COURSE GUIDE

BIO206
STATISTICS FOR AGRICULTURE AND BIOLOGICAL SCIENCES

Course Team: Dr Iliya S. Ndams (Course Writer)- Ahmadu Bello University, Zaria
Olatunji Arowolo (Course Editor)- School Of Science And Tech Lagos State Polytechnic
Mohammed Kabir (PhD) (Content Reviewer)- IBB University Lapai

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

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

e-mail: centralinfo@nou.edu.ng

URL: www.nou.edu.ng

Published by
National Open University of Nigeria
Printed 2009. Reprinted 2017
Reviewed: 2023

ISBN: 978-978-058-945-5

Course Guide:

Introduction

Statistics for Agriculture and Biological Sciences (Biostatistics) referred to as BIO 206 is a second semester year two course. It is a compulsory two-credit unit course hosted in the Department of Biological Sciences.
Statistics as the science of estimates and probabilities; as it is the collection, presentation, analysis and interpretation of numerical data. It is also a body of methods for making wise decision in the face of uncertainty. Thus, Biostatistics is the application of statistics with biological (living) system. Therefore, Biostatistics have applications in many fields such as Biology, Botany, Microbiology, Zoology, Biochemistry, Nursing, Physiology, and Medicine amongst others.

The course shall be looking at use of statistical methods in Biology and Agriculture; continuous and discrete variables; sampling procedures, sampling size; presentation of statistical results. Frequency distribution, law of probability, binomial, Poisson and normal frequency distributions. Measure of central tendency and measure of variability/dispersion. Estimations and tests of hypothesis. Design of simple Agricultural and Biological experiments. Analysis of variance and co-variance, simple regression and correlation, contingency tables, some non-parametric tests and ecological statistics. The use of statistical packages in statistical analysis.

Course Competencies

The course will provide general overview of the course synopsis; this course material shall be divided into appropriate sections to help the learners understand and assimilate the contents of the course. The course guide will help students to understand how to go about Tutor- Marked-Assignment which will form part of the overall assessment at the end of the course.

Similarly, structured on-line facilitation classes in this course shall increase the comprehension of the course thus students are encouraged to activity participate. This course exposes students to data collection, management and analysis, the knowledge will be helpful during your project data collection and analysis, it is indeed very interesting field of Biology.

Course Objectives

This course is aimed at providing students the knowledge of biostatistics and its application in the fields of life sciences especially Biological and Agricultural sciences.
The course objectives are to;

- understand the importance and significance of statistics in life sciences;
- differentiate and understand the different types of distributions in biostatistics;
- determine the measure of central tendency and variability/dispersion.
- differentiate methods of estimation and testing hypotheses;
- ascertain methods of determining relationships between variables;
- understand methods of sampling and experimental design;
- understand statistical tools for non-parametric and ecological tests;
- use computer approach to Biostatistics.

Working Through this Course

The successful completion of this course entails the studying of the course guide and the reference textbooks/materials as well as other materials provided by the National Open University of Nigeria. The course guide is divided into sections, each section has self-assessment exercise. The practice of the assessment will positively influence your academic performance in the course. The course is expected to cover a minimal period of 8 weeks to complete.

Study Units

The Modules of this course shall be in accordance with the course objectives thus;

Module 1: Basics of Biostatistics.
Unit 1: Concept of Biostatistics
Unit 2: Frequency Distribution
Unit 3: Probability Distribution
Unit 4: Methods of Estimation & Sampling
Unit 5: Concept of Hypotheses Formulation & Experimental Design

Module 2: Biostatistics Application I
Unit 1: Measures of Central Tendency
Unit 2: Measures of Dispersion
Unit 3: Student's t-distribution

Unit 4: Contingency Table
Unit 5: Analysis of Variance & Co-variance

Module 3: Biostatistics Application II
Unit 1: Simple Linear Regression
Unit 2: Simple Linear Correlation
Unit 3: Non-Parametric Tests
Unit 4: Ecological Statistics
Unit 5: Computer Approach to Biostatistics

References and Further Readings

In every section or Module, Reference materials shall be provided for further reading.

Presentation Schedule

Assignment	**Marks**
TMA 1-4	Four T M A s , best three marks of the four count at 10% each - 30% of course marks.
End of course examination	70% of overall course marks
Total	100% of course materials

Assessment

In every section or Module, self-assessment questions shall be provided for further practice.

How to get the Most from the Course

The course guide is designed in a simplified form to assist self comprehension. In addition, further references with web links are provided in each sectio/module or unit. Similarly, the course has facilitation session that will provide information on any grey areas.

Online Facilitation

Eight weeks is scheduled for online facilitation. This facilitation is divided into two session (synchronous and asynchronmous). The synchronous session is a live session that is provided by a facilitator through University approved source (Zoom) for 1 hour. While the asynchronous session is an alternative interaction session that may not be live. In the facilitator dashboard, students have access to the course materials, recorded online facilitation, weblinks, virtual library and host of others that would improve the course comprehension.

Course Information

Course Code: BIO 206
Course Title: Biostatistics
Credit Unit: 2
Course Status: Core
Course Blub: This course exposes students to data collection, management and analysis, the knowledge will be helpful them during their projects' data collection and analysis.
Semester: Second
Course Duration: 2 hours per week (16 hours per semester)
Required Hours for Study: 3 x 2hours x 8 week (48hrs)

Ice Breaker

Dr. Kabir Mohammed Adamu, is an Associate Professor of Hydrobiology, Fish Nutrition and Physiology, Department of Biological Sciences, Ibrahim Badamasi Babangida University, Lapai, (IBBUL) Niger State, Nigeria. He is an external Facilitator with the Department of Biological Sciences, National Open University of Nigeria (NOUN), where he facilitates the BIO 206 (Biostatistics) amongst other courses. He has been teaching/lecturing Biostatistics for the past fifteen (15) years in various level (undergraduate and postgraduate) of students in different tertiary institutions (IBBUL, NOUN, Nasarawa State University, Keffi, Nile University of Nigeria, Abuja). Dr. Kabir's research interest is in circular economy by understanding the interaction of freshwater fisheries with the environment, using both phenotypic and genotypic techniques in characterization of fisheries resources and their roles in healthy aquatic ecosystem. Understanding the protein requirement of fish and seeking for protein (especially insect protein) resource fish growth, nutrition and physiology.

Module 1: Basics of Biostatistics

Unit 1: Concept of Biostatistics

CONTENTS

1.1 Introduction

Biostatistics is playing an increasing important role in nearly all phases of living organisms. This section has been designed for insights into the study of biostatistics. There are different definitions for statistics such as: Boddington defines statistics as the science of estimates and probabilities. Croxton and Cowden defines statistics as the collection, presentation, analysis and interpretation of numerical data while Wallis and Robert define it as a body of methods for making wise decision in the face of uncertainty. It is therefore, used to referred to as any numerical characteristics of a set of data based on a sample rather than the population. **Thus, in summary Statistics is concerned with the scientific methods for collecting, organizing, summarizing, presenting and analyzing data, as well as drawing valid conclusions and making reasonable decisions on the bases of such analysis**. Therefore, the term **biostatistics** or biometry is simply statistics applied to biological problems.

1.2 Intended Learning Outcomes (ILOs)

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

- Understand the background and essence of biostatistics.
- Be acquitted with the common terminologies in biostatistics.

- Understand the requirements for data processing and presentations and,
- Know the limitations of biostatistics

1.3: Concept of Biostatistics

1.3.1: Types of Statistics

There are four types of statistics thus;

i. Descriptive: it describes the properties of the observed frequency distribution concisely and accurately after having gathered the data or facts from a sample population. It involves the summary and orderly presentation of such collected data in tabular or graphical forms. Descriptive information of these data is achieved with the use of measures of central tendency and measures of dispersion.

ii. Inferential: the drawing of inference or making some generalized conclusions from the summarized data about the characteristics of a whole population from the characteristics of its part, in this case the sample. This type of statistics is considered to be more probabilistic than deterministic due to its subjectivity.

iii. Enumeration: it deals with discrete numbers and attributes data. The answer here is qualitative 'yes' or 'no' situation e.g., spore forming bacteria versus non-spore bacteria, or smokers versus non-smokers.

iv. Measurement: it deals with continuous data. E.g., height of 5cm, 5.8cm, 8.4cm etc. or weights of 110.5kg, 128.5kg, etc.

1.3.2: Usefulness of Statistics

a. It permits the presentation of facts/data in a vivid/concise and definite form. It is a substitute for the 'rule of thumb' by concise and unambiguous representation of quantitative and qualitative data and facts.

b. It permits for the reduction/condensation of facts/data to a more manageable size so that they could convey much clearer and distinct meanings. It specializes in the contraction and simplification of quantitative data.

c. Statistics is used to verify formulated tests and hypotheses. Tests of hypotheses are specific assumptions made about the overall population, and they are crucial in virtually all statistical investigation.

d. It is used in the developing of alternative design for biological experiment. A good research plan must have good statistical analysis thus in experimental procedure the statistics required is a main factor to the success of the research.

1.3.3: Terminologies

i. Population/Universe is defined as the entire collection of measurements about which the statistician wishes to draw conclusion. If one for example wishes to draw conclusions about the heights of students of a National Open University of Nigeria, then the population under consideration comprises the heights of all students in the University.

ii. Samples is a subset of all the measurements in the population of interest. Since the populations of interest are generally so large, making it unfeasible to obtain all the measurements desired, the statistician normally resorts to obtaining a subset of all the measurements in the population. A sample is a subgroup of the population selected for study. When a sample is chosen at random from a population, it is said to be an unbiased sample. That is, the sample for the most part, is representative of the population. But if a sample is selected incorrectly, it may be a biased sample when some type of systematic error has been made in the selection of the subjects. However, the sample must be random in order to make valid inferences about the population. The importance of samples are;

a) It saves the researcher time and money.
b) It enables the researcher to get information that he or she might not be able to obtain otherwise.
c) It enables the researcher to get more detailed information about a particular subject.

iii. Variable is an observable quantity/attribute that varies from one biological entity (member of the population being studied) to another. Sometimes, the term variate is used, though its usage is less frequent. There are two types of variables;

a) Qualitative variables are variables that can be placed into distinct categories, according to some characteristic or attribute. For example, if subjects are classified according to gender (i.e. male or female), then the variable gender is qualitative.
b) Quantitative variables are numerical and can be ordered or ranked. For example, the variable age is numerical, and people can be ranked according to the value of their ages. Quantitative variables can be grouped into two:

a. Discrete variables these are variables that can be assigned values such as 0,1,2,3 (integers) and are said to be variables that assume values that can be counted. Examples include number of children in a family, number of birds in a pen, number of trees in a garden, number of animals per litter etc.
b. Continuous variables these are variables that can assume all values between any specific values. They are obtained by measuring. This applies to variables such as length, weight, height, yield,

temperature and time,that can be thought of as capable of assuming any value in some interval of values.

iv. Parameters is a descriptive property that describes/characterize a population such as descriptive property is referred to as a parameter (e.g., means). Parameters are constant for a population, but statistics may vary from sample to sample taken from the same population. A statistic is an estimator of the population parameter from the sample.

v. Data/Facts: these are numerical facts obtained from observations. Once it is decided what type of study is to be made, it becomes necessary to collect information about the concerned study, mostly in the form of data. For this, information has to be collected from certain individuals directly or indirectly such a technique is known as survey method. Another way of collecting data is by experimentation i.e., an actual experiment is conducted on certain individuals/units about which the inference is to be drawn.

1.3.4: Processing of Statistics Data

The processing of data involves the following:

a. Completeness: this ensures that the data are complete and no missing information.
b. Consistency: this avoids contradictory information/data.
c. Accuracy: collection of accurate data and its computation
d. Editing: to maintain homogeneity, the information sheets are checked to see whether the unit of information/measurement is the same in all the schedules.

1.3.5: Presentation of Data

Data can be represented in any of the four measurement scales:

a. Nominal scale: a nominal measurement scale is a set of mutually exclusive categories that varies qualitatively but not quantitatively. The variables under consideration are classified by some quality rather than by numerical measurement.
b. Ordinal scale: this is also a set of mutually exclusive categories that varies qualitatively but not quantitatively, however, the variables are arranged in a rank order/hierarchy.
c. Interval scale: this is similar to the aforementioned but varies in having size interval with an arbitrary zero point.
d. Ratio scale: this is measurement scale having both constant size interval and true zero point. For instance, the total lengths of a group of earthworms are one variable on the other hand the number of segments per earthworm is another variable. Regardless of

method of measurement and of counting, there are two fundamentally important characteristics of this scale of measurement, thus;

a) There is a constant size interval between successive units on the measurement scale.

b) There is a true zero point on the measurement scale and there is a physical significance to this scale.

1.3.6: Accuracy of Measurement

Accuracy is the nearness of any measurement to the actual value of the variable being measured. This depends on the precision required, the tools available for the measurement and the skill of the person undertaking the task.

1.3.7: Rounding of Figure

Sometimes the figures (values) are rounded by reducing one or more decimal places to the unit place/nearest to the ten or hundredth of a number. The universal rules of rounding are;

a. if the decimal place value to be rounded is less than 5, it should be deleted straight away
   3.23**4**6 ======== **3.23**
b. if it is greater than 5, the preceding number is increased by 1.
   3.23**6**6 ==========**3.24**
c. in case the decimal place value to be rounded off is 5, the rule is to delete it if the preceding number is even and increase the preceding number by 1 if it is odd. The rule is applicable in general for any numerical value.
   3.23**5***4* ================== **3.23**
   3.23**5***3* ==================== **3.24**

1.3.8: Limitations of Statistics

The limitations of statistics are;

a. the methods are applicable to qualitative characters like ‘yes’ or ‘no’ as they cannot be coded.
b. It does not deal with a single character/value or observation.
c. The conclusion of analyses is not application for all samples but for most the sample/population.
d. It does not take care of the changes occurring to the individuals but it does reveal the changes occurring in a mass or group of individuals.

The collections of quantitative information/data are called?
The variable that can be placed into distinct categories, according to some characteristics is called?

Self-Assessment Exercises

1. Variables that are assigned integers are called?
2. The measurements and observations that variables can assume is called?

