the pretest performance across groups. The scores so corrected by this method is called residuals or adjusted scores. ANCOVA helps us to find out the significance of the difference between the pretest and post test scores called covariates. Let us summarise the major uses below.

- i) It increases precision of data obtained from an experimental study;
- ii) It removes bias which may result from using intact groups whose equivalence on certain measures have not been determined;
- iii) It removes the effects of intervening variables or stabilizes independent variables to the point that their effects have not been unduly influenced by intervening variables;
- iv) It allows for certain trends to be observable from descriptive data so analysed; and
- v) It increases the power of statistical test merely by reducing within group variance error.

3.3 Illustration of ANCOVA

Let us use the experimental study of s researcher who compared three methods of evaluation of students learning outcomes, to illustrate the application of ANCOVA. The pretest, post test results of three groups of students are given below:

G1	X ₁	3, 9, 7, 9, 8, 5, 3, 6, 7, 8
	Y ₁	17, 16, 18, 9, 20, 12, 15, 14, 13, 19
G2	X ₂	8, 5, 7, 8, 6, 9, 4, 9, 3, 6
	Y ₂	14, 11, 9, 15, 12, 10, 14, 13, 7, 12
G3	X ₃	7, 4, 4, 9, 8, 5, 9, 6, 3, 8
	Y ₃	13, 10, 6, 12, 8, 11, 14, 9, 10, 10

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_1Y_1	X_2Y_2	X_3Y_3	X_1^2	Y_1^2	X_2^2	Y_2^2	X_3^2	Y_3^2
3	17	8	14	7	13	51	112	91	9	289	64	196	49	169
9	16	5	11	4	10	144	55	40	81	256	25	121	16	100
7	18	7	9	4	6	126	63	24	49	324	49	81	16	36
9	9	8	15	9	12	81	120	108	81	81	64	225	81	144
8	20	6	12	8	8	160	72	64	64	400	36	144	64	64
5	12	9	10	5	11	60	90	55	25	144	81	100	25	121
3	15	4	14	9	14	45	56	126	9	225	16	196	81	196
6	14	9	13	6	9	84	117	54	36	196	81	169	36	81
7	13	3	7	3	10	91	21	30	49	169	9	49	9	100
8	19	6	12	8	10	152	72	80	64	361	36	144	64	100
65	153	65	117	63	103	994	778	672	467	2445	461	1425	441	1111

In order to find the F – ration here, we will start by forming a composite table.

- Calculate total sum of squares for X, $SSX_t = \sum \sum X^2 - \frac{(\sum \sum X)^2}{N}$

$$= \frac{(467 + 461 + 441) - (65 + 65 + 63)^2}{60}$$

$$= \frac{1369 - \frac{37249}{60}}{60} = \frac{1369 - 620.82}{60} = \underline{\underline{748.18}}$$
- Calculate sum of squares between for X, $SSX_b = \sum \frac{(\sum X)^2}{n} - \frac{(\sum \sum X)^2}{N}$

$$= \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{(\sum \sum X)^2}{N}$$

$$= \frac{(65)^2}{10} + \frac{(65)^2}{10} + \frac{(63)^2}{10} - \frac{(193)^2}{60}$$

$$= (422.5 + 422.5 + 196.9) - (620.82) = 1041.9 - 620.82$$

$$= \underline{\underline{421.08}}$$
- Calculate sum of squares within for X = $SSX_t - SSX_b$

$$= 748.18 - 421.08 = \underline{\underline{327.10}}$$
- Calculate sum of squares total for Y, $SSY_t = \sum \sum Y^2 - \frac{(\sum \sum Y)^2}{N}$

$$= \frac{(2445 + 1425 + 1111) - (153 + 117 + 103)^2}{60}$$

$$= 4981 - 2318.8167$$

$$= \underline{\underline{2662.1833}}$$

5. Calculate sum of squares between for Y = $SSY_b = \sum(\sum Y)^2 - \frac{(\sum \sum Y)^2}{N}$

$$\begin{aligned} &= \frac{(\sum Y_1)^2}{n_1} + \frac{(\sum Y_2)^2}{n_2} + \frac{(\sum Y_3)^2}{n_3} - \frac{(\sum \sum Y)^2}{N} \\ &= \frac{(153)^2}{10} + \frac{(117)^2}{10} + \frac{(103)^2}{60} - \frac{(193)^2}{60} \\ &= (2340.9 + 1368.9 + 1060.9) - (2318.8167) \\ &= 4770.7 - 2318.8167 \\ &= \underline{\underline{2451.8833}} \end{aligned}$$

6. Calculate sum of squares within for Y = $SSY_t = SSY_t - SSY_b$

$$= 2662.1833 - 2451.8833 = \underline{\underline{210.30}}$$

7. Calculate sum of squares total for the cross product of X and Y

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

$$\begin{aligned} &= \frac{(994 + 778 + 672) - (193 \times 373)}{60} \\ &= \frac{2444 - 71989}{60} = \frac{2444 - 1199.8167}{60} \\ &= \underline{\underline{1244.1833}} \end{aligned}$$

8. Calculate sum of squares for X and Y between = $SSPXY_b$

$$\begin{aligned} &= \frac{(\sum X)(\sum Y) - (\sum \sum X)(\sum \sum Y)}{n} \\ &= \frac{(65 \times 153) + (65 \times 117) + (63 \times 103) - (193 \times 373)}{10} \\ &= 994.5 \times 760.5 + 648.9 - 1199.8167 \\ &= 2403.9 - 1199.8167 \\ &= \underline{\underline{1204.08}} \end{aligned}$$

9. Calculate sum of squares for X and Y within, $SSPXY_w$

$$\begin{aligned} &= SSPXY_b = SSPXY_t - SSPXY_b \\ &= 1244.1833 - 1199.8167 \\ &= \underline{\underline{44.3666}} \end{aligned}$$

10. Calculate sum of squares adjusted mean SSY_1

$$\begin{aligned} &= SSY_b + \frac{(SSPXY_w)^2}{SSX_w} - \frac{(SSPXY_t)^2}{SSX_t} \\ &= 2451.8833 + \frac{(44.3666)^2}{327.10} - \frac{(1244.1835)^2}{748.18} \end{aligned}$$

$$= 2457.901 - 2069.0102$$

$$= \underline{\underline{308.89}}$$

11. Calculate sum of squares error SSY_1 error.

$$= \frac{SSY_w - (SSPXY_w)^2}{SSX_w} = \frac{210.30 - (44.3666)^2}{327.10}$$

$$= \frac{210.30 - 6.018}{327.10}$$

$$= \underline{\underline{204.282}}$$

$$\text{Mean sum of square (1) for adjusted mean} = \frac{SSY_b^1}{df} = \frac{308.89}{2}$$

$$= \underline{\underline{154.445}}$$

$$(2) \text{ for error} = \frac{SSY^1 \text{ error}}{df}$$

$$= \frac{204.282}{58}$$

$$= \underline{\underline{3.522}}$$

$$F - \text{ratio} = \frac{\text{mean sum of square, adjusted mean}}{\text{mean sum of square error}} = \frac{154.445}{3.522}$$

$$= \underline{\underline{43.85}}$$

Now let us put the result in a summary table below

Square of Variation	Sum of Squares	df	Mean Squares	F	P
Adjusted Means	308.89	2	154.445	43.85	0.05
Error	204.282	58	3.522		
Total		60			

Decision:

$$F_{tab} = F(2:58, 0.05) = 3.17$$

$$F_{cal} = 43.85$$

$F_{cal} > F_{tab}$ we reject the null hypothesis.

Activity 1

Calculate the F – ratio using the data below:

X_1	3, 9, 7, 9, 8, 3
Y_1	10, 12, 9, 14, 10, 8
X_2	4, 3, 5, 6, 3, 2
Y_2	8, 7, 9, 10, 6, 5
X_3	2, 3, 3, 5, 2, 4

Y_3	9, 8, 7, 10, 6, 5
-------	-------------------

Answer to Activity 1

S/N	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_1Y_1	X_2Y_2	X_3Y_3	X_1^2	Y_1^2	X_2^2	Y_2^2	X_3^2
1	3	10	4	8	2	9	30	32	18	9	100	16	64	4
2	9	12	3	7	3	8	108	21	24	81	144	9	49	9
3	7	9	5	9	3	7	63	45	21	49	81	25	81	9
4	9	14	6	10	5	10	126	6	50	81	196	36	100	25
5	8	1	3	6	2	6	8	18	12	64	100	9	36	4
6	3	8	2	5	4	5	24	10	20	9	64	4	25	16
Σ	39	63	23	45	19	45	431	186	145	293	685	99	355	67

- Calculate total sum of squares for X, $SSX_t = \sum \sum X^2 - \frac{(\sum \sum X)^2}{N}$

$$= \frac{(293 + 99 + 67) - (39 + 23 + 19)^2}{36}$$

$$= 459 - 182.25 = \underline{\underline{276.75}}$$
- Calculate sum of squares between for X, $SSX_b = \sum \frac{(\sum X)^2}{n} - \frac{(\sum \sum X)^2}{N}$

$$= \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{(\sum \sum X)^2}{N}$$

$$= \frac{(39)^2}{6} + \frac{(23)^2}{6} + \frac{(19)^2}{6} - \frac{(81)^2}{36}$$

$$= (253.5 + 88.17 + 60.17) - (182.25) = \underline{\underline{219.59}}$$
- Calculate sum of squares within for X = $SSX_t - SSX_b$

$$= 276.75 - 219.59 = \underline{\underline{57.16}}$$
- Calculate sum of squares total for Y, $SSY_t = \sum \sum Y^2 - \frac{(\sum \sum Y)^2}{N}$

$$= (685 + 355 + 355) - \frac{(63 + 45 + 45)^2}{36}$$

$$= 1395 - 650.25$$

$$= \underline{\underline{744.75}}$$
- Calculate sum of squares between for Y = $SSY_b = \sum \frac{(\sum Y)^2}{n} - \frac{(\sum \sum Y)^2}{N}$

$$= \frac{(\sum Y_1)^2}{n_1} + \frac{(\sum Y_2)^2}{n_2} + \frac{(\sum Y_3)^2}{n_3} - \frac{(\sum \sum Y)^2}{N}$$

$$\begin{aligned}
&= \frac{(63)^2}{6} + \frac{(45)^2}{6} + \frac{(45)^2}{6} - \frac{(153)^2}{36} \\
&= (661.5 + 337.5 + 337.5) - (650.25) \\
&= 1336.5 - 650.25 = \mathbf{686.25}
\end{aligned}$$

6. Calculate sum of squares within for Y = $SSY_t = SSY_t - SSY_b$

$$= 744.75 - 686.25 = \mathbf{58.50}$$

7. Calculate sum of squares total for the cross product of X and Y

$$\begin{aligned}
&= SSPXY_t = \frac{\sum \sum XY - (\sum \sum X^2)(\sum \sum Y)}{N} \\
&= \frac{(431 + 186 + 145) - (81 \times 153)}{36} \\
&= 762 - 344.25 = \mathbf{417.75}
\end{aligned}$$

8. Calculate sum of squares for X and Y between = $SSPXY_b$

$$\begin{aligned}
&= \frac{(\sum X)(\sum Y) - (\sum \sum X)(\sum \sum Y)}{n \quad N} \\
&= \frac{(39 \times 63) + (23 \times 45) + (19 \times 45) - (81 \times 153)}{6 \quad 6 \quad 6 \quad 36} \\
&= 409.5 \times 172.5 + 142.5 - 344.25 \\
&= \mathbf{300.25}
\end{aligned}$$

9. Calculate sum of squares for X and Y within, $SSPXY_w$

$$\begin{aligned}
&= SSPXY_b = SSPXY_t - SSPXY_b \\
&= 417.75 - 300.25 \\
&= \mathbf{117.50}
\end{aligned}$$

10. Calculate sum of squares adjusted mean SSY_1

$$\begin{aligned}
&= SSY_b + \frac{(SSPXY_w)^2}{SSX_w} - \frac{(SSPXY_t)^2}{SSX_t} \\
&= 686.25 + \frac{(117.50)^2}{57.16} - \frac{(417.75)^2}{276.75} \\
&= 927.79 - 630.59 \\
&= \mathbf{297.20}
\end{aligned}$$

11. Calculate sum of squares error SSY_1 error.

$$\begin{aligned}
&= SSY_w - \frac{(SSPXY_w)^2}{SSX_w} = 58.50 - \frac{(117.50)^2}{57.16} \\
&= 58.50 - 241.54 = \mathbf{183.04}
\end{aligned}$$

12. Calculate Mean sum of squares (1) for adjusted Mean
(2) for error = SSY_b^1

$$\begin{aligned}
 & \text{df} \\
 \text{Mean sum of square (1) for adjusted mean} &= \frac{SSY^1_b}{df} = \frac{297.20}{2} \\
 &= \mathbf{148.60} \\
 \text{(2) for error} &= \frac{SSY^1_{\text{error}}}{df} = \frac{183.04}{34} \\
 \mathbf{ABSOLUTE VALUE} &= \mathbf{5.38}
 \end{aligned}$$

13. Calculate F – ratio

$$\begin{aligned}
 \text{F – ratio} &= \frac{\text{mean sum of square, adjusted mean}}{\text{mean sum of square error}} = \frac{148.60}{5.38} \\
 &= \mathbf{27.62}
 \end{aligned}$$

14. Draw the summary table for the results.

Square of Variation	Sum of Squares	df	Mean Squares	F	P
Adjusted Means	297.20	2	148.60	27.62	0.05
Error	-183.04	34	5.38		
Total		36			

Decision:

$$\begin{aligned}
 F_{\text{tab}} &= F(2:34, 0.05) = 3.28 \\
 F_{\text{cal}} &= 27.62 \\
 F_{\text{cal}} &> F_{\text{tab}} \text{ we reject the null hypothesis.}
 \end{aligned}$$

It implies that the F – ratio is significant from 3.28

4.0 CONCLUSION

You have now seen that when you have the difficulty of using intact group instead of subjects selected through random sampling or randomization, you can rely on the use of ANCOVA to solve the problem and increase the power of your test. Although it is more complex in terms of computations but the advantages outweigh the disadvantage. It is a viable alternative when randomization is not used.

5.0 SUMMARY

In this unit, you have worked through the analysis of covariance which we have described as a statistical tool used for analysing differences between experimental treatment and control groups on the dependent variable based on a pretest – post test design especially when intact groups are used. You have seen the uses and applications of the ANCOVA. In this unit, you have seen only the raw score method which is very simple and short.

6.0 TUTOR-MARKED ASSIGNMENT (TMAs)

- i) What is ANCOVA?
- ii) What are the uses of ANCOVA?

7.0 REFERENCES/FURTHER READINGS

Ali, A (1996). Fundamentals of Research in Education. Awka, Meks Publishers (Nig.).

Olaitan, S.O. and Nwoke, G. I. (1988). Practical Research Methods in Education. Onitsha. Summer Educational Pub. Ltd.

Ogomaka, P.M.C. (2004). Inferential Statistics for Research in Education and Social Sciences. Owerri. Peacewise Systems & Print.