1.4: Summary

An introduction to the unit cum Biostatistics was provided, thereafter the usefulness of Biostatistics to life sciences were enumerated. The different terminologies used in Biostatistics were listed and explained. Data presentation and its requirements were also discussed in the Unit. Similarly, the rules of rounding figures were stated and the limitations of biostatistics were also discussed.

1.5: References/Further Reading/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp

Basic Concepts for Biostatistics: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_BiostatisticsBasics/

Introduction to Biostatistics Some Basic Concepts: http://fac.ksu.edu.sa/sites/default/files/introduction_to_biostatistics-106_1.pdf

Biostatistics Concept & Definition – SlideShare: https://www.slideshare.net/bijayabnanda/ls-bs-1concept-definition

https://www.youtube.com/watch?v=a23NhTCyxUY
https://www.youtube.com/watch?v=_e4mwlqCQrc

1.6: Possible Answers to Self-Assessment Exercises

1- Discrete variables
2- Data

Unit 2.: Frequency Distribution

Unit Structure

2.1: Introduction

Very often, large amount of data are encountered in research. The best way to present these data is by the use of frequency tables, which assists to summarize the data, making them more meaningful. This involves the listing of all observed values of the variables under consideration and stating how many times each value is observed (i.e., the frequency).
Measurements or counting gives rise to raw data. Raw data itself is difficult to comprehend because it lacks organization, summarization, which renders it meaningless. Thus, the raw data has to be put in some order through classification and tabulation so as to reduce its volume and heterogeneity. To describe situations, draw conclusions or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a *frequency distribution.*

2.2: Intended Learning Outcomes

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

- Understand ways of data collections and how to arrange data.
- Describe how to prepare frequency distribution data.
- Know how to present data on graphs.

2.3: Frequency Distribution

2.3.1: Raw data & Array

The collected data that have not been organized numerically is called RAW DATA. OR they are data recorded in the original way they were sampled. An example is the set of masses of 100 male students obtained from an alphabetical listing of University record. While the arrangement of raw data in either ascending or descending order of magnitude is called an ARRAY.

2.3.2: Graphical Presentation

Graph has the ability to show at a glance the main characteristics/quality of a set of presented data. There are two major types of graphical presentation of quantitative data thus;

Pie Charts
A **pie chart** is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The whole pie (or circle) represents the total sample or population. It is defined as a circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories.

Bar Graphs
A graph made of bars whose heights represent the frequencies of respective categories is called a *bar graph*. The bar graphs for relative frequency and percentage distributions can be drawn simply by marking the relative frequencies or percentages, instead of the frequencies, on the vertical axis. Sometimes a bar graph is constructed by marking the categories on the vertical axis and the frequencies on the horizontal axis.

2.3.3: Frequency preparation

When summarizing large masses of raw data, it is often useful to distribute the data into classes or categories and to determine the number of individuals belonging to each class, called the **class frequency.** A tabular arrangement of data by classes together with the corresponding class frequencies is called a **frequency distribution** or **frequency table.** Data organized and summarized in frequency table are often referred to as **grouped data.** A *frequency distribution* for qualitative data lists all categories and the number of elements that belong to each of the categories.

a. Categorical Frequency This is used for data that can be placed in specific categories, such as nominal or ordinal-level data. It is useful to know the proportion of values that fall within a group, category or observation rather than the number of values or frequencies. To get the *relative frequency*, the frequency of occurrence of each number is divided by the total number of values and multiplied by hundred. This can be expressed as follows: $\frac{F}{X}$ x100 Where f = Frequency of the category class and n = total number of values.

b. Ungrouped Frequency Distribution
This is a list of the figures in array form, occurring in the raw data, together with the frequency of each figure, i.e., a frequency is constructed for a data based on a single data value for each class.

c. Grouped Frequency Distribution
The heights in inches of commonly grown herbs are shown below. Organize the data into a frequency distribution with six classes, and make useful suggestions.

The reasons for constructing a frequency distribution are:

a. To organize the data in a meaningful, intelligible way.
b. To enable the reader to determine the nature and shape of the distribution.
c. To facilitate computational procedures for measures of average and spread.
d. To enable the researcher to draw charts and graphs for the presentation of data.
e. To enable the reader to make comparisons among different data sets.

General rules for forming frequency distributions

a. Determine the largest and smallest numbers in the raw data and thus find the range.
b. Divide the range into a convenient number of class intervals having the same size. If this is not feasible, use class intervals of different sizes or open class intervals. Class intervals are chosen so that the class marks or midpoints coincide with actually observed data. This tends to lessen the so-called grouping error involved in further analysis. However, class boundaries should not coincide with actually observed data.
c. Determine the number of observations falling into each class interval, i.e., find the class frequencies. This best done by using tally or score sheet.

Class interval and class limits:
Class interval is the symbol defining a class; while the end number is called class limit. The lower number is the lower-class limit and the higher number the upper-class limit. The term class and class interval are often used interchangeably, although the class interval is actually a symbol for the class.

The size or width of a class interval
The size or width of a class interval is the difference between the lower- and upper-class boundaries and is also referred to as class width, class size or class strength.

The class marks
This is the midpoint of the class interval and is obtained by adding the lower- and upper-class limits and divide by two. The class mark is also called class midpoint.

Graphing Grouped Data
Grouped (quantitative) data can be displayed in a *histogram* or a *polygon*. A **histogram** can be drawn for a frequency distribution, a relative frequency distribution, or a percentage distribution. A histogram is called a **frequency histogram**, a **relative frequency histogram**, or a **percentage histogram** depending on whether frequencies, relative frequencies, or percentages are marked on the vertical axis. A *histogram* is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on the vertical axis. The frequencies, relative frequencies, or percentages are represented by the heights of the bars. In a histogram, the bars are drawn adjacent to each other.

Histograms and frequency polygons:

1. A histogram or frequency histogram consist of a set of rectangles having:

a. bases on a horizontal axis (the x-axis) with center at the class marks and length equal to the class interval sizes.
b. areas proportional to class frequencies.

2. A frequency polygon is a line graph of class frequency plotted against class mark. It can be obtained by connecting midpoints of the tops to the rectangles in the histogram.

Polygons
A **polygon** is another device that can be used to present quantitative data in graphic form. A polygon with relative frequencies marked on the vertical axis is called a *relative frequency polygon*. Similarly, a polygon with percentages marked on the vertical axis is called a *percentage*

polygon. A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a *polygon*.

Relative frequency distributions

The relative frequency of a class is the frequency of the class divided by the total frequency of all classes and is generally expressed as a percentage. If frequencies are replaced by corresponding relative frequency the resulting table is called a relative frequency distribution, percentage distribution or relative frequency table.

Graphical representations of relative frequency distributions can be obtained from the histogram or frequency polygon by simply changing the vertical scale from frequency to relative frequency, keeping exactly the same diagram. The resulting graphs are called relative frequency histograms or percentage histograms and relative frequency polygons or percentage polygon respectively.

Cumulative frequency distribution (ogives)

The total frequency of all values less than the upper-class boundary of a given class interval is called the cumulative frequency up to and including the class interval. A table presenting such cumulative frequencies is called a cumulative frequency distribution, cumulative frequency table or briefly a cumulative distribution. A graph showing the cumulative frequency less than any upper-class boundary plotted against the upper-class boundary is called cumulative frequency polygon or ogives.

Elative cumulative frequency distribution (percentage ogives)

The relative cumulative frequency or percentage cumulative frequency is the cumulative frequency divided by the total frequency. If relative cumulative frequencies are used in place of cumulative frequencies, the results are called relative cumulative frequency distributions or percentage cumulative distributions and relative cumulative frequency polygon or percentage ogives respectively.

Frequency curve (Smooth ogives)

Frequency polygon or relative frequency polygon for a large population have small broken line segments that are closely approximate curves, which are called frequency curves or relative frequency curves respectively. Thus, frequency curve is sometimes called smoothed frequency polygon.

Arrange the numbers 17, 45, 38, 27, 6,58,48, 32, 19, 22, 34 in an array and determine the range.

1. The number of occurrences of an element in a sample is called?
2. The frequency distribution used for data can be placed at a specific group is called?
3. When classes do not have overlapping class limit is one of the rules of?
4. To organize data in a meaningful, intelligible way is one of the rationales of?
5. If there exist A, B, O and AB blood group amongst 40 students determine the percentage of A with 12 frequencies.
6. If there exist A, B, O and AB blood group amongst 40 students determine the percentage of B with 11 frequencies.
7. If there exist A, B, O and AB blood group amongst 40 students determine the percentage of O with 10 frequencies.
8. If there exist A, B, O and AB blood group amongst 40 students determine the percentage of AB with 7 frequencies.
9. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, determine the range of the distribution.
10. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, assuming the number of class to be used is 6, determine the class width
11. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, determine the cumulative frequency of the distribution
12. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, determine the percentage frequency of the distribution
13. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, determine the mean of the distribution
14. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, determine the mode of the distribution
15. The length of herbs grown in the garden were 18, 20, 16, 14, 31, 21, 16, 8, 10, 28, 34mm, determine the median of the distribution.
16. The data below represents the blood groups of 40 students in a Biostatistics class. Construct a frequency distribution for the data

Self-Assessment Exercises

A	AB	B	O	O	A	B	AB	A	B
O	O	O	A	AB	B	B	A	O	AB
A	O	O	A	AB	B	B	A	A	B
AB	A	O	B	AB	O	A	B	A	B

2.4: Summary

In this Unit, the way to present of large data were discussed and instances provided appropriately. This Unit discuss the methods and requirements for preparation of frequency tables and graphs.

76	73	75	73	74	74	72	75	76	73
68	72	78	74	75	72	76	76	77	70
78	72	70	74	76	75	75	79	75	74
75	70	73	75	70	74	76	74	75	74
78	74	75	74	73	74	71	72	71	79

2.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Bailey, N.T.J. (1994). *Statistical Methods in Biology*. Third Edition. Cambridge University Press. United Kingdom.
- Bluman, A.G. (2004). *Elementary Statistics*. A Step by Step Approach. Fifth Edition. McGraw-Hill Companies Incorporated. London.
- Daniel, W.W. (1995). *Biostatistics*: a foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.
- Harper, W.M. (1991). *Statistics.* Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.
- Helmut F. van Emden.(2008). *Statistics for Terrified Biologists.* Blackwell Publishing Limited. USA.

Frequency Distribution - Definition, Types, Examples: https://www.cuemath.com/data/frequency-distribution/

Frequency Distribution - Quick Introduction - SPSS tutorials: https://www.spss-tutorials.com/frequency-distribution-what-is-it/

https://www.toppr.com/guides/maths/statistics/frequency-distribution/

https://www.youtube.com/watch?v=amLYLq73RvE

https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cumulative-frequency-distribution/

2.6: Answers to Self-Assessment Exercises

1. Frequency
2. Categorical frequency
3. Construction of frequency
4. constructing a frequency distribution
5. 30
6. 27.5
7. 25
8. 17.5
9. 26mm
10. 5
11. 216mm
12. 100
13. 19.63mm
14. 16mm
15. 18mm

Unit 3: Probability Distribution

Unit Structure

3.1: Introduction

This section explains the occurrence of event by chance, thus, the different types of probability distribution. Probability can be defined as the chance of an event occurring. It is the basis of inferential statistics. Probability is a branch of mathematics which as a general concept can be defined as the chance of an event occurring. It is the basis of inferential statistics. A distribution is a scatter of related values, such as the assortment of weights in a group of cattle. A *Probability distribution* shows how probable given random variable values in a range of such values are. Therefore, a *probability distribution* is simply a complete listing of all possible outcomes of an experiment, together with their probabilities.

3.2: Intended Learning Outcomes

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

- understand probability distribution
- distinguish the different types of probability distribution.
- understand the circumstances of using the different forms of probability distribution

3.3: Probability distribution

3.3.1: Normal distribution

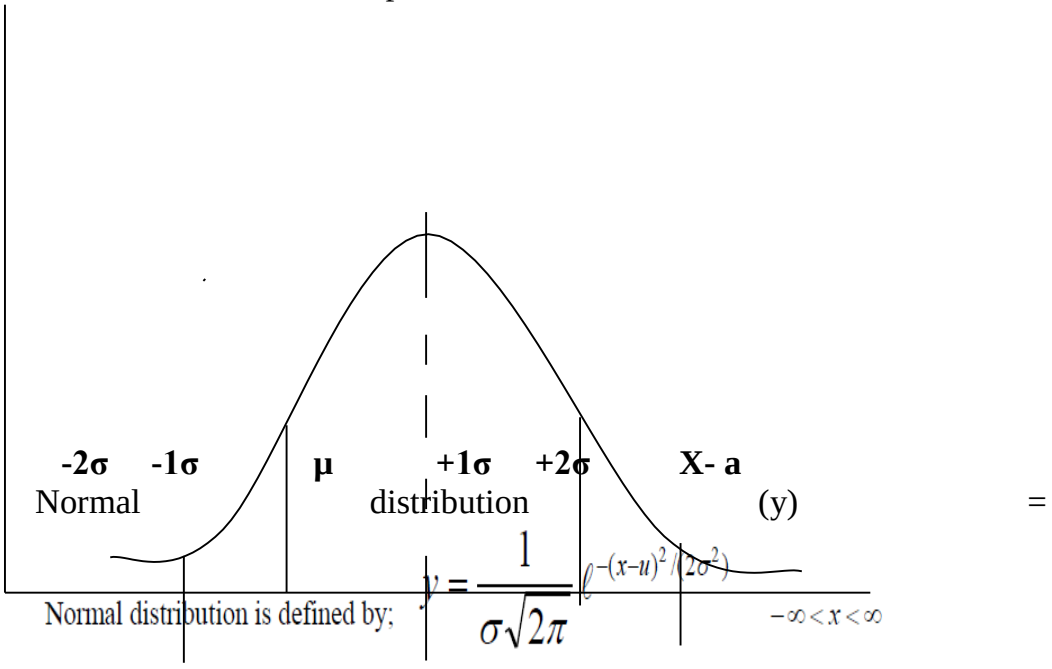
This is the most important distribution in statistics. It is also known as the Gaussian distribution named after Gauss, a German astronomer who

showed its use in statistics. The normal distribution is defined by just two statistics, the *mean* and the *standard deviation*. Normal distribution is concerned with results obtained by taking measurements on continuous random variable (i.e., the quantified value of a random event) like weight, yield etc. The *normal distribution* is a particular pattern of variation of numbers around the mean. It is symmetrical (hence we express the standard deviation as ±) and the frequency of individual numbers falls off equally away from the mean in both directions. It so happens that the curve given by this probability's distribution approximates very closely to a *Mathematical curve.* This curve is called the *Normal curve.*

In checking for normality, it is important to know whether an experimental data is an approximate fit to a normal distribution. This is easily checked with large samples. There should be roughly equal numbers of observations on either side of the mean. Things are more difficult when we have only a few samples. In experiments, it is not uncommon to have no more than three data per treatment. However, even here we can get clues. If the distribution is normal, there should be no relationship between the magnitude of the mean and its standard deviation.