UNIT 5 PREDICTION AND REGRESSION

Table of Contents

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Prediction
 - 3.2 The concept of Regression
 - 3.3 Regression Equation
 - 3.4 Linear Regression Equation: Example
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment (TMAs)
- 7.0 References/Further Readings

1.0 INTRODUCTION

The provision for predictions and the most exacting test of any hypothesis in statistics is a very important benefit of scientific investigation. You will have to note that statistical reasoning applies to all types of predictions in the behavioural sciences. Statistical ideas guide us to frame statements of a predictive nature; to say something definite about how trustworthy our predictions are and about how much error we should expect in the phenomenon predicted. In some research studies, you may wish to investigate the relationships between two or more variables or to establish a basis for making predictions. In this unit therefore, we shall look at prediction and regression, the regression equation and the coefficient of regression.

2.0 OBJECTIVES

After working through this unit, you should be able to:

- i) Explain the concept of prediction.
- ii) Describe the concept of regression.
- iii) Explain regression towards the mean.
- iv) Use regression equation to calculate the coefficients of regression.

3.0 MAIN CONTENT

3.1 Prediction

In your course EDU 701: Statistical Methods I, you treated correlation coefficients. You have seen that two variables can be said to be correlated. If this is so, it means that we can make estimates or predictions about what an individual's score will be on one variable given his score on the other variable. In other words, when we talk about the existence of a functional relation between two variables or variable quantities, we mean that they are connected by a simple formula. It means that they if the value of one of the variables is specified or given, the corresponding value of the other variable got through substitution in the formula. For instance, if we say that there is a positive relationship between IQ and achievement scores of students, then we can predict that a student, who is an average student as regards IQ, should have an average score in the achievement test. But some of the times, we are not perfect in our predictions. If we say that there is positive correlation between IQ and scores in achievement tests, it does not mean that every student who is above average in IQ must score above average in the achievement test. Some of the times, in the school system we use some subjects in the junior secondary schools as predictors of students' performance in their senior schools. If we are using scores in Integrated Science as a predictor variable to students' performances in the Sciences at the senior secondary school level, then students' performance in the sciences is the predicted variable. You will have to note that the predictor scores cannot ensure that every student will perform as predicted.

Except in cases where the correlation between variables is known to be perfect. It can however provide a reasonable estimate which can then be used as the basis for action. This implies that the accuracy of our prediction depends in part on the strength of the correlation. If the correlation between the variables in question is perfect, we can then make perfectly accurate predictions. In the sciences, we have pairs of associated variables which have perfect of functional relationships and that many of such relationships are linear. But in education and social sciences, variables that have functional or perfect linear relationships are very rare.

3.2 The concept of Regression

It may interest you to have an idea of the historical origin so as to help you understand the regression equations which we shall discuss in the next section. The idea of regression was introduced before correlation followed. It all started when Sir. Francis Galton making some studies of heredity

looked at the relationship between heights of children and heights of their parents. He plotted the first scatter-diagram. This he did in order to put parents and children on a common measuring scale. He transformed all heights to standard z – scores corresponding to certain fixed parents' heights. He also used the raw scores and found that the means of columns fell along a straight line trend. He was struck to note that the means of offspring height did not increase as rapidly as those of parents' heights. The mean heights of offspring deviated less from their general mean than the heights of the parents from which they came. The 'falling back' of heights of offspring toward the general mean has been called the Law of Filial regression. This is more generally known as regression toward the mean. It is a good illustration of imperfect correlation. When Galton wanted a single value which would express the amount of this regression phenomenon in any particular relationship problem. But Karl Pearson solved the problem with the formula which you have seen in the Pearson's' Product Moment Correlation Coefficient.

3.3 Regression toward the mean

Regression is an interesting phenomenon which is noticed when there is less than perfect relationship between the variables involved in prediction. Regression effect is the tendency for students who make extremely scores high or extremely low scores on a test to make less extreme scores or scores close to the mean on a second administration of the same test or on some predicted measure. Regression refers to the fact that the predicted score on a variable will be closer to the mean of the sample than is the predictor score. For instance, if you select a number of students in your class and these students are all alike on the X variable, you will find that these students will tend to be closer to the mean on the variable Y than they are to the overall mean on the X variable. Now let us use an example. If you select a number of students who are superior on an IQ test, you will see that while most of them are also above average on a class performance test, say in mathematics, only a very few will be as far above average in the performance test as they in intelligence. At the same time, if you select a group of students with low IQ scores, you will find that as a group, their scores on the performance test will lie closer to the mean than did their intelligence test scores. Therefore, unless two variables are perfectly correlated, there is a tendency for a group scoring at a given level above or below the mean on the first variable to be closer to the mean on the second variable. The effect on the scores is called regression effect while regression towards the mean of the second variable is called regression toward the mean. Regression is an inherent part of prediction. Predicted Y scores are closer to the mean than the X scores. The extent of correlation

between two variables determines how much regression will occur. If the correlation is perfect, $P = +1.00$ or -1.00 , then every measure in the first distribution is paired with another with the same relative positions in the second distribution and there may be no regression. Remember that regression occurs only when two variables have less than perfect correlation. If the correlation is high but not perfect, there is a slight tendency for the mean score of a group selected on the first variable to move toward the mean of the second variable. If the correlation is low, the tendency is for a more pronounced movement toward the mean of the second distribution. If the correlation is zero, there is complete regression to the mean.

3.4 Regression Equation

The main use of regression equation is to predict the most likely measurement in one variable from the known measurement in another. If we have the correlation between X and Y variables to be perfect, then we will make predictions of Y from X or of X from Y . The errors of prediction would be zero. You know that if the correlation is zero, predictions would be futile. It means that between the two extremes of perfect and zero correlation predictions will be possible with varying degrees of accuracy so, the higher the correlation, the greater the accuracy of prediction and the smaller the errors of prediction. Note that predictions or estimations are special forms of regression equations. Most of the times, variables of measurement are not expressed as deviations from their mean. In our discussion therefore we are going to use the raw scores whose variables are denoted by X and Y . To this effect we have:

$$\frac{Y - \bar{Y}}{\sigma_y} = r \frac{(X - \bar{X})}{\sigma_x}$$

where σ_y and σ_x are population's standard deviation of the variables. r is the Pearson Product Moment Correlation Coefficient of the two variables.

$$\square \quad Y = \frac{\sigma_y r}{\sigma_x} (X - \bar{X}) + \bar{Y}$$

This can be expanded to be in the form $Y = MX + C$. Then we have:

$$Y = \frac{(\sigma_y r)}{\sigma_x} X + (\bar{Y} - \frac{\sigma_y r}{\sigma_x} \bar{X})$$

$$\text{This means that for } X, \text{ we have } X = \frac{(\sigma_y r)}{\sigma_y} Y + (\bar{X} - \frac{\sigma_y r}{\sigma_y} \bar{Y})$$

$$\text{From these equations we can say, let } [Y - \frac{\sigma_y r}{\sigma_y} \bar{X}] = A_0 \text{ and } (\frac{\sigma_y r}{\sigma_y}) = A_1$$

so the regression equation obtained is $Y = A_1X + A_0$. Similarly if we say let

$$\left[\frac{\bar{X} - \frac{\sigma_y r}{\sigma_x}}{\sigma_y} \right] = b_0 \text{ and } \left[\frac{\sigma_y r}{\sigma_x} \right] = b_1 \text{ then } X = b_1 Y + b_0$$

You will have to note that regression equations are obtained from samples' scores by the calculation of samples' means, samples' standard deviations and samples' correlation coefficient. The results of these can be used to calculate A_0 and A_1 or b_0 and b_1 . Now let us reason that the required line has the equation $Y = A_1 X + A_0$; where Y is the predicted value. If Y is the observed value, the error $Y - \hat{Y}$ and $Y - A_1 X - A_0$. But this error $Y - \hat{Y} = (Y - A_1 X - A_0)$ could be positive or negative. So to remove the effects of negative values, we square both $Y - \hat{Y}$ and $Y - A_1 X - A_0$ so as to get $(Y - \hat{Y})^2 = (Y - A_1 X - A_0)^2$. For all such squares, we have $\sum(Y - \hat{Y})^2 = \sum(Y - A_1 X - A_0)^2$. From all these equations, we come out with the derivative which are simply put:

$$A_1 = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - (\sum X)^2/n} \quad \text{while } A_0 = \bar{Y} - A_1 \bar{X}$$

$$\text{On the other hand } A_1 = \frac{\sum XY - (\sum \bar{X})(\sum \bar{Y})/n}{\sum Y^2 - (\sum Y)^2/n}$$

$$\text{while } b_0 = X - b_1 Y$$

$$\text{But } \sum X^2 = \sum X^2 - (\sum X)^2/n \text{ and } \sum XY = (\sum X)(\sum Y)/n$$

$$\text{Where } X = X - \bar{X} \text{ and } y = Y - \bar{Y} \text{ therefore } A_1 = \frac{\sum XY}{\sum X^2}$$

There are two main methods of getting the coefficients of regression. The first is the equations given above. The method is known as method of least squares. But the second method which uses the normal equation of the line of best fit is given as: $\sum Y = (A_0)n + A_1 \sum X$ and $(\sum XY) = A_0 \sum X + A_1 \sum X^2$.

These equations will result to having:

$$A_1 = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} \quad A_0 = \frac{\sum Y \sum X^2 - (\sum Y) \sum XY}{n \sum X^2 - (\sum X)^2}$$

Activity 1

- i) What is the relationship between prediction and correlation?
- ii) What is the relationship between regression and correlation?
- iii) Regression equations are obtained from samples by calculating what?

iv) The two main methods of getting coefficients of regression are:

Answers to Activity 1

- i) The relationship between prediction and correlation is that the accuracy of our predictions depends on the strength of the correlation. If the correlation between variables is perfect, the predictions will be perfect and accurate. But when correlations are less than perfect, predictions are good estimates rather than exact predictions.
- ii) The extent of correlation between two variables determines how much regression takes place. If the correlation is perfect, there will be accurate prediction and no regression, but when two variables are less than perfect correction, there will be regression.
- iii) Regression equations are obtained by calculating (i) samples' means (ii) samples' standard deviations and (iii) samples' correlation coefficients.

3.5 Linear Regression Equation example:

Using the following data, develop prediction equations.

S/N	1	2	3	4	5	6	7	8	9	10	11	12
X	10	15	8	5	18	12	6	16	17	12	11	13
Y	13	9	10	11	12	8	7	6	1	2	3	4

1. Complete the composite table as follows:

						$X - \bar{X}$	$Y - \bar{Y}$			
S/N	X	Y	X^2	Y^2	XY	x	y	X^2	y^2	xy
1	10	13	100	169	130	-2.25	5.87	5.06	33.99	-13.21
2	15	9	225	81	135	2.75	1.83	7.56	3.35	5.03
3	8	10	64	100	80	-4.25	2.83	18.06	8.01	-12.03
4	5	11	25	121	55	-7.25	3.83	52.56	14.67	-27.77
5	18	12	324	144	216	5.75	4.83	33.06	23.33	27.77
6	12	8	144	64	96	-0.25	0.83	0.06	0.69	-0.21
7	6	7	36	49	42	-6.25	0.17	39.06	0.03	1.06
8	16	6	256	36	96	3.75	-1.17	14.06	1.37	-4.39
9	17	1	289	01	17	4.75	-6.17	22.56	38.07	-29.31
10	12	2	144	04	24	-0.25	-5.17	0.06	26.73	1.29
11	11	3	121	09	33	-1.25	-4.17	1.56	17.39	5.21
12	13	4	169	16	52	0.75	-3.17	0.56	10.05	-2.38
Σ	147	86	1897	794	976			194.22	177.68	-48.94

$$\bar{X} = \frac{147}{12} = \underline{\underline{12.25}} \quad \bar{Y} = \frac{86}{12} = \underline{\underline{7.17}}$$

2. Calculate the correlation coefficient.

$$\begin{aligned}
 &= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2] [n\sum Y^2 - (\sum Y)^2]}} \\
 &= \frac{12 \times 976 - 147 \times 86}{\sqrt{[12 \times 1897 - (147)^2] [12 \times 794 - (86)^2]}} \\
 &= \frac{11712 - 12642}{\sqrt{[22764 - 21609] [9528 - 7396]}} \\
 &= \frac{-930}{\sqrt{[2462460]}} \\
 &= \frac{-930}{1569.22} \\
 &= -0.5926501 = \sigma_{xy} = \underline{\underline{-0.59}}
 \end{aligned}$$

3. Calculate the standard deviation.

$$\begin{aligned}
 S_x &= \frac{\sqrt{[\sum X^2 - (\sum X)^2]}}{N} \\
 &= \frac{\sqrt{[1897 - (147)^2/12]}}{12} \\
 &= \frac{\sqrt{[1897 - 21609/12]}}{12} \\
 &= \frac{\sqrt{[1897 - 1800.75]}}{12} \\
 &= \frac{\sqrt{[96.25]}}{12} \\
 &= \sqrt{[8.0208333]} = \underline{\underline{2.83}}
 \end{aligned}$$

$$\begin{aligned}
S_y &= \frac{\sqrt{[\sum X^2 - (\sum X)^2]} }{N} \\
&= \frac{\sqrt{[794 - (86)^2/12]}}{12} \\
&= \frac{\sqrt{[794 - 7396/12]}}{12} \\
&= \frac{\sqrt{[794 - 616.33]}}{12} \\
&= \frac{\sqrt{[177.67]}}{12} \\
&= \sqrt{[14.8058333]} = \underline{\underline{3.8478349}} = \underline{\underline{3.85}}
\end{aligned}$$

4. Calculate A_1

$$\begin{aligned}
X &= \frac{\sigma_y r_{xy}}{\sigma_x} = \frac{S_y r_{xy}}{S_x} \\
&= \frac{3.85 X - 0.59}{2.83} \\
&= \frac{-2.2715}{2.83} \\
&= \underline{\underline{-0.803}}
\end{aligned}$$

5. Calculate A_0

$$\begin{aligned}
&= \bar{Y} - \frac{S_y r_{xy}}{S_x} \bar{X} \\
&= 7.17 - \frac{3.85 \bar{X} - 0.59 \times 2.83}{2.83} \\
&= 7.17 - \frac{(-6.428)}{2.83} \\
&= 7.17 + 2.2715 \\
&= \underline{\underline{9.44}}
\end{aligned}$$

Remember that $Y = A_1 X + A_0 = -0.8X + 9.44$.

4.0 CONCLUSION

You have learnt that prediction and regression are very important in statistics. You have seen that these can be done depending on the extent or degree of correlation of the two variables concerned. You have seen the relationships between correlation and prediction and correlation and regression. You can now calculate the coefficients. You can apply them now.

5.0 SUMMARY

In this unit, you worked through prediction and regression. You learnt that you can use a simple formula and by substitution get the value of a corresponding value of a variable if the value of the other variable is given. You have known that we can predict perfectly if the correlation between the two variables is perfect. But if there is no correlation we have regression. You have seen that the first man to work on regression was Sir Francis Galton and you have learnt what we mean by regression toward the mean. The regression equations are given by $Y = A_1X + A_0$ and $X = b_1Y + b_0$. With this we conclude our discussions on inferential statistics. In the next units, we shall be looking at how to test hypothesis using the non-parametric test, such as the chi square.