Properties of a Normal Curve

i. It is a Unimodal symmetrical curve.
ii. The mean, mode & median all coincide, thereby dividing the curve into two equal parts.
iii. Most items on the curve are clustered around the mean
iv. No kurtosis or skewness in the curve
v. The area beneath the curve is proportional to the observation associate with the part.



The important aspect of the curve is the area in relationship with probability, if perpendiculars are erected at a distance of 1σ from the mean and in both directions, the area covered by these perpendiculars and the curve will be about 68.26% of the total area. (It means that 68.26% of all the frequencies are formed within one standard deviation of the mean). The total probability encompassed by the area under the curve is 1 (100%).
The measures of central tendency (μ = Population mean) and dispersion (σ = Population standard deviation) are the parameters of the distribution and once they have been estimated for a particular population, the shape of its distribution curve can be worked out using the normal curve formula. Usually, we do not know the values of μ and σ and have to estimate them from a sample as *x* and *s*. If the number of observations in the sample exceeds about 30, then *x* and *s* are considered to be reliable estimates of the parameters.

Standardizing the Normal Curve
Any value of an observation *X* on the baseline of a normal curve can be standardized as a number of standard deviation units, the observation is away from the population mean, μ. This is called a *z*-score. To transform *x* into *z* the formula is given by:
$z = ((x - \mu))/\sigma$
If the population mean μ is larger than the sample mean *x*, the *z* is negative. But if the sample size is more than about 30 observations, the sample mean (x) and standard deviation (s) are considered to be good estimates of μ and σ, and z is given by:
$z = ((x - \mu))/s$
If the calculated value of z is larger than 1.96 (i.e., $P< 0.05$ or 95% confidence coefficient) then this is regarded as unlikely or statistically significant.

3.3.2: Poisson distribution

A Poisson distribution is a discrete probability distribution that is useful when n is larger and p is small and when the independent variables occur over a period of time. It can be used when a density of items is distributed over a given area or volume, such as the number of plants growing per acre. It can also be used to discover whether organisms are randomly distributed. For example, in ecological studies, Poisson distribution is used to describe the spread of organisms like insects, trees, and snails' etc. by the following:

i. Divide the large area into small squares of equal size.
ii. Count the particular animal or plant species under study in each square.

iii. You can also randomly select a number of squares, if the area is two large.
iv. The probability of *X* occurrences in an interval of time, volume, area etc. for a variable where λ (lambda) is the mean number of occurrences per unit (time, volume, area etc) is given by:

$$P(x,\lambda) = \frac{\ell^{-\lambda}\lambda^{x}}{x!} \qquad where \quad x = 0,1,2,3,.....$$

e = constant, approximately equal to 2.7183

3.3.3: Binomial distribution

A *binomial distribution* is a special probability distribution that describes the distribution of probabilities when there are only two possible outcomes for each trial of an experiment. A *binomial* experiment is a probability experiment that satisfies the following four requirements:

i. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. i.e., these outcomes can either be success or failure. No two events can occur simultaneously.
ii. There must be a fixed number of trails.
iii. The outcomes of each trial must be independent of each other.
iv. The probability of a success must remain the same for each trial.The binomial probability formula is given by

$$P(x) = \frac{n!}{(n-x)!x!} p^{n} q^{n-x}$$

Where: p = Numerical probability of a success = P(s)
q = Numerical probability of a failure = P(F)
n = Number of trials
x = The number of successes in n trials
! = Mathematical symbol called 'factorial'. So n! means multiple all the numbers in a count down from the total number in the sample.
State the formula of normal distribution.
A survey on birds showed that one out of five fire finch was trapped using mix net, in a given season, if 10 birds were selected at random, find the probability that 3 of the birds were trapped in the previous season.

Self-Assessment Exercises

1. The term used to describe how probably given random variable values in a range of such values occur is?
2. The other name for normal distribution is?
3. A particular pattern of variation of number around the mean is?
4. Normal distribution is easily checked with…….sample.
5. Unimodal symmetrical curve is a property of?
6. State the formula for poison distribution.
7. The discrete probability distribution that is useful when n is large and p is small is?
8. A survey on birds showed that one out of five fire finch was trapped using mix net, in a given season, if 10 birds were selected at random, what type of probability distribution is applicable here?
9. A survey on birds showed that one out of five fire finch was trapped using mix net, in a given season, if 10 birds were selected at random, find the numerical probability of success?
10. A survey on birds showed that one out of five fire finch was trapped using mix net, in a given season, if 10 birds were selected at random, find the numerical probability of failure?

3.4: Summary

This unit is able to discuss about the three different types of probability distribution with worked examples. An introduction to probability distribution was provided. The three types of probability distributions were discussed with appropriate exercises.

3.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Bailey, N.T.J. (1994). *Statistical Methods in Biology*. Third Edition. Cambridge University Press. United Kingdom.
- Bluman, A.G. (2004). *Elementary Statistics*. A Step by Step Approach. Fifth Edition. McGraw-Hill Companies Incorporated.

London.

- Daniel, W.W. (1995). *Biostatistics*: a foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.

What is Probability Distribution? https://byjus.com/maths/probability-distribution/

Probability Distribution | Formula, Types, & Examples https://www.scribbr.com/statistics/probability-distributions/

Probability Distribution | Types of Distributions

https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/

https://www.youtube.com/watch?v=UnzbuqgU2LE

https://www.khanacademy.org/math/precalculus/x9e81a4f98389efdf:prob-comb/x9e81a4f98389efdf:probability-distributions-introduction/v/discrete-probability-distribution

https://www.youtube.com/watch?v=CfZa1daLjwo

https://www.coursera.org/lecture/statistics-international-business/probability-and-probability-distributions-an-introduction-ASvPO

3.6: Possible Answers to Self-Assessment Exercises

1. Probability distribution
2. Gaussian distribution
3. Normal distribution, Gaussian distribution
4. Large
5. Normal curve
6. $P(x, \lambda) = \frac{\ell^{-\lambda}\lambda^{x}}{X!}$
7. Poisson distribution
8. Binomial
9. 0.2
10. 0.8

Unit 4: Methods of Estimation & Sampling

Unit Structure
4.1: Introduction
4.2: Intended Learning Outcomes
4.3: Methods of Estimation and Sampling
4.4: Summary
4.5: References/Further Readings/Web Sources
4.6: Possible Answers to Self-Assessment Exercises

4.1: Introduction

The use of estimation has been discussed in this section as well as its application in life sciences. In order to obtain unbiased samples, several sampling methods have been developed. The most common methods are random, systematic, stratified, and clustered sampling.

4.2: Intended Learning Outcomes

By the end of this unit, students should be able to:
- understand the essences of estimation
- understand the methods of sampling

4.3: Methods of Estimation and Sampling

Estimation is the entire process of using an estimator to produce an estimate of a parameter. Estimation and hypothesis testing are interrelated. An estimate is any specific value of a statistic while an estimator is any statistic used to estimate a parameter. For example, the sample mean x is used to estimate the populations mean μ. A Point estimate is obtained when a single number is used to estimate a population parameter. For example s = 30. An Interval estimate is obtained when a range of values is used to estimate a population parameter. For example, a range of values between 20 and 30 allows evaluation of the estimate unlike the point estimate of a single value.

In sampling, there are different methods of sampling such as;

i. **Random Sampling**: For a sample to be a random sample, every member of the population must have an equal chance of being selected. Therefore, a random sample is one that has the same chance as any other of being selected. Randomness assists in

avoiding various forms of conscious and unconscious bias and can be achieved by these two ways:

a. Number each element of the population and then place the numbers on cards. Place the cards in a hat or bowl, mix them, and then select the sample by drawing the cards. You must ensure that the numbers are well mixed.

b. The second and most preferred way of selecting a random sample is to use random numbers e.g. Table of random numbers by Fisher and Yates. The table comprises of a series of digits 0, 1, 2…. up to 9 arranged as such that each number had the same chance of appearing in any given position.

ii. **Stratified Sampling:** A stratified sample is a sample obtained by dividing the population into subgroups, called strata, according to various homogenous (alike) characteristics and then selecting members from each stratum for the sample. For example, you can group the items on basis of their age, size, colour etc. The advantage of stratified sampling is that it increases precision because all types of groups are represented through stratification and a heterogeneous population is made into a homogenous one.

iii. **Cluster Sampling:** A cluster sample is a sample obtained by selecting a preexisting or natural group, called a *Cluster* and using the members in the cluster for the sample. For example a habitat, or a large area or field is divided into smaller units and a number of such units are randomly selected and used as a sample. There are three advantages to using a cluster sample instead of other types of sample;

a. A cluster sample can reduce cost.

b. It can simplify field work.

c. It is convenient.

The major disadvantage of cluster sampling is that the elements in a cluster may not have the same variations in characteristics as elements selected individually from a population.

iv. **Systematic OR Skip Sampling:** This method involves taking an item as a sample from a larger population at regular intervals. For example, when sampling from a poultry farm, every third or fifth or tenth chick coming out of the cage is taken and included in the sample. This is done after the first number is selected at random for counting to start.

v. **Proportionate Sampling:** This type of sampling involves selecting a sample in proportion to the different groups in the population under study. In sequence sampling – successive units taken from production lines are sampled to ensure that products meet certain standards set by the manufacturing company. This is used in quality control. In double sample, a

very large population is given a questionnaire to determine those who meet the qualifications for a study. After the questionnaires are reviewed, a second, smaller population is defined. Then a sample is selected from this group. In multistage sampling, the researcher uses a combination of sampling methods.

What is obtained when a range of values is used to estimate a population parameter?
Any specific value of a statistician is called _____

Self-Assessment Exercises 1

1. Any statistics used to estimate a parameter is called?
2. When a single number is used to estimate a population, parameters is called?
3. An unbiased sample is when a sample is?
4. The most common methods to obtain unbiased sample is?
5. A sample that has the same chance as any of other of being selected is?

4.4: Summary

The Unit discussed estimation in biostatistics. Estimation was discussed and exercises were considered. The Unit examined the different methods of sampling and their applications. The different methods of sampling and their applications were dealt.

4.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Bailey, N.T.J. (1994). *Statistical Methods in Biology*. Third Edition. Cambridge University Press. United Kingdom.
- Bluman, A.G. (2004). *Elementary Statistics*. A Step by Step

Approach. Fifth Edition. McGraw-Hill Companies Incorporated. London.

- Daniel, W.W. (1995). *Biostatistics*: a foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.
- Harper, W.M. (1991). *Statistics.* Sixth Edition. Pitman Publishing, Longman Group, United Kingdom.
- Helmut F. van Emden.(2008). *Statistics for Terrified Biologists.* Blackwell Publishing Limited. USA.

Sampling and estimation techniques - European Union
https://ec.europa.eu/eurostat/ramon/nace-cpa-webpage/manuals_nat_experiences/docs/estimation_NL_1.pdf

Estimation - Biostatistics - University of Florida
https://bolt.mph.ufl.edu/6050-6052/unit-4/module-11/

5 Sampling and Estimation - The National Academies Press
https://www.nap.edu/read/25098/chapter/7

Methods of Sampling and Estimation
https://www.stat.go.jp/english/data/shakai/2006/pdf/suikei.pdf

https://www.youtube.com/watch?v=pqEPtona94A

https://www.youtube.com/watch?v=5P57_sobsnk

https://www.coursera.org/lecture/statistical-inference-for-estimation-in-data-science/estimators-and-sampling-distributions-2Zfz7

4.6: Possible Answers to Self-Assessment Exercises

1. Estimator
2. Point estimate
3. Chosen at random from a population
4. Random, Systematic, Stratified, Clustered
5. Random

Unit 5: Concept of Hypotheis formulation & Experimental Design

Unit Structure

5.1: Introduction

The use of hypotheses testing has been discussed in this section as well as its application in life sciences. To understand the concept of hypothesis, there is need to understand the steps in the scientific methods as it relates to the conduct of a research. Conducting research in Biology and other natural sciences involves the following steps:

a. Forming a research question
b. Performing background research or preliminary study
c. Creating a ***hypothesis*** for an experiment
d. Designing and conducting an experiment
e. Collecting data from the experiment
f. Analyzing the results with the appropriate methods from the collected data
g. Making an inference or drawing conclusions from the analyzed results
h. Presenting the results through proper communication.

Hypothesis can be defined as a tentative prediction about the nature of the relationship between two or more variables. OR as a tentative explanation of the research problem, a possible outcome of the research, or an educated guess about the research outcome. OR as declarative sentence form, and they relate, either generally or specifically, variables to variables. OR as a statement or explanation that is suggested by knowledge or observation but has not, yet, been proved or disproved.

Experimental design is a field of investigation concerned with the structure and planning of experiments. Experiment is nothing more

than a procedure governed by a set of rules to draw a sample from a population. This set of rules is the experimental design, pragmatically defined as the complete sequence of steps taken well in advance to ensure that the appropriate data are collected in a way that permits objective analysis leading to valid inferences about the problem been investigated. The requirements of a good experiment are;

A. Absence of bias (i.e., the association of a particular kind of error with a particular treatment.
B. Internal estimate of experimental error in order to carry out the test of significance.
C. Precision (i.e. repetition of measurements).
D. Scope (result should have a wide ranges of validity/possibilities).
E. Simplicity (designs should be consistent with the objective of an experiment).

5.2: Intended Learning Outcomes

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

- understand the importance of hypotheses and experimental design.
- understand the sources of hypotheses and experimental design.
- distinguish the different types of hypotheses and experimental design.