6.0 TUTOR-MARKED ASSIGNMENT (TMAs)

Given two sets of scores X and Y below, calculate the coefficient of regression.

S/N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	5	11	21	18	13	9	19	15	25	17	8	12	20	10	22
Y	15	12	8	9	11	13	7	9	8	9	14	12	6	11	5

7.0 REFERENCES/FURTHER READINGS

Ary, D. and Jacobs, L.C. (1976). Introduction to Statistics: Purposes and Procedures. New York. Chicago...London, Sydney. Holt, Rinehart & Winston.

Guilford, J.P. and Fucliter, B. (1981). Fundamental Statistics in Psychology and Education. Sixth Edition. Auckland, Bogota...Sydney, Tokyo. McGraw-Hill. Inc.

Ogomaka, P.M.C. (2004). Inferential Statistics for Research in Education and Social Sciences. Owerri: Peacewise Systems and Prints.

Module 6

Chi – Square tests

Unit 1 Introduction to chi-square

Unit 2 The chi-square test

Unit 3 The chi-square – a continuation

UNIT 1 THE CHI – SQUARE (χ^2) TESTS:

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
	3.1 General features of Chi square
	3.2 The basic nature of Chi square
	3.3 Test of significance
4.0	Conclusion
5.0	Summary
6.0	Tutor-Marked Assignment (TMAs)
7.0	References /Further Readings

1.0 INTRODUCTION

In the previous units on inferential statistics, you have learnt how to compute statistical tests from representative samples as estimates of the corresponding population parameters. These statistical tests are known as parametric tests of significance. Can you recall some of the tests? You remember the t – test, z – test and the F – ratio! You will also remember that these parametric tests are based on certain assumptions regarding the parameters. Now, let us remember some of these assumptions which are implicit in the use of the parametric tests. These are:

- (i) That the data represent population in other words the data are interval or ratio;
- (ii) That the variable has a normal or near normal distribution in the population; and
- (iii) That the sample statistic provides an estimate of the population parameter.

But some of the times there is a need to test hypothesis with data which are ordinal or nominal or even with interval data which are not normally distributed or which fail to meet the necessary assumptions for the use of parametric tests. Therefore, non-parametric tests which require fewer assumptions than the parametric tests can be used in a wide variety of situations where parametric test cannot be used. These non-parametric tests are used for testing hypotheses involving nominal or ordinal data. In this unit, you will be working through one of such non-parametric tests which is called chi – square (pronounced – Kai square).

2.0 OBJECTIVES

After working through this unit, you will be able to:

- i) Explain the features of chi square;
- ii) Calculate the chi square one variable method; and
- iii) Take decisions about the null hypothesis using the chi square.

3.0 MAIN CONTENT

3.1 General features of Chi – Square

The chi – square which is symbolised by X^2 is a procedure which is used to test hypothesis about the independence of frequency counts in various categories. In other words, it is used with data in the form of frequencies or data which can be easily transformed into frequencies. This includes proportions and probabilities. Chi – square has very important features, one of which is its additive property, which makes possible the combination of several statistics or other values in the same test.

3.2 The basic nature of Chi – Square

You have already learnt about the z score, which is a standard score or measure in your EDU 701. Recall the formula for finding a z score. If you have done this, then note that the fundamental nature of chi – square can be very simply or completely explained on the basis of the z score. The chi – square is identified with the square of z, i.e. z^2 when there is one degree of freedom. Therefore the mathematical relation of chi – square X^2 to squared z score z^2 with one degree of freedom is given by:

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

where X is any measurement in a normally distributed population, μ is its mean and σ is its standard deviation.

Now think of a situation where we have a sampling situation in which there are k mutually independent measures of x . We also have k mutually independent z values and k mutually independent X^2 values. Then, a most useful property of X^2 is that a sum of k mutually independent chi – square values is also a X^2 , with k degrees of freedom. If we put this statement into equation, we shall have that:

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

This implies that if we have some values of X such as $X_1, X_2, X_3, \dots, X_n$ then we should have that:

$$\begin{aligned} X^2(n) &= \frac{(X_1 - \mu)^2}{\sigma^2} + \frac{(X_2 - \mu)^2}{\sigma^2} + \frac{(X_3 - \mu)^2}{\sigma^2} + \dots + \frac{(X_n - \mu)^2}{\sigma^2} \\ &= \sum_{i=1}^n \frac{[X_i - \mu]^2}{\sigma^2} \end{aligned}$$

You will have to recall that S^2 which is variance is given by:

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

This implies that $\sum (X - \bar{X})^2 = (n - 1)S^2$ or $S^2 = \frac{\sum (X - \bar{X})^2}{n - 1} = nS^2$

Note that the above is just to give you an idea about the relationship between the z and the X^2 . It may not be necessary for you to dwell or spend most of your time on it. The next section will be of immense interest to you.

3.3 Test of Significance

Earlier, we have said that the chi- square is a test of independence and it is used for analyzing data that are in the form of frequency counts occurring in two or more mutually exclusive or discrete variables being compared. It is used for comparing the significance of the difference between two proportions that are actually observed and expected. The observed frequency is data obtained from the actual frequency count while the expected frequency is data that are expected when equal numbers of

response to the same variables equally. This means that the larger the value between the observed and the expected, the higher the chi – square value. The calculated chi – square value is compared against a given critical value from the table to determine whether it is significant. Now let us use some examples to illustrate the use of chi – square as a test of significance.

Example 1

Etiti U.C.B. Ventures is an organisation which has many business outfits. A section of this business venture has produced a type of pomade, which they want to push into the market. They want to find out first which colour of the pomade best appeals to ladies. The organisation has contracted a research fellow who used the process of randomization to select 600 ladies who are interested in the pomade. Each of the 600 ladies are given the three colours of pomade. The ladies are asked to indicate their colour preference for the pomade. They are asked to be very objective and honest in their indications. The data showing preferred colour are shown as follows:

Colour of Pomade	White	Blue	Pink
Number of Ladies Choosing	150	350	100

In this type of problem, you will note that chi – square tries to determine the difference between the expected which is theoretical and the observed which is the actual values, and therefore enables us to test the null hypothesis. Here, the expected for all the colours would be 200 i.e. $600 \div 3$.

The null hypothesis would be: There is no significant difference between choices made according to colours.

Colour of Pomade	White	Blue	Pink	Total
Expected Choices f_e	200	200	200	600
Observed Choices f_0	150	350	100	600

To test the null hypothesis, we use the formula:

$$\begin{aligned}
 X^2 &= \sum \frac{(f_0 - f_e)^2}{f_e} \\
 X^2 &= \frac{(150 - 200)^2}{200} + \frac{(350 - 200)^2}{200} + \frac{(100 - 200)^2}{200} \\
 &= \underline{-50^2} + \underline{150^2} + \underline{-100^2}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{200}{2500} + \frac{200}{22500} + \frac{200}{10000} \\
 &= \frac{200}{2500} + \frac{200}{22500} + \frac{200}{10000} = \underline{\underline{175}}
 \end{aligned}$$

The degree of freedom is given by $K - 1$ where K is the number of categories. In our example $K = 3$ so the degree of freedom is $3 - 1 = 2$.