5.3: Hypotheses Formulation

5.3.1: Importance of Hypotheses

a. Hypotheses provide the researcher with rational statements, consisting of elements expressed in a logical order of relationships which seek to describe or to explain conditions or events, that have not yet been confirmed by facts. The hypotheses enable the researcher to relate logically known facts to intelligent guesses about unknown conditions. It is a guide to the thinking process and the process of discovery. It is the investigator's eye – a sort of guiding light in the work of darkness.

a. Hypotheses facilitate the extension of knowledge in an area. They provide tentative explanations of facts and phenomena, and can be tested and validated. It sensitizes the investigator to certain aspects of situations which are relevant from the standpoint of the problem in hand.

b. Hypotheses provide direction to the research. It defines what is relevant and what is irrelevant. The hypotheses tell the researcher

specifically what he needs to do and find out in his study. Thus, it prevents the review of irrelevant literature and the collection of useless or excess data. Hypotheses provide a basis for selecting the sample and the research procedures to be used in the study. The statistical techniques needed in the analysis of data, and the relationships between the variables to be tested, are also implied by the hypotheses. Furthermore, the hypotheses help the researcher to delimit his study in scope so that it does not become broad or unwieldy.

c. Hypothesis has a very important place in research although it occupies a very small pace in the body of a thesis. It is almost impossible for a research worker not to have one or more hypotheses before proceeding with his work.

d. Hypotheses provide the basis for reporting the conclusions of the study. It serves as a framework for drawing conclusions. The researcher will find it very convenient to test each hypothesis separately and state the conclusions that are relevant to each. On the basis of these conclusions, he can make the research report interesting and meaningful to the reader. It provides the outline for setting conclusions in a meaningful way.

5.3.2: Sources of Hypotheses

a. Review of similar studies in the area or of the studies on similar problems;

b. Examination of data and records, if available, concerning the problem for possible trends, peculiarities and other clues;

c. Discussions with colleagues and experts about the problem, its origin and the objectives in seeking a solution.

d. Exploratory personal investigation which involves original field interviews on a limited scale with interested parties and individuals with a view to secure greater insight into the practical aspects of the problem.

e. Intuition is often considered a reasonable source of research hypotheses -- especially when it is the intuition of a well-known researcher or theoretician who “knows what is known”

f. Rational Induction is often used to form “new hypotheses” by logically combining the empirical findings from separate areas of research

g. Prior empirical research findings are perhaps the most common source of new research hypotheses, especially when carefully combined using rational induction

Thus, hypotheses are formulated as a result of prior thinking about the subject, examination of the available data and material including related studies and the council of experts.

5.3.3: Types of Hypotheses

Hypothesis can be described simply as a statement made about a population which may be true or false; and as a matter of fact, we may not know which one is it. However, on the basis of sample data and decision taking after experimentation, we either accept or reject the hypothesis. The process of accepting or rejecting hypothesis is called '***hypothesis testing'* the** aim of every hypothesis testing is to enable the researcher conclude on a more and effective decision. Thus, hypothesis testing is grouped into two categories:

Null Hypothesis
Designated by: **H_0 or H_N**. Pronounced as "H oh" or "H-null". The **null hypothesis** represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. It has serious outcome if incorrect decision is made! It assumes that non-sampling errors such as bias is absent in the measurement and that the differences is solely due to chance. Null hypothesis is always about the current state of the affairs. Statistically given as:
$H_o : X_1 = X_2$
H_o is the null hypothesis; X_1 and X_2 are the sample mean.
For example, a researcher may be interested in finding out the influence of the feeding on the weights of children according to the parental type (single versus dual parenting). The null hypothesis in this case will be there is no significant influence of feeding on the weights of the children from the single parents and the dual parents.

Alternative Hypothesis
Designated by**: H_1 or H_a.** The **alternative hypothesis** is a statement of what a hypothesis test is set up to establish. Opposite of Null Hypothesis. Only reached if H_0 is rejected. Frequently "alternative" is actual desired conclusion of the researcher! The alternative hypothesis is sometimes arbitrary. Statistically denoted as:
$H_{a:}\ X_1 < X_2$
H_a is the alternative hypothesis; X_1 and X_2 are the sample mean.
In the given example above, the alternative hypothesis will be stated as there is significant influence of feeding on the weights of the children from single and dual parents since surely, one category of the children from one parent type will have a different weight to the other category. It should be stated that the occurrence of a highly improbable difference does not disprove the null hypothesis since such a difference owing to chance will be highly unlikely. Thus, the null hypothesis will be rejected.

Working Hypothesis

The working or trail hypothesis is provisionally adopted to explain the relationship between some observed facts for guiding a researcher in the investigation of a problem. A Statement constitutes a trail or working hypothesis (which) is to be tested and conformed, modifies or even abandoned as the investigation proceeds.

5.3.4: Formulating Hypotheses

There are no precise rules for formulating hypotheses and deducing consequences from them that can be empirically verified. However, there are certain necessary conditions that are conducive to their formulation. Some of them are:

a. ***Richness of background knowledge***. A researcher may deduce hypotheses inductively after making observations of behaviour, noticing trends or probable relationships. Background knowledge, however, is essential for perceiving relationships among the variables and to determine what findings other researchers have reported on the problem under study. New knowledge, new discoveries, and new inventions should always form continuity with the already existing corpus of knowledge and, therefore, it becomes all the more essential to be well versed with the already existing knowledge. Hypotheses may be formulated correctly by persons who have rich experiences and academic background, but they can never be formulated by those who have poor background knowledge.

b. ***Versatility of intellect***: Hypotheses are also derived through deductive reasoning from a theory. Such hypotheses are called deductive hypotheses. A researcher may begin a study by selecting one of the theories in his own area of interest. After selecting the particular theory, the researcher proceeds to deduce a hypothesis from this theory through symbolic logic or mathematics. This is possible only when the researcher has a versatile intellect and can make use of it for restructuring his experiences. Creative imagination is the product of an adventure, sound attitude and agile intellect. In the hypothesis's formulation, the researcher works on numerous paths. He has to take a consistent effort and develop certain habits and attitudes. Moreover, the researcher has to saturate himself with all possible information about the problem and then think liberally at it and proceed further in the conduct of the study.

c. ***Analogy and other practices***. Analogies also lead the researcher to clues that he might find useful in the formulation of hypotheses and for finding solutions to problems. The researcher, however, should use analogies with caution as they are not fool proof tools

for finding solutions to problems. At times, conversations and consultations with colleagues and expert from different fields are also helpful in formulating important and useful hypotheses.

5.3.5: Characteristics of Good Hypotheses

a. Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
b. Hypothesis should be capable of being tested. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."
c. Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.
d. Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.
e. Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.
f. Hypothesis should be consistent with most known facts i.e. it must be consistent with a substantial body of established facts. In other words, it should be one which judges accept as being the most likely.
g. *The hypotheses selected should be amenable to testing within a reasonable time*. The researcher should not select a problem which involves hypotheses that are not agreeable to testing within a reasonable and specified time. He must know that there are problems that cannot be solved for a long time to come. These are problems of immense difficulty that cannot be profitably studied because of the lack of essential techniques or measures.
h. Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalizations, one should be able to deduce the original problem condition. Thus, hypothesis must actually explain what it claims to explain, it should have empirical reference.

5.3.6: Errors in Hypotheses

When there exists the difference between the population parameters and their estimates, the process is called 'sampling error'. In choosing between H_o and H_a the researcher can choose

correctly or incorrectly leading to an error. Thus, the two types of error in hypotheses are:

a. **Type I:** A type I error occurs when the null hypothesis (H_0) is wrongly rejected or to choose H_o when actually H_a should be chosen. For example, A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them. The probability of type 1 error is also referred to as level of significance, the size of the test or the size of the critical region.

b. **Type II:** A type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false. It is the opposite of type 1 error. For example: A type II error would occur if it were concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact they produced different ones.

There are four possible decisions in hypothesis testing. They are illustrated in the table below.

		H_o	**Decision**	**H_a**
Truth	H_o	Correct decision		Incorrect decision (**Type I error**)
	H_a	Incorrect decision (**Type II error**)		Correct decision

The hypothesis that states that there is no difference between a parameter and a specific value is?
The symbolic representation of null hypothesis is _____;

Self-Assessment Exercises - 1

1. The hypothesis that is often used in research study is?
2. The degree of difference between sample means and population mean is?
3. How many probabilities occurs when you reject null hypothesis?

5.4: Experimental Design Types

a. Preliminary experiment: this is the initial step of experimental design.
b. Formal experiment: this can be divided into;
i. Simple experiment: this is concerned with one factor and its variation.
ii. Factorial experiment: this is the investigation of the effect of more than one factor simultaneously to reveal the effect of interaction.

Components

a. Control group: a change in the dependent variable could result in a change in the dependent variable, but a change in the dependent variable could also be the result of a random variable which may find its way into the experiment thus to ensure this does not happen, a controlled experiment is run:

a. Control group is an additional experimental run or treatment.

b. It is a separate experiment, set up like the others with exactly the same conditions.

c. The only difference is that test variable is changed.

d. In an experiment, a control is a treatment which is included to provide a reference set of data which can be compared with data obtained from the experimental treatments. For example, an investigation into the effect of increased copper in the water in fish growth would have as a control a group of fish cultured in copper free growth medium. The effect of copper can be determined by comparing the growth rates of fish in growth medium with various levels of copper with the control. For this comparison to be valid, it is critical there are no other variables apart from the independent variable that differs between the control and experimental groups.

b. Variables: in designing an experiment it is important to consider all the factors that can change in some way. These are called variables. There are three types of variables in an experimental design:

a) Independent variability: it can also be referred to as the manipulated or the variable which is deliberately altered in some way by the person carrying out the investigation.

b) Dependent variable: it can also be referred to as responding variable or the variable which is measured/directly observed by the experimenter.

c) Fixed/controlled variables: the major types of fixed/controlled variables that needs to be controlled are:

✓ Factors affecting the organism: if organisms are used in the experiments, then many of the variables relating to them will need to be controlled. If the investigation uses small numbers in the same size, then all due care needs to be taken to ensure that the control group and experimental group are matched as closely as possible with regards to things such as the number of organisms, their age, size, sex and any other relevant characteristics. However, if the sample size is large then these variables often average themselves out.

✓ The environment factors: in all the experiments the environmental factors will need to be controlled as much as possible. This is much easier to achieve if the investigation is carried out in a laboratory environment but not as easy in a field situation. For instance, in the

laboratory situation factors such as temperature, light, dosage of a chemical and the amount of food and water administered can be controlled easily. In field situation these factors may have diurnal or seasonal variation descending on the length of the experiment.

c. Replication: replication is having more than one treatment at the same type. It is important in that:

a) Replication with a treatment shows how variable the response can be

b) Resources may limit the total number of replicates. If this occurs then a compromise between the number of treatments and the number of replicates within a treatment has to be found.

c) Depending/the statistics used in analysis it is possible to work out the number of replicated needed to show a significant difference between pairs of means.

d) Statistics play a vital role in the analyzing of results obtained from investigations. The experimental in scientific research is very important for reliable result to be obtained.

d. Sample size: in most experiments it is rarely possible to take measurements from every individual in the population either in a laboratory situation or in the field. A sub set or sample is used to estimate the values that might have been obtained had we measured every individual in the population. A sample is made up of a series of sampling units which depends on the type of variable being measured.

Characteristics of a good sample

Good sampling design should take into account both of these and should

a. Relate to the objectives of the investigation

b. Be practical and achievable

c. Be cost effective in terms of equipment and labour

d. Provide estimates of population parameters that are truly representative and unbiased

e. An ideal representative samples should be:

a) Taken at random so that every member of the population of data has an equal chance of selection.

b) Large enough to give sufficient precision

c) Unbiased by the sampling procedures or equipment

It is very important in sampling procedures to take into account relevant factors such as: location, habitat, time, age, sex, physiological condition and disease status. These also need to be noted in the design as otherwise a wrong interpretation may arise from the result.

Errors in Experiments

There are many potential sources of error when designing and carrying out experiments. Error can arise from:

a. The design of the experiment
b. The measurement and sampling of data
c. Measurement error can arise for a number of reasons:

a) Instrument error: calibration of the instrument has not been carried out or is faulty consequently accuracy and precision are affected
b) Personal error: observer making in accurate observations. This type of error can be overcome by taking an average measurement, especially if data is collected by two or more independent observers.
c) Sampling errors: these can also arise because of the size or nature of the sample used. Sample size can either be too small or not random enough. Replication of experiment also reduces errors.

The term precision is the closeness of repeated measurements to each other, which accuracy is the closeness of a measured or derived data value to its true value.

Types of Experimental Design

a. Complete randomize experimental design (single factor analysis of variance). This is a design in which treatments are assigned completely at random to the experimental units. The randomization gives every group of units an opportunity of receiving the treatment. It is known as *one way classification* because the data are classified according to one criterion viz treatment only. For example, the Complete randomize design for three replicates of six treatments is as follows.

A	B	A
A	E	B
D	D	E
C	C	C
F	F	E
F	B	D

Advantages of CRD

i. The design is very flexible and can be used for any number of treatments.
ii. The statistical analysis is comparatively easy and straightforward.
iii. It is unaffected by missing observations for any treatment for some purely random accidental reason.

Disadvantage of CRD

i. The design is inherently less informative than other more sophisticated layouts.

b. Randomize block design: this design compares treatments in the plots with each treatment replicated. The plots are divided into groups or block such that the blocks are as uniform as possible and the known sources of variation among the plots being assigned to the blocks. Data from it can be classified in a two-way table, the rows are blocks and the columns are treatments vice versa hence called *a two-way classification.*

Advantages and Disadvantages of Randomize Block Design

1. With heterogeneous material the residual variance can be reduced by choosing blocks of plots such that the plots within the blocks are fairly similar. i.e., the design reduces the effect of heterogeneous material.
2. There is no restriction on the number of blocks or treatments, but in each block, there must be the same number of plots, one to each treatment.
3. If some yields are accidentally lost, the analysis is again without due complications, although special modifications are required.

c. Latin Square Design: the treatments are arranged in complete groups of two durations, the two classifications being orthogonal to each other and to the treatment. Latin square should be not less than 5 x 5 and not more than 10 x 10 where it becomes clumsy and difficult to manage. The design is a special case of a 3-way classification, with rows, column and treatments. The randomization of a Latin square may be carried out as follows;

a. Permute the rows at random
b. Permute the columns at random
c. Allot the treatments at random to the letter A, B, C etc.

The phenomena that deal with the methods of constructing and analysing comparative experiment is called? The variation among varieties that cannot be accounted for by the variation due to treatment is called?

Self-Assessment Exercises - 2

1. The phenomena used to describe the ability to estimate the effect of the treatment so that valid conclusion can be drawn is called?
2. The type of experimental design that can be applied when same experimental material is used on different experimental unit is called?
3. The experimental design in which the total area is divided into blocks and all treatments are arranged within each block in a random order is?

5.5: Summary

In this Unit the essence, of hypotheses were discussed with the various types. The Unit also examined the different components, types of experimental designs and their appropriate uses. This Unit looks at the importance of hypotheses, sources, types characteristics and the methods of formulating hypotheses. In the text, students will be able to learn how to design experiment and take into account the different components in the experiment as well as the possible statistical tools required.

5.6: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Helmut F. van Emden.(2008). *Statistics for Terrified Biologists.* Blackwell Publishing Limited. USA.

Link for Textbooks/Journals

Hypothesis Formulation in Research - Reading Craze
https://readingcraze.com/index.php/hypothesis-formulation-research/

Formulating the Research Hypothesis and Null Hypothesis
https://study.com/academy/lesson/formulating-the-research-hypothesis-and-null-hypothesis.html

Experimental Design - an overview | ScienceDirect Topics
https://www.sciencedirect.com/topics/earth-and-planetary-sciences/experimental-design#:~:text=Experimental%20design%20is%20the%20process,has%20on%20a%20dependent%20variable.

A Quick Guide to Experimental Design | 5 Steps & Examples
https://www.scribbr.com › ... › Methodology

https://www.scribbr.com/methodology/experimental-design/

https://www.scribbr.com/methodology/experimental-design/

https://study.com/academy/lesson/formulating-the-research-hypothesis-and-null-hypothesis.html

5.7: Possible Answers to Self-Assessment Exercises

Self-Assessment Exercises - 1

1. Null hypothesis
2. Significance difference
3. Five

Self-Assessment Exercises - 2

1. Sensitivity
2. Completely randomized design
3. Randomized block design

Glossary

- **Alternate hypothesis:** the opposite of the null hypothesis. It is the conclusion when the null hypothesis is rejected.
- Bar chart or Bar graph: a chart or graph used with nominal characteristics to display the numbers or percentages of observations with the characteristics of interest.
- Bell-shaped distribution: a term used to describe the shape of the normal (Gaussian) distribution
- Bias: A systematic error
- Biostatistics: the application of research study design and statistical analysis to application in life sciences.
- Categorical variable: A variable having only certain possible values for which there is no logical ordering of the values. Also called a nominal, polytomous, discrete categorical variable or factor.
- Class limits: the subdivisions of a numerical characteristics (or the widths of the classes) when it is displayed in a frequency table or graph
- Continuous variable: A variable that can take on any number of possible values.
- **distribution**: he values of a characteristic or variable along with the frequency of their occurrence. Distributions may be based on empirical observations or may be theoretical probability distributions (eg, normal, binomial, chi-square).
- Estimate: A statistical estimate of a parameter based on the data.
- **Estimation**: The process of using information from a sample to draw conclusions about the values of parameters in a population.
- **frequency polygon**: A line graph connecting the midpoints of the tops of the columns of a histogram. It is useful in comparing two frequency distributions.
- **frequency table**: A table showing the number or percentage of observations occurring at different values (or ranges of values) of a characteristic or variable.
- **Histogram**: A graph of a frequency distribution of numerical observations.
- **hypothesis test**: An approach to statistical inference resulting in a decision to reject or not to reject the null hypothesis
- **modal class**: The interval (generally from a frequency table or histogram) that contains the highest frequency of observations.
- null hypothesis: Customarily but not necessarily a hypothesis of no effect.
- **percentage polygon**: A line graph connecting the midpoints of the tops of the columns of a histogram based on percentages instead of counts. It is useful in comparing two or more sets of observations when the frequencies in each group are not equal.

- probability: The probability that an event will occur, that an invisible event has already occurred, or that an assertion is true, is a number between 0 and 1 inclusive such that (1) of all possible outcomes (including non-events) the probability of some possible outcome occurring is 1, and (2) the probability of any of a set of mutually exclusive events (i.e., union of events) occurring is the sum of the individual event probabilities

End of the module Questions

1. The collection of quantitative information/data on living things is called?
2. The important aspect of biostatistics that describe on how to collect, organize and analyse data is called?
3. The type of variable exemplified by classifying students into gender is?
4. The type of variable exemplified by the nature of children in a family is called?
5. The variable dealing with all values are called?
6. The data below represents the blood groups of 40 students in a Biostatistics class. Construct a frequency distribution for the data.

A	AB	B	O	O	A	B	AB	A	B
O	O	O	A	AB	B	B	A	O	AB
A	O	O	A	AB	B	B	A	A	B
AB	A	O	B	AB	O	A	B	A	B

7. Given below, are the wing length measurements (to the nearest whole millimeter) of 50 laughing doves.

76	73	75	73	74	74	72	75	76	73
68	72	78	74	75	72	76	76	77	70
78	72	70	74	76	75	75	79	75	74
75	70	73	75	70	74	76	74	75	74
78	74	75	74	73	74	71	72	71	79

Construct a frequency distribution table.

8. A flower is drawn at random from garden containing 6 red flowers, 4 white flowers and 5 blue flowers, determine the probability that it is a red flower?

9. A flower is drawn at random from garden containing 6 red flowers, 4 white flowers and 5 blue flowers, determine the probability that it is a white flower?

10. A flower is drawn at random from garden containing 6 red flowers, 4 white flowers and 5 blue flowers, determine the probability that it is a blue flower?

11. A flower is drawn at random from garden containing 6 red flowers,

4 white flowers and 5 blue flowers, determine the probability that it is not red flower.

12. A flower is drawn at random from garden containing 6 red flowers, 4 white flowers and 5 blue flowers, determine the probability that it is a red or white flower?

13. A special probability distribution that describes the distribution of probabilities when there are only two possible outcomes for each trial experiment is?

14. The sampling technique that increases precision is?

15. How many advantages exist between cluster sampling and others?

16. The sampling technique that is more current is?

17. The sampling technique that involves taking an item as a sample from a large population at regular interval is?

18. When a large population is given a questionnaire to determine those who meet the qualification for a study is called

19. The type of error that occurs when one rejects the null hypothesis when it is true is?

20. The experimental design in which the number of rows, columns and treatments are equal and each treatment occurs just once in each row and column is termed?

Answers

1. Biostatistics
2. Experimental design
3. Qualitative variables
4. Discrete variable
5. Continuous variable
6. Since the data are categorical, the blood groups: A, B, O and AB can be used as the classes for the distribution.

Class	**Tally**	**Frequency**	**Percent**
A	///,///,//	12	30
B	///,////,/	11	27.5
O	///,///	10	25
AB	///,//	7	17.5
TOTAL			**100**

Therefore, it can be concluded that in the sample more students have **type A** blood group because its frequency is the highest.

7. The measurements above are presented in the order in which the observations were recorded. This can be represented in an ordered array so that the minimum and maximum values can easily be read.

68 70 70 70 70 71 71 72 72 72

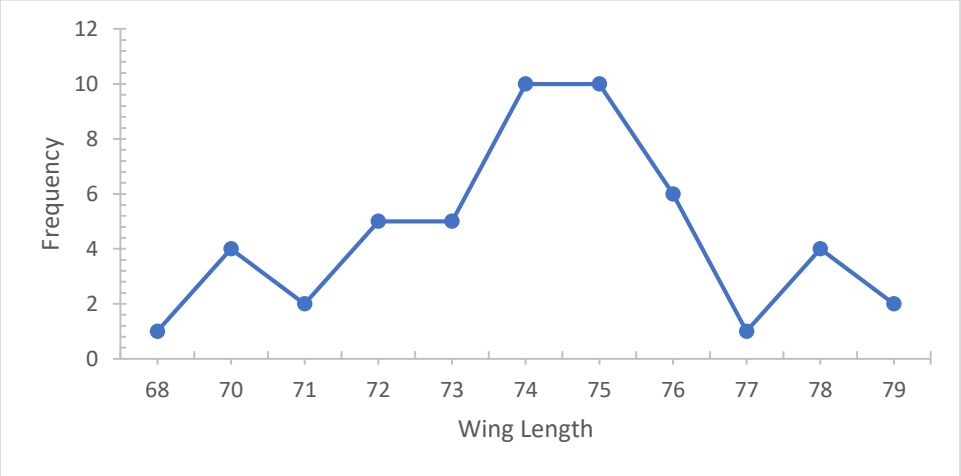
72	72	73	73	73	73	73	74	74	74
74	74	74	74	74	74	74	75	75	75
75	75	75	75	75	75	75	76	76	76
76	76	76	77	78	78	78	78	79	79

Find the range of the data: *Highest value – lowest value* (79 – 68 =11). Since the range of the data is small, classes of single data values can be used.

Table 2.1: A tally of frequency of the wing length (mm) of 50 laughing doves.

Class limits	**Tally**	**Frequency**	**Cumulative frequency**	**Relative frequency (%)**
68	/	1	1	2
70	////	4	5	8
71	//	2	7	4
72	~~////~~	5	12	10
73	~~////~~	5	17	10
74	~~////~~, ~~////~~	10	27	20
75	~~////~~, ~~////~~	10	37	20
76	~~////~~,/	6	43	12
77	/	1	44	2
78	////	4	48	8
79	//	2	50	4
				100

Frequency distribution of wing-length



8. 0.4
9. 0.27
10. 0.33
11. 0.6
12. 0.67
13. Binomial
14. Stratified sampling
15. Three
16. Cluster
17. Systematic, Skip
18. Double sampling
19. Type I error
20. Latin square design

Module 2: Biostatistics Application 1

Unit 1: Measure of Central Tendency

Unit Structure

1.1: Introduction

A measure of central tendency also called measure of location indicates the location or position of the sample or population among all possible values of the variable. These measures are descriptive parameters in that they describe a property of population. There are various measures of central tendency but the most commonly used ones are discussed here. A measure of central tendency also called measure of location indicates the location or position of the sample or population among all possible values of the variables. These measures are descriptive parameters in that they describe a property of populations.

1.2: Intended Learning Outcomes

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

- describe the different types of measures of central tendencies.
- understand the usage of the different types of central tendencies

1.3: Measures of Central Tendency

1.3.1: Characteristics

a. It should be based on all the observations.
b. It should not be affected by extreme values.
c. It should be as close to the maximum number of observed values as possible.
d. It should be defined rigidly which means that it should have a definite value.

e. It should not be subjected to complicated and tedious calculations.
f. It should be capable of further algebraic treatment.
g. It should be stable with regard to sampling. This means that if a number of samples of the same size are drawn from a population, the measure of central tendency having minimum variation among the different calculated values should be prepared.

1.3.2: Types

Arithmetic Mean/Average

If mean is mentioned, it implies arithmetic mean as the other means are identified by their full names. Mean can be defined as the sum of the observed values of a set divided by the number of observations in the set. If X_1, X_2…..Xn are N observed values, the mean

Mean = $(\sum \boldsymbol{FX})/(\sum \boldsymbol{F})$

For weighted mean, in case K variate values X_1, X_2….X_k have known weight W_1, W_2…W_k respectively then the weighted mean (μ) = $(W1X1 + W2X2 + \cdots WkXk)/(W1 + W2 + \cdots Wk)$ = $(\sum WX)/(\sum W) = 1/W \sum WiXi$ i=1,2...k

Weighted mean is commonly used in the construction of index numbers.

Properties

i) The algebraic sum of the deviations of a set of numbers from their arithmetic mean is zero.

ii) The sum of the square of the deviations of a set of numbers from any number above is a minimum if and only if a is equal to mean.

iii) If Fi numbers have Yi and Fk has Yk, then the mean of all the number is $(F1Y1 + F2Y2 + \cdots FkYk)/(F1 + F2 + \cdots Fk)$

Merits and Demerits

a) The algebraic sum of the deviations of the given values from their arithmetic mean is always zero i.e. $\sum [\![(X - /X)]\!] = 0$

b) The sum of the squares of the deviations of the given values from their arithmetic means is minimum i.e. $\sum [\![(X - /X)]\!]^2$ is minimum.

c) An average possesses all the characteristics of a central value given earlier except no.2 which is greatly affected by extreme values.

d) In case of grouped data if any class interval is open, arithmetic mean cannot be calculated

Median:

It has been pointed out that mean cannot be calculated whenever there is frequency distribution with open end intervals. Also, the mean is to a great extent affected by the extreme values of the set of observations. Hence in such cases, there has been a search for some better measure of central tendency. Median is the value of the variable which divides it into two equal halves; in an order series of data, median is an observation lying in the middle of the series, in a set of observations below it and remaining

half above it. The median for a set of observations can easily be found out after arranging them in either descending or ascending order.

Let X_1, X_2… Xn be N ordered observations. Now two possibilities are there:

a) N is an odd number say N=2p +1 where p is an integer. In this case (p+1)th observation will be the median value.

b) If N is even N =2p, then the average of pth and (p+1)th observations will be the median value.

The median for grouped data: if the data are given with class interval as:

Class interval	**Frequency**	**cumulative frequency**
$\mathbf{X_1 - X_2}$	$\mathbf{F_1}$	$\mathbf{F_1}$
$\mathbf{X_2 - X_3}$	$\mathbf{F_2}$	$\mathbf{F_2}$
.	.	.
.	.	.
$\mathbf{X_p - X_{p+1}}$	F_p	F_p
.	.	.
$X_k - X_{k+1}$	F_k	F_k

Where $F_k = N = \sum Fi$ for i = 1,2…..K, we can thereof calculate the median by

$$\mathbf{M_d = Lo + (}\frac{\frac{N}{2}-(\sum f)1}{fmedian}\mathbf{)}C$$

Where Lo = Lower class boundary of the median class i.e. lower limit of interval (the clas containing the median)

N = number of observation (i.e. total frequency)

$(\sum f)1$ = sum of frequencies (cumulative frequency) of all classes lower than the median class.

C = size of the median class interval i.e. the range of accuracy.