Decision: To take decision on the null hypothesis, go to the chi – square table and check the critical value under df 2 and alpha level of 0.05 (given).

$$X^2 \text{ cal} = 175, X^2 \text{ tab} = 5.991$$

□ Since $X^2 \text{ cal} > X^2 \text{ tab}$, we reject the null hypothesis. Note that this type X^2 test is often described as a goodness of fit technique. This is because it permits us to determine whether or not a significant difference exists between the observed number of cases falling into each category and the expected number of cases based on the null hypothesis. Now look at the example above. If there is no significant difference in colour preference, then we would expect 200 ladies to choose each of the coloured pomades. Note again that the example we have given represents the calculation of X^2 one variable case. But in most research studies, researchers are concerned with the interrelationships between and among variables. we shall look at this in the next unit. Meanwhile let us give another example.

Example 2

A random sample of students in Okigwe Education Zone was conducted by a researcher based on language preference in the school system. The question given was based on whether French should be compulsory in all classes of the secondary school system. The results are as follows:

Opinion	Agree	Undecided	Disagree	Total
Number of Ladies Choosing	300	50	650	1000

- i) Propose the null hypothesis.
- ii) Find the expected value under each category.
- iii) What is the chi – square?
- iv) What is the degree of freedom?
- v) What is the decision at 0.05 level?

Solution

i. $H_0 =$ There is no significant difference between the observed and the expected frequencies in the opinion of the students from Okigwe Education Zone on making French compulsory in the school system.

ii. The expected value under each category would be:

$$1000 \div 3 = \text{Agree} = 333.33, \text{Undecided} = 333.33, \text{Disagree} = 333.33$$

$$\begin{aligned} \text{iii. } X^2 &= \frac{(300 - 333.33)^2}{333.33} + \frac{(50 - 333.33)^2}{333.33} + \frac{(650 - 333.33)^2}{333.33} \\ &= \frac{1108.89}{333.33} + \frac{80258.89}{333.33} + \frac{100298.98}{333.33} \\ &= 3.327 + 240.801 + 300.927 \\ &= \mathbf{545.055} \end{aligned}$$

$$\text{iv. } df = K - 1 = 3 - 1 = 2$$

v. Decision:

$$X^2_{ca;} = 545.055. X^2_{tab} (2.005) = 5.991.$$

Since $X^2_{cal} > X^2_{tab}$, we reject the null hypothesis H_0 .

Now do the following activity.

Activity 1

Using the table below, calculate the chi – square and take a decision on the H_0 at 0.05 level.

Opinion	B.A.	A	N	D	S.D.	Total
Frequency	200	250	100	150	100	800

Answer to Activity 1

You may have given the answer below.

1. $H_0 =$ There is no significant difference between the opinions expected and observed.

2. Frequency Table

Opinion	B.A.	A	N	D	S.D.	Total
Frequency Observed	200	250	100	150	100	800
Frequency Expected	160	160	160	160	160	800

The expected frequency in each category is $800 \div 5 = 160$.

$$\begin{aligned}
 3. X^2 &= \frac{(200-160)^2}{160} + \frac{(250-160)^2}{160} + \frac{(100-160)^2}{160} + \frac{(150-160)^2}{160} + \frac{(100-160)^2}{160} \\
 &= \frac{40^2}{160} + \frac{90^2}{160} + \frac{-60^2}{160} + \frac{-10^2}{160} + \frac{-60^2}{160} \\
 &= \frac{1600}{160} + \frac{8100}{160} + \frac{3600}{160} + \frac{100}{160} + \frac{3600}{160} \\
 &= 10 + 50.625 + 22.5 + 0.625 + 22.5 \\
 &= \mathbf{106.25}
 \end{aligned}$$

Decision:

$$X^2 \text{ cal} = 106.25, X^2 \text{ tab at } (4 : 0.05) = 9.488$$

We reject that there is no significant difference and accept that there is a significant difference.

4.0 CONCLUSION

You have now seen that most of the times you can use non-parametric test to verify your hypotheses especially in data treatment where the data are discrete and also when the study involves opinions observed and expected. Therefore, if your research data do not meet the requirements and assumptions in the parametric tests, the chi – square is there as one of the useful tests to employ.

5.0 SUMMARY

In this unit, you have learnt that the chi – square is a test of independence used for analyzing discrete data in the form of frequency counts occurring in two or more mutually exclusive variables, being compared. It is used for comparing the significance differences between the observed and the expected frequency especially in studies involving opinion polls. You have learnt how to calculate the one variable method. In the next unit, we shall be looking at the contingency table.

6.0 TUTOR-MARKED ASSIGNMENT (TMAs)

Use the table below to calculate the chi – square and take a decision on the H^0 .

Colour Preference	Blue	Brown	Black	Pink	Total
Frequency Observed	50	50	50	50	200
Frequency Expected	60	45	48	47	200

7.0 REFERENCES/FURTHER READINGS

- Ali, A. (1996). Fundamentals of Research in Education. Awka, Mekis Publishers (Nig.)
- Ary, P. and Jacobs, L.C. (1976). Introduction to Statistics, Purposes and Procedures. New York. Chicago .. London, Sydney. Holt Rinehart & Winston.
- Denga, I.D. and Ali, A. (1983). An Introduction to Research Methods and Statistics in Education and Social Sciences. Jos, Savannah Publishers Limited
- Ogomaka, P.M.C. (2004). Inferential Statistics for Research in Education and Social Sciences. Owerri, Peacewise Systems and Prints.

UNIT 2 CHI – SQUARE: A CONTINUATION

Table of Contents

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Test of Significance
 - 3.2 Contingency table
 - 3.3 Illustrations
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment (TMAs)
- 7.0 References/Further Readings

1.0 INTRODUCTION

In the last unit, you worked through the introduction to the chi – square test. You learnt that a wide class of non-parametric tests find their basis in the chi – square distribution. Some of these tests resemble the ANOVA tests in a vague way and offer a non-parametric alternative to testing in multivariable situations. A common situation which often arises in multiple classification studies is one in which each individual of some finite population can be categorized into exactly one of K mutually exclusive categories. The problem of interest here is to ascertain whether or not the frequencies observed fit some preconceived probabilities attached to the categories. In other words, we try to find out if the differences between the observed frequency counts and the expected frequency counts are significant or not. In this unit, you will continue to work through the chi – square especially in the area of test of independence and the contingency table.

2.0 OBJECTIVES

After working through this unit, you should be able to:

- i) explain the meaning of contingency table;
- ii) calculate the chi – square using the contingency tables; and
- iii) describe the restrictions in the use of the chi – square.

3.0 MAIN CONTENT

3.1 Test of independence

In the chi – square test, most of the times, investigations involve the frequency distributions of various categories or classes of two variables exhibited by or associated with some group of individuals. Each of these two variables may be dichotomized, trichotomised or polychotomised according to Ogomaka (2004). It may be possible that the two variables may not have equal number of categories or classes. Generally, data collected from such investigation are presented in contingency tables or cells. Before we continue, let us look at a run down of the procedures you will follow when applying the chi – square test of independence.

- i) You will propose a null hypothesis to state that the two variables are independent of each other; that is knowledge of an individual's classification on one variable would indicate nothing of his classification on the other variable.
- ii) Draw a random sample and classify the subjects into two or more variables.
- iii) Set up a contingency table with the variables indicated in the rows and columns of the table.
- iv) Record the observed frequencies in the proper cells and determine the marginal totals.
- v) Derive the expected frequencies from the observed data. The sum of the expected frequencies will equal the sum of the observed frequencies.
- vi) Calculate the chi – square and compare it with the critical value on the table at a predetermined level of significance and with degree of freedom (rows – 1) (columns – 1).
- vii) If the calculated X^2 value equals or exceeds the value in the table, the finding is significant and the null hypothesis of independence is rejected. You then conclude that the two variables are dependent or related at the given level of significance. But if the calculated X^2 values is less or smaller than the tabled value, the null hypothesis of independence is accepted or retained. The conclusion is that there is no sufficient evidence of relationship between the two variables.