Fmedian = frequency of median class

I = Class interval of the median class

Merits and Demerits

i) Median is a positional average and hence it is not influenced by the extreme value.

ii) Median can be calculated even in the case of open and end intervals.

iii) Median can be located even if the data are incomplete.

iv) It is not a good representative of data if the number of items is small.

v) It is not amenable to further algebraic treatment.

vi) It is susceptible to sampling fluctuations.

Mode

Mode is a value of a particular type of items which occurs most frequently. It can be defined as a variate value which occurs most frequently in a set of values. In case of discrete distribution one can find

mode by inspection. The variate value having the maximum frequency is the modal value.

Merits and Demerits

i) It is not affected by extreme values of a set of observation
ii) It can be calculated for distributions with open end classes
iii) The main drawback of mode is that often it does not exist
iv) Often value is not unique
v) It does not fulfill most of the requirements of a good measures of central tendency.

The grades of a student on six examinations were 84, 91, 72, 68, 87 and 78 find the arithmetic mean of the grades.

The grades of a student on six examinations were 84, 91, 72, 68, 87 and 78 find the total frequency of the grades.

Self-Assessment Exercises

1. The measures of the diameter of a plant leaf were 38.8, 40.9, 39.2. 39.7, 40.2, 39.5, 40.3, 39.2, 39.8, 40.6mm, find the number of measures?
2. The measures of the diameter of a plant leaf were 38.8, 40.9, 39.2. 39.7, 40.2, 39.5, 40.3, 39.2, 39.8, 40.6mm, find the total number of frequency?
3. The measures of the diameter of a plant leaf were 38.8, 40.9, 39.2. 39.7, 40.2, 39.5, 40.3, 39.2, 39.8, 40.6mm, find the arithmetic mean?
4. Find the mean of the set of figures obtained from measuring the weight to nearest gram of fingerlings: 3, 5, 2, 6, 5, 9, 5, 2, 8, 6

1.4: Summary

The different measures of central tendency were discussed with their advantages and disadvantages. The characteristics of measures of central tendency were discussed, the different types, their advantages and disadvantages were also discussed.

1.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp

- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp

Mean, Mode and Median - Measures of Central Tendency
https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

Measures of Central Tendency: Mean, Median, and Mode
https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

https://www.thoughtco.com/measures-of-central-tendency-3026706

https://study.com/academy/lesson/central-tendency-measures-definition-examples.html

1.6: Possible Answers to Self-Assessment Exercises

1. 10mm
2. 398.2mm
3. 39.8mm
4. 5.1

Unit 2: Measure of Dispersion/Variability

Unit Structure

2.1: Introduction

A measure of dispersion or variability is an indication of the spread of measurement around the centre of the distribution. It shows how variable the measurements are. In addition to the mean, they are used in summarizing set of data. Examples of such measures are range, variance, standard deviation, standard error and coefficient of variation or variability. Measures of dispersion or variability are indication of the spread of measurement around the center of the distribution. It shows how variable the measurements are. In addition to the mean, they are used in summarizing set of data. Example of such measures includes the **range**, standard deviation, standard error, variances and coefficient of variation.

2.2: Intended Learning Outcomes

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

- describe the different types of measures of dispersion.
- understand the usage of the different types of dispersion

2.3: Measures of Dispersion

2.3.1: Purposes

a) To have an idea about the reliability of central value. In a way, it is a measure of degree of scatteredness. If scatter is large, an average is less reliable. If the value of dispersion is small, it indicates that a central value is a good representative of all the values in the set

b) To compare two or more sets of values with regard to their variability: two or more sets can be compared by calculating the

same measure of dispersion having the same unit of measure. A set with smaller value possesses lesser variability

c) To provide information about the structure of a series: a value of measure of dispersion gives an idea about the spread of the observations.

d) To control the variation: in many situations a measure of dispersion provides the basis for controlling the causes which lead to greater variation.

e) To pave way to the use of other statistical measures: measures of dispersion especially variance and standard deviation lead to many statistical techniques like correlation, regression, analysis of variation etc.

Properties

i) It should be based on all values of a series.

ii) It should not be susceptible to fluctuation of sampling.

iii) It should be rigidly defined i.e.; each investigator should arrive at the same value for the same set of data.

iv) It should be capable of further algebraic treatment.

v) It is preferable that the unit of measurement of dispersion should be the same as the unit of measurement of observations

vi) It should be calculable with reasonable ease i.e., the formular should be such that it does not complicate the computation of a measure of dispersion.

vii) It should be least affected by extreme values.

2.3.2: Types

Range

Range is defined as the difference between the largest and the smallest observation in a set

Range (R) = L – S

Where L = largest observation

S = Smallest observation

A relative measure known as coefficient of range is given as

Coefficient of Range = $(L - S)/(L + S)$

The lesser the range or coefficient of range, the better the result

Properties

a) It is the simplest measure and can easily be understood.

b) Besides the above merit, it hardly satisfies any property of a good measure of dispersion e.g., it is based on two extreme values only, ignoring the others.

c) It is not liable to further algebraic treatment.

Variance (δ^2, S^2)

The variance is the average of the squares of deviations taken from mean population variances (δ^2) =(∑▒ 〖(Xi〗 − μ)^2)/N where μ is replaced by /X and N replaced by n -1 (called the degree of freedom) in the above equation, it is an unbiased estimate of δ^2 which is called the sample variances thus:

Sample variance = (∑▒ 〖(Xi〗 −/X)^2)/(n − 1) = SS/(n − 1) = (∑▒ 〖X^(2)− ((∑▒X)^2)/n〗)/(n − 1)

Properties

a) It has mostly removed the lacunae which are present in the measures of dispersion given before it.
b) The main demerit is that its unit is the square of the unit of measurement of variate values. E.g., the variable X is measured in cm, the unit of variance is cm^2
c) The variance gives more weight to the extreme values as compared to those which are near to mean value, because the difference is squared in variance

Standard deviation

The positive square root of the variance is called deviation (S.D) = √(δ^2) = √(S^2)In simple words, we can say that standard deviation explains the average amount of variation on either side of the mean. It has the same S.I unit as the measurement.

Properties

a) It is considered to be the best measure of dispersion and is used widely.
b) There is however, one difficulty with it. If the unit of measurement of variables of two series is not the same, then their variability cannot be compared by comparing the values of standard deviation.

Coefficient of variation (CV)

Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100.

C.V = δ/mean x 100

Properties

a) It is one of the most widely used measured of dispersion because of its virtues.
b) Smaller the value of C.V than the C.V of other series is more consistent i.e.; it has less variability.
c) For field experiment C.V is generally reported. If C.V is low, it indicates more reliability of experimental findings.

d) It is a relative measure of variability

Standard Error

It tells how close the values of means are to the population mean S.E = square root of variance divides by the frequency of sampling (n). The unit of the standard error is the same as the unit of the individual measurements

Which of the following are methods under measures of dispersion?

a. Standard deviation
b. Mean deviation
c. Range
d. All of the above

Which of the following are characteristics of a good measure of dispersion?

a. It should be easy to calculate
b. It should be based on all the observations within a series
c. It should not be affected by the fluctuations within the sampling
d. All of the above

Self-Assessment Exercises

1. **If all the observations within a series are multiplied by five, then _____**
    a. The new standard deviation would be decreased by five
    b. The new standard deviation would be increased by five
    c. The new standard deviation would be half of the previous standard deviation
    d. The new standard deviation would be multiplied by five
2. **The coefficient of variation is a percentage expression for _____.**
    a. Standard deviation
    b. Quartile deviation
    c. Mean deviation
    d. None of the above
3. **While calculating the standard deviation, the deviations are only taken from _____**
    a. The mode value of a series
    b. The median value of a series
    c. The quartile value of a series

2.4: Summary

The measures that determine the closeness of values to the centre were considered in this Unit. The different types of measures of dispersion were considered and treated in this Unit.

2.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp

Measures of dispersion
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198538/#:~:text=Standard%20deviation%20(SD)%20is%20the,by%20the%20number%20of%20observations.

Measures of Dispersion in Statistics (Definition & Types) - Byju's
https://byjus.com/maths/dispersion/

Measures of Dispersion - Definition, Formulas, Examples
https://www.cuemath.com/data/measures-of-dispersion/

https://www.toppr.com/guides/business-mathematics-and-statistics/measures-of-central-tendency-and-dispersion/measure-of-dispersion/

https://www.youtube.com/watch?v=YvGeUSeQGYU

https://www.youtube.com/watch?v=dAwRlYhEWOs

2.6 Possible Answers to SAEs

1. : d
2.: a
3. d
4. d

Unit 3.0: Student's t-Distribution

Unit Structure

3.1: Introduction

The t distributions were discovered by William Sealy Gossets in 1908 under the pseudonym 'student', hence it is commonly referred to as Student's t-distribution or Student's t-test. This has revolutionary statistics of small samples. The student's *t* test (also called T test) is used to compare the means between two groups and there is no need of multiple comparisons as unique *P* value is observed. A *t*-test is used to infer on statistical grounds whether there are differences between group means for an experimental design with;

(i) one parametric dependent variable and

(ii) one independent variable with two levels, i.e., there is one outcome measure and two groups.

It is one of the most popular statistical techniques used to test whether mean difference between two groups is statistically significant. T test are three types i.e., one sample *t* test, independent samples *t* test, and paired samples *t* test. The t-distribution is characterized by the following properties, which are similar to that of a standard normal distribution;

- It is bell-shaped.
- It is symmetric about the mean
- The mean, median and mode are equal to 0 and are located at the center of the distribution.
- The curve never touches the *x*-axis.
- The properties that differentiates it from the standard normal distribution are;.
- The variance is greater than 1.
- The t-distribution is a family of curves based on the sample size (*n*), with the degrees of freedom *n*-1.
- As the sample size increases, the t-distribution approaches the normal distribution.

3.2: Intended Learning Outcomes

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

- describe the different types of t-test
- understand the circumstance of using the different types of the test

3.3: Student's test

3.3.1: One Sample t-test

The one sample t test is a statistical procedure used to determine whether mean value of a sample is statistically same or different with mean value of its parent population from which sample was drawn. To apply this test, mean, standard deviation (SD), size of the sample (Test variable), and population mean or hypothetical mean value (Test value) are used. Sample should be continuous variable and normally distributed. One-sample *t* test is used when sample size is <30. In case sample size is ≥30 used to prefer one sample z test over one sample *t* test although for one sample z test, population SD must be known.
If population SD is not known, one sample *t* test can be used at any sample size. In one sample Z test, tabulated value is z value (instead of *t* value in one sample *t* test).

3.3.2: Independent Sample t-test

The independent *t* test, also called unpaired *t* test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated (independent) groups? To apply this test, a continuous normally distributed variable (Test variable) and a categorical variable with two categories (Grouping variable) are used. Further mean, SD, and number of observations of the group 1 and group 2 would.

3.3.3: Paired Sample t-test

The paired samples *t* test, sometimes called the dependent samples *t*-test, is used to determine whether the change in means between two paired observations is statistically significant? In this test, same subjects are measured at two time points or observed by two different methods. To apply this test, paired variables (pre-post observations of same subjects) are used where paired variables should be continuous and normally distributed. Further mean and SD of the paired differences and sample size (i.e., no. of pairs) would be used to calculate significance level.

***t*-test:** tn-1 = $(\mathbf{x} - \boldsymbol{\mu})/(\mathbf{s}/\sqrt{\mathbf{n}})$
Where: x = Sample mean.
μ = Population mean.
s = Sample standard deviation.
n = Sample size.
Testing of hypotheses using the t-test follows the same procedure as for the z-test, except that you use the t- table instead.
When the variance is greater that 1, it is a characteristics of ……test.
Testing of hypothesis using the t-test follows the same procedure as for the z-test except that you use the ……. instead.

Self-Assessment Exercises

i)

1. Which test approaches the normal distribution as sample size increases?
2. The inappropriate test for testing hypothesis when the sample size is less than 30 is?

3.4: Summary

This Unit is able to explain the simplest method of comparing two sets of variables. The introduction the Unit was provided as well the different types of the distribution. The rationale for the application of the different types were also provided.

3.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Helmut F. van Emden.(2008). *Statistics for Terrified Biologists.* Blackwell Publishing Limited. USA.

Student's t-distribution https://www.investopedia.com/terms/t/tdistribution.asp#:~:text=T

he%20T%20distribution%2C%20also%20known,distributions%2C%20hence%20the%20fatter%20tails.

T-Distribution / Student's T: Definition, Step by Step Articles ...
https://www.statisticshowto.com/probability-and-statistics/t-distribution/

T Distribution (Definition and Formula) - Byju's
https://byjus.com/maths/t-distribution/

https://www.statisticshowto.com/probability-and-statistics/t-distribution/

https://www.youtube.com/watch?v=32CuxWdOlow

https://study.com/academy/lesson/student-t-distribution-definition-example-quiz.html

3.6: Possible Answers to Self-Assessment Exercises

1. t-distribution
2. z-test

Unit 4: Contigency Table

Unit Structure

4.1: Introduction

Contingency tables are usually constructed for the purpose of studying the relationship between two or more variables of classification. One may wish to know whether the two variables are independence or there is association between them. By means of chi square. Chi-square ($\chi 2$) is the general method for testing compatibility based on a measure of the extent to which the observed and expected frequencies agree. Chi-square is also, referred to as test for homogeneity randomness, association, independence and goodness of fit.

4.2: Intended Learning Outcomes

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

- understand the principles behind the use of chi-square.
- describe the methods of using the test tool.
- understand the application of the tool.

4.3: Chi-Square

Chi-square (χ^2) is the general method for testing compatibility based on a measure of the extent to which the observed and expected frequencies agree. Chi- square is also, referred to as test for homogeneity randomness, association, independence and goodness of fit. The assumptions for the chi-square goodness- of-fit test are:

- The data are obtained from a random sample.
- The expected frequency for each category must be 5 or more.

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Where o_i and e_i denote the observed and expected frequencies, respectively, for the ith cell, and k denotes the number of cells.

The frequencies we *observe* are compared to those we *expect* on the basis of some null hypotheses. If the differences between the observed and expected frequencies are great and exceed the critical value at appropriate degrees of freedom, we are then obliged to the reject the null hypothesis (H_O) and accept the alternative hypothesis (H_1).

4.3.1: Properties of Chi-square

a) It is concerned with the sequences of normally distribution observations, thereby estimating the variance.
b) Chi-square values tend to become normal as the sample size (n) increases. Its degree of freedom is n - 1.
c) It is non-symmetrical.
d) $\chi 2$ can take any value from zero to in infinity.
e) $\chi 2$ is additive and always positive.
f) Chi-square can be:

- One-way classification as in testing the goodness of fit of a hypothesis, with two or more classes.
- Two-way classification as in determining association or differences between two different classes or the test may involve more than two classes.

4.3.2: Chi-Square testing

Chi-square distributions are used in a procedure that involves the comparison of the differences between the *sample frequencies* of occurrences or percentages that are *actually observed* and the hypothetical or theoretical *population frequencies* of occurrences or percentages that are expected if the hypothesis is true. Steps in the general $\chi 2$ testing procedure is;

a. Formulate the null and alternative hypotheses.
b. Select the level of significance to be used in the particular testing situation.
c. Take random samples from the populations, and *record the observed frequencies* that are actually obtained.
d. Compute the frequencies of percentages that would be *expected* if the null hypothesis is true.
e. Use the observed and the expected frequencies to compute the **$\chi 2$.**

f. Compare the value of $\chi 2$ computed in step 5 with the $\chi 2$ table value at the specified level of significance (step 2).

The general method for testing compatibility based on a measure of the extent to which the observed and expected frequencies agree is?
The assumption that data are obtained from a random sample is?

Self-Assessment Exercises

1. In an experiment to test the effectiveness of three different traps for catching birds, the number of birds captured in each trap design over the study period was recorded as follows:

Design	Observed Frequency
A	10
B	**27**
C	15
Total	52

2. The offspring of a certain cross gave the following colours: Red, Black or white in the ratio 9:3:4. Assuming the experiment gave 74, 32, and 38 offspring respectively in those categories, is the theory substantiated?

4.4: Summary

The Unit explain another statistical tool used to compare two variables. Chi-square significance, methods and application was discussed in this Unit.

4.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Bluman, A.G. (2004). *Elementary Statistics*. A Step by Step Approach. Fifth Edition. McGraw-Hill Companies Incorporated. London.

Chi-squared test: https://www.investopedia.com/terms/c/chi-square-statistic.asp#:~:text=A%20chi%2Dsquare%20(%CF%872)%20statistic%20is%20a%20measure,especially%20those%20nominal%20in%20nature.

The Chi squared tests - The BMJ
https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests
https://www.statisticshowto.com/what-is-a-contingency-table/

https://study.com/academy/lesson/contingency-table-statistics-probability-examples.html

https://www.youtube.com/watch?v=9KIQC9Npndg

4.6: Possible Answers to Self-Assessment Exercises

1. The *observed frequencies* are, of course, 10, 27 and 15. The question we are interested in is "Does the observation that more birds were caught in trap design B really reflect a genuine difference, or could the difference be due to chance scatter or sampling error?" Another question could be "Is the distribution of frequencies between the traps homogenous (i.e evenly spread)?" From these, the hypotheses are:
H_O: The observed frequencies are homogenous and any departure can be accounted for by chance scatter or sampling error;
H_1: The observed frequencies depart from those expected of a homogenous (even) distribution by an amount that cannot be explained by sampling error.
Then, what frequencies would have been *expected* if H_O is indeed true. That means if the frequencies reflect a homogenous distribution, we would expect the 52 birds to be equally distributed in all the three trap designs. That is 52/3 =17.33 birds in each design is our *expected frequency*. To calculate χ^2, we can summarize our frequencies as shown below.

Sample	Frequency (observed)	Frequency (expected)	Difference	Chi-square
A	10	17.33	-7.33	3.10
B	27	17.33	9.67	5.40
C	15	17.33	-2.33	0.313
				χ^2 = 8.813

The calculated χ^2 = 8.813 can then be compared with the critical or table χ^2 value at 0.05 or 0.01 levels of significance. The degrees of freedom = n – 1 ⇒ 3 – 1 = **2.** From the χ^2 table at 2 df, we have 5.99 under the 0.05 (5%) level of significance and 9.21 under the 0.01(1%). Our calculated χ^2 value of 8.813 is bigger than the first but smaller than the second. We conclude that the difference between the observed and expected frequencies is statistically 'significant' but not 'highly significant'. This simply means that **"the trap of design B was shown in a trial to be more effective in catching birds than the other two traps tested**; $\chi^2_{(2d.f)}$ =8.813, *P*<0.05".

2.

Red	Black	White	**Total**
74	32	38	**144**
9	3	4	**16**

The expected frequencies are calculated as follows:

Red: 9/16 x 144 = 54
Black: 3/16 x 144 = 27
White: 4/16 x 144 = 36

$$\chi^2 = \frac{(74-54)^2}{54} + \frac{(32-27)^2}{27} + \frac{(38-36)^2}{36} = 8.45$$

χ^2 = 8.45
The D.F n – 1
3 – 1 = 2

The calculated $\chi 2$ of 8.45 is higher than the table χ^2 value of 5.99 at α= 0.05 and d.f 2. Therefore, we conclude that the number of offspring in the 3 colours is not compatible with the given ratios. i.e. we reject the null hypothesis.

Unit 5: Analysis of Variance and Co-Variance

Unit Structure

5.1: Introduction

Analysis of variance (ANOVA) is a statistical tool used to detect differences between experimental group means. ANOVA is warranted in experimental designs with one dependent variable that is a continuous parametric numerical outcome measure, and multiple experimental groups within one or more independent (categorical) variables. In ANOVA terminology, independent variables are called *factors*, and groups within each factor are referred to as *levels*.

5.2: Intended Learning Outcomes

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

- understand the principles behind the use of ANOVA.
- describe the methods of using the test tool.
- understand the application of the tool.

5.3: Analysis of Variance

5.3.1: Assumptions of ANOVA

Specifically, a data set should meet the following criteria before being subjected to ANOVA:

- *Parametric data:* A parametric ANOVA, the topic of the article, requires parametric data (ratio or interval measures). There are non-parametric, one-factor versions of ANOVA for nonparametric ordinal (ranked) data, specifically the Kruskal-Wallis test for independent groups and the Friedman test for repeated measures analysis.

- *Normally distributed data within each group:* ANOVA can be thought of as a way to infer whether the normal distribution curves of different data sets are best thought of as being from the same population or different populations.

5.3.2: Mechanism of Calculation

ANOVA evaluates differences in group means in a round-about fashion, and involves the "partitioning of variance" from calculations of "Sum of Squares" and "Mean Squares." Three metrics are used in calculating the ANOVA test statistic, which is called the *F* score (named after R.A. Fisher, the developer of ANOVA):

(i) ***Grand Mean***, which is the mean of all scores in all groups;

(ii) ***Sum of Squares***, which are of two kinds, the sum of all squared differences between group means and the Grand Mean (between- groups Sum of Squares) and the sum of squared differences between individual data scores and their respective group mean (within-groups Sum of Squares), and

(iii) ***Mean Squares***, also of two kinds (between-groups Mean Squares, within-groups Mean Squares), which are the average deviations of individual scores from their respective mean, calculated by dividing Sum of Squares by their appropriate degrees of freedom.

The systematic procedure for obtaining two or more estimate of variance and comparing them is?

Analysis of variance is a statistical method of comparing the _____ of several populations. a. standard deviations b. variances c. means d. proportions e. none of the options

Self-Assessment Exercises

1. The table below shows the number of seeds for five varieties of garden egg to three level of Indo-acetic acid (IAA)

IAA\varieties	**A**	**B**	**C**	**D**	**E**
I	3	5	10	7	8
II	2	4	7	4	5
III	4	5	8	6	7

5.4: Summary

This Unit, looked in another test tool used to compare two variables. The single factor analysis of variance was considered here

5.5: References/Further Readings/Web Sources

- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Bailey, N.T.J. (1994). *Statistical Methods in Biology*. Third Edition. Cambridge University Press. United Kingdom.

Analysis of variance https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/

Analysis Of Variance (ANOVA) - Analytics Vidhya
https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

https://www.statisticssolutions.com/analysis-of-covariance-ancova/

https://www.investopedia.com/terms/a/anova.asp

https://www.youtube.com/watch?v=ZSwjaIUPBRg

5.6: Possible Answers to Self-Assessment Exercises

State the null hypothesis: There is no significant difference between seed number in five varieties of garden egg and levels of IAA. Calculate the totals from the table as below:

IAA\varieties	**A**	**B**	**C**	**D**	**E**	**Total**
I	3	5	10	7	8	**33**
II	2	4	7	4	5	**22**
III	4	5	8	6	7	**30**
Total	**9**	**14**	**25**	**17**	**20**	**GT=85**

Then calculate the Correction Factor (CF) as: $CF = GT^2/N = 85^2/15 = 481.7$

Calculate the Mean Squares (SS)

$BLOCKss = (33^2 + 22^2 + 30^2/5) - CF = 12.9$

$VARIETIESss = (9^2 + 14^2 + \ldots\ldots\ldots + 20^2)/3 - CF = 48.6$

$TOTALss = (3^2 + 5^2 + \ldots\ldots\ldots + 7^2) - CF = 65.3$

$ERRORss = TOTALss - (BLOCK + VARIETIESss) = 3.8$

Then calculate:

$BLOCK_{MS} = BLOCKss / BLOCK_{DF} = 12.9/2 = 6.45$

$VARIETIES_{MS} = VARIETIESss/VARIETIES_{DF} = 48.6/4 = 12.15$

$Block\ F\text{-}value = Block_{MS} / Error_{MS} = 6.45/0.475 = 13.58$

$Varieties\ F\text{-}value = Varieties_{MS} / Error_{MS} = 12.15/0.475 = 25.5$

The calculated F-values are compared with the F-distribution table, using their respective degrees of freedoms.

SOURCE	**DF**	**SS**	**MS**	**F**
Block	2	12.9	6.45	13.58**
Varieties	4	48.6	12.15	25.58**
Error	8	3.8	0.475	
Total	14	65.3		

** indicates that the values are highly significant.

Conclusion: Since the F-values are highly significant, we reject the null hypothesis. It means that the three levels of IAA have effect on the seed number of the five varieties of garden egg.

Complete the Anova table below and draw your conclusions.

Source	**Sum of squares (SS)**	**Degrees freedom (DF)**	**Mean squares (MS)**	**F-ratio**
Varieties	123.44	*	*	*
Residual	*	15	*	
Total	210.21	18		

RESIDUAL**ss** = 210.21 – 123.44 = 86.8

VARIETIES**DF** = 18 – 15 = 3

VARIETIES**MS** = 123.44/3 = 41.15

RESIDUAL**MS** = 86.8/15 = 5.79

F- ratio = 41.15/5.79 = 7.11

The complete table is as shown below.

Source	**Sum of squares (SS)**	**Degrees of freedom (DF)**	**Mean squares (MS)**	**F-ratio**
Varieties	123.44	3	41.15	7.11
Residual	86.8	15	5.79	
Total	210.21	18		

Conclusion:
The observed variance ratio of 7.12 is greater than the table values at both 5% (3.29) and 1% (5.42). That means there is high-significant difference among the varieties. Therefore, we reject the null hypothesis that the varieties are the same.

Glossary

- ANOVA: Analysis of variance usually refers to an analysis of a continuous dependent variable where all the predictor variables are categorical.
- Chi-square distribution: the distribution used to analyze counts in frequency tables.
- **descriptive statistics**: Statistics, such as the mean, the standard deviation, the proportion, and the rate, used to describe attributes of a set of data.
- goodness of fit: Assessment of the agreement of the data with either a hypothesized pattern (e.g., independence of row and column factors in a contingency table or the form of a regression relationship) or a hypothesized distribution (e.g., comparing a histogram with expected frequencies from the normal distribution).
- **mean ($\bar{X}$)**: The most common measure of central tendency, denoted by l' in the population and by in the sample. In a sample, the mean is the sum of the *X* values divided by the number *n* in the sample ($\sigma X/n$).
- median: Value such that half of the observations' values are less than and half are greater than that value.
- P-value: The probability of getting a result (e.g., t or $\chi 2$ statistics) as or more extreme than the observed statistic had H0 been true.
- significance level: A preset value of α against which P-values are judged in order to reject H0
- standard deviation: A measure of the variability (spread) of measurements across subjects.
- standard error: The standard deviation of a statistical estimator.
- two-sided test: A test that is non-directional and that leads to a two-sided P-value.
- variance: A measure of the spread or variability of a distribution, equaling the average value of the squared difference between measurements and the population mean measurement.

End of the module Questions

1. Find the median of the set of figures obtained from measuring the weight to nearest gram of fingerlings: 3, 5, 2, 6, 5, 9, 5, 2, 8, 6
2. Find the mode of the set of figures obtained from measuring the weight to nearest gram of fingerlings: 3, 5, 2, 6, 5, 9, 5, 2, 8, 6
3. Find the mean of the set of figures obtained from the measure of the length of fish juvenile: 51.6, 48.7, 50.3, 49.3, 48.9
4. Find the median of the set of figures obtained from the measure of the length of fish juvenile: 51.6, 48.7, 50.3, 49.3, 48.9
5. **The numerical value of a standard deviation can never be _____.**

a. Negative
b. Zero
c. Larger than the variance
d. None of the above

6. **The average of squared deviations from the arithmetic mean is known as _____.**

a. Quartile deviation
b. Standard deviation
c. Variance
d. None of the above

7. **Which of the following is not a characteristic of a good measure of dispersion?**

a. It should be rigidly defined
b. It should be based on extreme values
c. It should be capable of further mathematical treatment and statistical analysis
d. None of the above

8. **Which of the following cannot be calculated for open-ended distributions?**

a. Standard deviation
b. Mean deviation
c. Range
d. None of the above

9. The statistical tool for the mean of a population used when the population is normally or approximately normally distributed is?

10. Two plant extracts are claimed to be effective in curing stomach ulcer were tested on patients. The patients' reactions to treatment were recorded in the table below:

EXTRACTS	**EFFICACY**		
	HELPED	**HARMED**	**NO**
Anona	62	84	24
Bauhinia	34	44	22

Test the data whether the two extracts have the same effects.

11. The table below is an outcome of a survey of Ahmadu Bello University Zaria graduates working in Abuja. They were divided into four groups on the basis of their classes of degree and their income in practice, ten years after graduation

	Income		
	High	**Medium**	**Low**
First	22	10	10
Second	10	13	7
Third	20	6	6
Pass	5	9	15

Determine the relationship between the class of degree and their income.