3.2 Contingency Table

You would have noticed that the concept contingency table has been appearing in our discussions of the chi – square. Let us explain it in some

detail here. A contingency table is a planner grid having m rows and n columns. It has $m \times n$ cells and is therefore regarded as $m \times n$ contingency table. The number of rows m , is always given first before the number of columns, n . Note again that m may not be equal to n . The word contingency itself means “by chance”. So in a contingency table we are basically interested in examining whether the frequencies we observe in the cells of the table are more or less what we would expect by chance. This implies that we want to find out if the two classifications are independent.

3.3 Illustrations

Example 1

In a recent poll conducted by a teacher in a state school system trying to find out the opinions of the students on the introduction of one uniform system of dressing in the secondary schools in the state. He used a sample of 1000 students randomly selected from all the local government areas of the state. The results classified according to their ages are given below. Test the claim that opinion is independent of age group at 0.05.

Age / Opinion	Yes	No	Undecided	
Under 13	106	186	41	
13 – 16	156	120	57	
Above 16	152	101	80	

Solution

1. Complete the observed contingency table.

Age / Opinion	Yes	No	Undecided	Total
Under 13	106	186	41	333
13 – 16	156	120	57	333
Above 16	152	101	80	334
Totals	414	408	178	1000

2. Complete the expected contingency table. The expected frequency in each cell of the table is given by $\frac{\text{column total} \times \text{row total}}{\text{overall total}}$

Expected Contingency table.

Age / Opinion	Yes	No	Undecided	Total
Under 13	i) 137.862	ii) 135.864	iii) 59.274	333
13 – 16	iv) 137.862	v) 135.864	vi) 59.274	333
Above 16	vii) 138.276	viii) 136.272	ix) 59.452	334
Totals	414	408	178	1000

$$\text{For cell 1} = \frac{414 \times 333}{1000} = 137.862.$$

$$\text{For cell 2} = \frac{408 \times 333}{1000} = 135.864.$$

$$\text{For cell 3} = \frac{59.274 \times 333}{1000} = 59.2.74$$

$$\text{For cell 4} = \frac{414 \times 333}{1000} = 137.862.$$

$$\text{For cell 5} = \frac{408 \times 333}{1000} = 137.864.$$

$$\text{For cell 6} = \frac{178 \times 333}{1000} = 59.274.$$

$$\text{For cell 7} = \frac{414 \times 334}{1000} = 138.276.$$

$$\text{For cell 8} = \frac{414 \times 334}{1000} = 136.272.$$

$$\text{For cell 9} = \frac{414 \times 334}{1000} = 59.452.$$

3. Computer the chi – square.

You will note the formula $X^2 = \sum \frac{(O - E)^2}{E}$ can best used in a table when there are many cells to treat.

S/N	Observed	Expected	O – E	(O – E) ²	(O – E) ² /E
1	106	137.862	- 31.862	1015.187	7.364
2	186	135.864	50.136	2513.619	18.501
3	41	59.274	- 18.274	333.939	5.634
4	156	137.862	18.138	328.987	2.386
5	120	135.864	- 15.864	251.667	1.852
6	57	59.274	- 2.274	5.171	0.087
7	152	138.276	13.724	188.348	1.362
8	102	136.272	- 34.272	1174.570	8.619
9	80	59.452	20.548	422.220	7.102
					52.907

4. Find the degree of freedom = $(c - 1)(r - 1) = (3 - 1)(3 - 1) = 4$
5. Decision:
 $X^2_{cal} = 52.907$ X^2_{tab} at $(4 : 0.05) = 9.488$
 We reject the H_0 .

Example 2

In a recent survey carried out in one of the commercial towns in Nigeria on whether the number of cars owned by individuals depends on the income yielded these results.

No. of cars / Annual income in Naira	Under 50,000	50,000 – 100,000	100,000 – 250,000	200,000 – 500,000	Above 500,000
One	130	330	110	140	80
Two	60	145	60	80	55
More than two	30	45	40	50	60

Test the independence hypothesis at both 0.01 and 0.05.

Solution:

1. Complete the observed table.

No. of cars / Annual income in Naira	Under 50,000	50,000 – 100,000	100,000 – 250,000	200,000 – 500,000	Above 500,000	Totals
One	130	330	110	140	80	790
Two	60	145	60	80	55	400
More than two	30	45	40	50	60	225
Totals	220	520	210	270	195	1415

2. Complete the expected frequency table.

No. of cars / Annual income in Naira	Under 50,000	50,000 – 100,000	100,000 – 250,000	200,000 – 500,000	Above 500,000	Totals
One	122.83	290.32	117.24	150.74	108.87	790
Two	62.19	147.00	59.36	76.33	55.12	400
More than two	34.98	82.69	33.39	42.93	31.01	225
Totals	220	520	210	270	195	1415

3. Complete the chi – square.

S/N	Observed	Expected	O – E	(O – E) ²	(O – E) ² /E
1	130	122.83	7.17	51.41	0.419
2	60	62.19	- 2.19	4.80	0.077
3	30	34.98	- 4.98	24.80	0.709
4	330	290.32	39.68	1574.50	5.423
5	145	147.00	- 2.00	4.00	0.027
6	45	82.69	- 37.69	1420.54	17.179
7	110	117.24	- 7.24	52.42	0.447
8	60	59.39	0.64	0.41	0.000
9	40	33.39	6.61	43.69	1.309
10	140	150.75	- 10.75	115.56	0.767
11	80	76.33	3.67	13.47	0.176
12	50	42.93	7.07	49.98	1.164
13	80	108.87	- 28.87	833.48	7.656
14	55	55.12	- 0.12	0.01	0.000
15	60	31.01	29.99	840.42	27.102
					62.455

4. Find the degree of freedom = $(c - 1)(r - 1) = (5 - 1)(3 - 1) = \mathbf{8}$

5. Decision:

$$X^2 \text{ cal} = 62.455 \quad X^2 \text{ tab at } (8 : 0.01) = 20.090$$

$$X^2 \text{ tab at } (8 : 0.05) = 15.507$$

$X^2 \text{ cal}$ is greater than $X^2 \text{ tab}$ at both 0.01 and 0.05 so we reject the null hypothesis at both levels.

Activity 1

A random sample of teachers classified according to the level of their schools and asked about their opinion on unionism in the school system, are as follows:

	Elementary	Junior Sec.	Senior Sec.	Teachers Coll.
Yes	10	20	25	30
No	30	20	15	10

Are the differences statistically significant at 0.01 level.

Solution:

1. Complete the observed frequency table.

	Elementary	Junior Sec.	Senior Sec.	Teachers Coll.	Totals
Yes	10	20	25	30	85
No	30	20	15	10	75
Totals	40	40	40	40	160

2. Complete the expected frequency table.

	Elementary	Junior Sec.	Senior Sec.	Teachers Coll.	Totals
Yes	21.25	21.25	21.25	21.25	85
No	18.75	18.75	18.75	18.75	75
Totals	40	40	40	40	160

3. Complete the chi – square table.

S/N	Observed	Expected	O – E	(O – E) ²	(O – E) ² /E
1	10	21.25	- 11.25	126.56	5.956
2	20	21.25	- 1.25	1.56	0.074
3	25	21.25	3.75	14.06	0.662
4	30	21.25	8.75	76.56	3.603
5	30	18.75	11.25	126.56	6.750
6	20	18.75	1.25	1.56	0.083
7	15	18.75	- 3.75	14.06	0.750
8	10	18.75	- 8.75	76.56	4.083
9					21.961

4. Find the degree of freedom = $(c - 1)(r - 1) = (4 - 1)(2 - 1) = \underline{3}$

5. Decision:

$$X^2 \text{ cal} = 21.961 \quad X^2 \text{ tab at } (3 : 0.01) = 11.341$$

The differences are statistically significant at 0.01 level, so we reject H_0 .

4.0 CONCLUSION

You have seen that some of the times the data in a study do not meet the parametric assumptions. In such cases, a variety of descriptive and inferential nonparametric procedures can be used. This chi – square which you have just studied in this unit is one of them. You will find it very interesting to use especially in your survey research.

5.0 SUMMARY

In this unit, we have looked at the chi – square in a contingency table. You have learnt that the goodness of fit chi – square is used to establish whether observed proportions differ significantly from the expected or a hypothesised distribution. When subjects are categorized on the basis of different classification variables, the chi – square test of independence tests the null hypothesis that the variables are independent of each other.

6.0 TUTOR-MARKED ASSIGNMENT (TMAs)

1) What is the number of degrees of freedom in the following types of contingency tables?

(a) 2 x 2 (b) 3 x 5 (c) 4 x 2 (d) 3 x 2 (e) 4 x 3.

2) Calculate the chi – square in the data given:

Gender / Opinion	Yes	No	Undecided	Total
Boys	25	30	15	70
Girls	40	15	20	75

7.0 REFERENCES/FURTHER READINGS

Ary, D. and Jacobs, L.C. (1976). Introduction to Statistics: Purposes and Procedures. New York. Chicago...London, Sydney. Holt, Rinehart & Winston.

Ogomaka, P.M.C. (2004). Inferential Statistics for Research in Education and Social Sciences. Owerri: Peacewise Systems and Prints.

Zehna, P.W. (1974). Introductory Statistics. Boston, Massachusetts. Prindle Weber and Schmidt, inc.

UNIT 3 CHI – SQUARE: CONCLUSION

Table of Contents

1.0	Introduction
2.0	Objectives
3.0	Main Content
3.1	Non parametric test of significance
3.2	Advantages and disadvantages of nonparametric tests
3.3	Restrictions in the use of chi – square
4.0	Conclusion
5.0	Summary
6.0	Tutor- Marked Assignment (TMAs)
7.0	References

1.0 INTRODUCTION

You have worked through the previous two units where we discussed the chi – square as a strong nonparametric test used in testing hypotheses in survey researches involving observed and expected frequencies, or opinion polls. In this unit, we shall be concluding the chi – square test as well as the course, by looking at the nonparametric tests, their advantages and disadvantages and the assumptions underlying the use of chi – square.

2.0 OBJECTIVES

After working through this unit, you should be able to:

- i) explain what is meant by nonparametric tests;
- ii) give the advantages and disadvantages of such tests; and
- iii) describe the assumptions underlying the use of the chi – square test.

3.0 MAIN CONTENT

3.1 Non-Parametric Test of Significance

In the statistical methods I: EDU 701, you treated many of the descriptive statistical tests such as the mean, mode, standard deviation among others. There are many more which you have not treated. These include the sign test, Wilcoxon, rank test, the median test, Mann whitney u – test, Kruskal – Wallis ANOVA by ranks phi – coefficient among others. Most of these

tests would be treated in a course called parametric and non-parametric tests. But note that non-parametric statistics are distribution free tests of significance because they make no assumption about the shape of the distribution tested. Unlike the parametric statistics which are statistical tests which make use of the normal probability model such as comparison involving setting up of confidence limit, degree of freedom, and also determining whether the obtained statistical value or the calculated value or the value on the table and accepting or rejecting a stated hypothesis. You would have noticed that the chi – square, though a non-parametric test, involves all these rigours. That is why it is a powerful non-parametric test of significance. Before we look at the assumptions of the chi – square, let us look at the advantages and disadvantages of nonparametric tests.

3.2 Advantages and disadvantages of Non-Parametric Tests

There are a number of advantages derived from using the nonparametric tests. Usually, there is the ease in calculation, the ease in presentation as tables and the ease in interpretation. They describe and give clearer pictures about events, results, observations etc. Because most studies use or entail the use of non-parametric and which use nominally scaled variables, the use of parametric tests will be less useful and more confusing. But the nonparametric tests have the disadvantage of not being rigorous tools for describing and testing data. This is because when we use them on data, we lose some information e.g. the mode or median. They do not give us the exact size of the scores involved.

Activity 1

Answer true or false.

- i) The chi – square is most appropriate for the analysis of data which are classified as frequency of occurrence within categories or nominal data.
- ii) The chi – square is usually preferred when analysing data at the ordinal or interval level.
- iii) The chi – square is a parametric test of significance.
- iv) Non-parametric statistics are distribution free tests.
- v) Mann whitney – u test is a parametric test of significance.

Answers to Activity 1

- i) True
- ii) False

- iii) False
- iv) True
- v) False

3.3 Restrictions in the use of Chi - Square

Although the chi – square is a nonparametric, inferential statistical test, it is not free from assumptions. Ogomaka (2004) listed such assumptions as:

- (1) Most times the working out of expected frequencies are based upon approximated probabilities. These approximated probabilities are assumed to follow the multinomial rule. It is therefore necessary that each and every observation categorized should be independent of each other observation. This is to say that chi – square test is not appropriate to use when there are dependent observations.
- (2) When the degrees of freedom for a chi – square test is greater than one, the least expected frequency in each cell should not be less than five. However, for chi – square test with one degree of freedom, a minimum expected frequency of 10 per cell is safer than otherwise.
- (3) When a chi – square test involves two categorized values of two variables, there are joint-observations, and the resulting joint frequency table must be complete. This implies that each of all the observations made must belong or fall into one and only one cell or joint-event possibility. In other words, each distinct observation made or that is possible must be identified with or without doubt belong to one and only one row, and one and only one column. It must inevitably belong to one and only one cell of the resulting contingency table.
- (4) Finally or indeed the whole idea of chi – square test rests upon the randomness of the observations or sample and the prior specification of the categories into which observations will fall. Chi – square test rests upon random sampling. It also demands that the categories of the variables with which the observations are identified must be chosen or specified in advance. It is therefore not sound to pool frequencies together after the observations have been made or after the data have been seen. This last note is very important considering the assumptions stated above.

4.0 CONCLUSION

You have seen that the chi – square, though not a parametric test, is a procedure used to test hypothesis about the independence of frequency counts in various categories. There are many other nonparametric tests

which are not covered in either EDU 701 or EDU 702. You will get those ones in statistics textbooks if you need to use them. In these courses, we have tried to discuss these tests which you will need including some you may not need in your projects.

5.0 SUMMARY

In this unit, you have worked through the concluding part of the chi – square and indeed the concluding part of this course. You have looked at the nonparametric tests, their advantages and disadvantages and the assumptions underlying the use of the chi – square, which include that the data must be categorized or at the nominal level and that in a contingency table, no cell should contain less than five frequency counts.

6.0 TUTOR-MARKED ASSIGNMENT (TMAs)

- 1) What are the advantages of using non-parametric tests.
- 2) Mention one disadvantage of nonparametric test.
- 3) What is the minimum number of frequencies required to use the chi – square test?

7/0 REFERENCES/FURTHER READINGS

Ary, P. and Jacobs, L.C. (1976). Introduction to Statistics, Purposes and Procedures. New York. Chicago .. London, Sydney. Holt Rinehart & Winston.

Ogomaka, P.M.C. (2004). Inferential Statistics for Research in Education and Social Sciences. Owerri, Peacewise Systems and Prints.

Olaitan, S.O. and Nwoke, G.I. (1988). Practical Research Methods in Education. Onitsha: Summer Educational Publishers Limited.