12. The _____ sum of squares measures the variability of the observed values around their respective treatment means. a. treatment b. error c. interaction d. total

13. The _____ sum of squares measures the variability of the sample treatment means around the overall mean. a. treatment b. error c. interaction d. total

14. If the true means of the k populations are equal, then MSTR/MSE should be: a. more than 1.00 b. close to 1.00 c. close to 0.00 d. close to -1.00 e. a negative value between 0 and - 1 f. not enough information to make a decision

15. If the MSE of an ANOVA for six treatment groups is known, you can compute a. df1 b. the standard deviation of each treatment group c. the pooled standard deviation d. b and c e. all answers are correct

16. To determine whether the test statistic of ANOVA is statistically significant, it can be compared to a critical value. What two pieces of information are needed to determine the critical value? a. sample size, number of groups b. mean, sample standard deviation c. expected frequency, obtained frequency d. MSTR, MSE

Answers

1. 5
2. 5
3. 49.8
4. 49.5
5. **Answer: a**
6. **Answer: c**
7. **Answer: d**
8. **Answer: b**
9. t-test
10. Our null hypothesis (H_O): The two plant extracts have the same effect on the patients. First calculate the expected frequencies.

	Helped	**Harmed**	**No effect**	**Total**
A. senegalensis	62	84	24	17
B. monandra	34	44	22	100
Total	96	128	46	270

Frequencies of each category:

Helped: *A. senegalensis*: 170 x 96 ÷ 270 = 60.4

B. monandra: 100 x 96 ÷ 270 = 35.6

Harmed: *A. senegalensis*: 170 x 128 ÷ 270 = 80.4

B. monandra: 100 x 128 ÷270 = 47.4

No effect: *A. senegalensis*: 170 x 46 ÷ 270 = 29.0
B. monandra: 10 0 x46 ÷ 270 = 17.0

Form a table of expected frequencies.

	Helped	**Harmed**	**No effect**
A. senegalensis	60.4	80.6	29.0
B. monandra	35.6	47.4	17.0
Total	**96**	**128**	**46**

$\chi^2 = \frac{(62-60.4)^2}{60.4} + \frac{(84-80.6)^2}{80.6} + \ldots\ldots\ldots\ldots \frac{(22-17.0)^2}{17.0} = 2.83$

Therefore, $\chi^2 = 2.83$

Find the degrees of freedom (DF): Rows (r) = 2; Columns (c) = 3

DF = (r – 1) (c – 1)

= (2 – 1) (3 – 1) = 2

The table χ^2 value at DF 2 is 5.99, at $\alpha = 0.05$

Since our calculated χ^2 value (2.83) is less than the χ^2 table value (5.99), it shows that the effect of the extracts is not significant. Therefore, we accept our H_O i.e the two extracts have the same effect on the patients.

11. **First, we state the hypotheses.**

1. Null hypothesis (H_O): There is no significant relationship between the class of degree and income of A.B.U. Zaria graduates in Abuja.
2. Alternative hypothesis (H_1): There is significant relationship between the class of degree and income of A.B.U. Zaria graduates in Abuja.

Compute the totals of the observed values.

	Income			
	High	**Medium**	**Low**	**Total**
First	22	10	10	42
Second	10	13	7	30
Third	20	6	6	32
Pass	5	9	15	29
Total	**57**	**38**	**38**	**133**

Compute the expected frequencies.

First – High 42 x 57 / 133 = 18
First - Medium 42 x 38 / 133 = 12
First – Low 42 x 38 / 133 = 12
Second – High 30 x 57 / 133 = 12.9

Second – Medium 30 x 38 / 133 = 8.6
Second – Low 30 x 38 / 133 = 8.6
Third – High 32 x 57 / 133 = 13.7
Third – Medium 32 x 38 / 133 = 9.1
Third – Low 32 x 38 / 133 = 9.1
Pass – High 29 x 57 / 133 = 12.4
Pass - Medium 29 x 38 / 133 = 8.3
First – Low 29 x 38 / 133 = 8.3

Place the computed expected values in the table of frequencies.

	Income		
	High	**Medium**	**Low**
First	18	12	12
Second	12.9	8.6	8.6
Third	13.7	9.1	9.1
Pass	12.4	8.3	8.3

We can compute the chi square value using the χ^2 formular.

$$\chi^2 = \frac{(22 - 18)^2}{18} + \frac{(10 - 12)^2}{12} + \frac{(10 - 12.9)^2}{12.9} \ldots\ldots. \frac{(15 - 8.3)^2}{8.3} = 19.65$$

Then determine the degrees of freedom (d.f).
d.f = (column -1)(row -1)
= (3-1)(4-1) = 6
Check the value in the chi square table at df 6 At df 6, the value is 12.59

Conclusion:
Since the calculated value of 19.65 in higher than the table value of 12.59, it shows that the relationship between the class of degree and income of the A.B.U Zaria graduates in Abuja is significant. Therefore, we reject our null hypothesis
(H_O) and accept our alternative hypothesis (H_1).

12. B
13. A
14. B
15. C
16. A

Module 3: Application of Biostatistics II

Unit 1: Simple Linear Regression

1.1: Introduction

The relationship between two dependable variables is considered in this unit as simple linear relationship.

1.2: Intended Learning Outcomes

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

- understand the principles behind simple linear relationship

1.3: Simple Linear Regression

Regression is defined as a dependent relationship between two or more variables. Simple in the sense means only two variables are involved. Linear means the relationship is a straight line. Therefore, simple linear relation is a straight-line dependent relationship between two variables. The standard procedure for doing the simple linear regression is to plot the dependent variable 'Y' on the ordinate and the independent variable 'X' on the abscissa known as Scattergram. The relationship is described as the regression of Y on X. There are other kinds of relationships that can occur between X and Y; if Y varies directly with X, then the relationship is linear, however, if Y changes with a different unit per unit change in X is called curvilinear.

In an experiment situation, such perfect relationship between X and Y does not occur. If the points plotted do not fall in a straight line, then there is need to construct the **line of best fit/regression line.** The line of best fit is defined as the line which best fits a series, passing through the points in such a way that the summation of the squared deviation between the points and the line is at minimum. The regression line for a linear relationship is represented by the equation:

Y = a + bx

Where a = intercept at Y
b = slope/gradient/regression coefficient.

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

a = y – bx

The line that best fit the point in a scatter diagram is called?
In a straight-line equation 'Y = M + bX', the 'b' denotes?

Self-AssessmentExercises

A scientist was interested in finding out the acute effect of neem leaf dust (mg/l) on African catfish during the 4 days experimental period and obtained the following result.

Concentration of neem leaf dust (mg/l)	0.00	1.00	2.00	3.00	4.00	5.00	6.00
Cumulative mortality (%)	0.00	30.00	40.00	56.70	63.60	76.70	90.00

1.4: Summary

The method of determining the relationship between two dependable variables was discussed in this Unit. Procedure for the determination of simple linear regression was determined and learnt

1.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp

Simple linear regression:
https://www.scribbr.com/statistics/simple-linear-regression/#:~:text=What%20is%20simple%20linear%20regression,Both%20variables%20should%20be%20quantitative.

What is Simple Linear Regression? | STAT 462
https://online.stat.psu.edu/stat462/node/91/

https://www.jmp.com/en_sg/statistics-knowledge-portal/what-is-regression.html

https://www.youtube.com/watch?v=owI7zxCqNY0

https://www.youtube.com/watch?v=GhrxgbQnEEU

1.6: Possible Answers to Self-Assessment Exercises

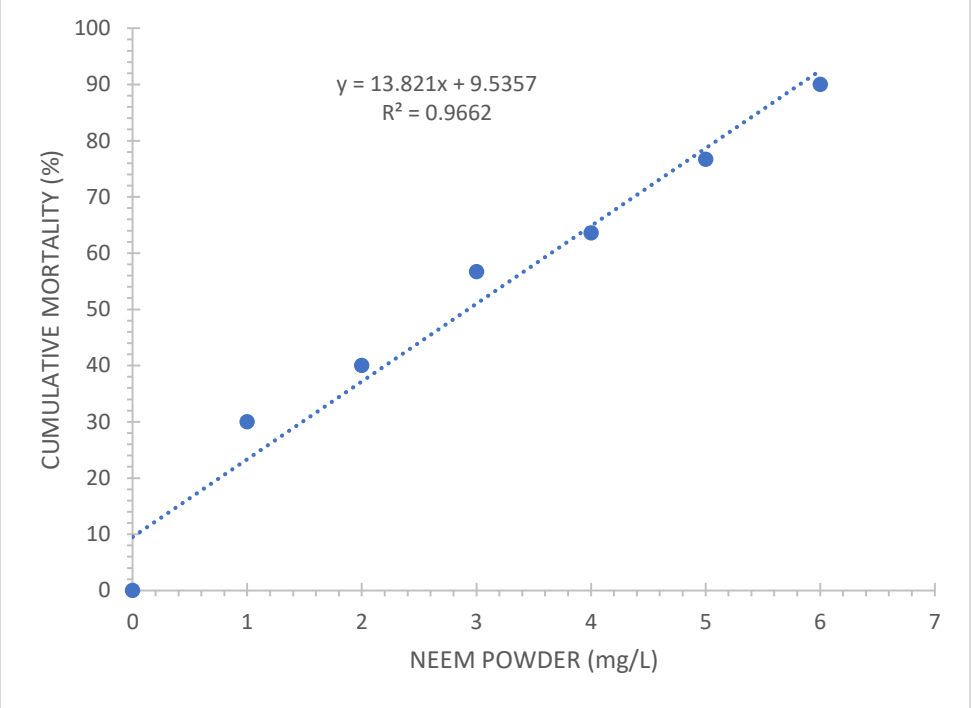
1.

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

X	Y	XY	X^2
0	0	0	0
1	30	30	1
2	40	80	4
3	56.7	170.1	9
4	63.6	254.4	16
5	76.7	383.5	25
6	90	540	36
21	**356.7**	**1458**	**91**

$$\mathbf{b} = (1458 - (21x356.7)/7)/(91 - ((21)^2)/7)$$
$$(1458 - (7490.7)/7)/(91 - 441/7)$$
$$(1458 - 1070.1)/(91 - 63)$$
$$(387.9)/28$$

b=13.85
Y = a + 13.85x
a = y – bx
mean of y = 356.7/7 = 50.96
mean of x = 21/7 = 3
a = 50.96 – 13.85 x 3
a = 50.96 – 41.55
a= 9.41
Y = 9.41 + 13.85x
When x = 6.5
Y = 9.41 + 13.85(6.5)
Y = 9.41 + 90.025
Y = 99.435
When x = -0.5
Y = 9.41 + 13.85(-0.5)
Y = 9.41-6.925
Y= 2.485


y = 13.821x + 9.5357
R² = 0.9662
CUMULATIVE MORTALITY (%)
NEEM POWDER (mg/L)
0
10
20
30
40
50
60
70
80
90
100
0
1
2
3
4
5
6
7

Unit 2: Simple Linear Correlation

Unit Structure

2.1: Introduction

The relationship between two simultaneously variables is considered in this unit as simple linear correlation.

2.2: Intended Learning Outcomes

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

- understand the principles behind simple linear correlation

2.3: Simple Linear Correlation

Simple linear correlation exists when the two variables changes simultaneously but they are functionally dependent on each other. To know whether two of these variables are correlated 'r' is calculated. There are different types of correlation coefficients;

a. Pearson's correlation coefficient (r): this is used for any bivariate populations which are normally distributed. The value of r can range from -1 to 0 to +1. If r is positive, the variables are positively correlated and if negative they are negatively correlated. Whether the calculated r is strongly correlated or not depends on the degree of freedom available, because a change in the degree of freedom will affect the degree of significance.

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

The degree of freedom for correlation is n-2, because on df goes for X and the other goes for Y variable. Coefficient of determination (r^2) tells the percentage amount of variation due to regression or correlation. It measures the part due to the dependence of one variable on the other.

b. Spearman's rank correlation coefficient = r_s: this is non parametric rank correlation. The ranking in this test is the same as in other non-parametric tests.

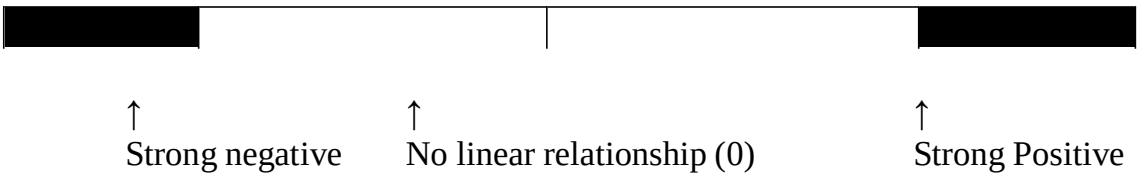
$r_{s = 1 -}(\frac{6 \sum dt^2}{n^3 - n})$

The values of the two stations can be compared to detect significant difference by using Mann-Whitney U test which has a similar approach to the problem and gives entirely equivalent results to the Wilcoxon two sample test.

The range of values for correlation coefficient is from -1 to +1 that is if there is

a. Strong positive linear relationship between the variables, the value or **r** will be close to +1.
b. Strong negative linear relationship between the variables, the value of **r** will be close to -1.
c. No linear relationship between the variables or only a weak relationship, the value of **r** will be close to 0.

Figure 6.1: Strength of linear relationships



Generally, simple correlation and simple linear regression may be:

1. **Positive correlation** – when an increase in one variable is associated to a greater or lesser extent with an increase in the other.
2. **Negative correlation** – when an increase in one variable is associated to a greater or lesser extent with a decrease in the other.
3. **Perfect correlation** – when a change in one variable is exactly matched by a change in the other variable. If both increase together, it is perfect positive correlation: if one decreases as the other increases, it is perfect negative correlation.
4. **High correlation** – When a change in one variable is almost exactly matched by a change in the other.
5. **Low correlation** – when a change in one variable is to a small extent matched by a change in the other.
6. **Zero correlation** – when the two variables are not in matched at all, and there is no relationship between changes in one variable and changes in the other

SPURIOUS CORRELATION

When interpreting correlation, r, it is important to realize that, there may be no direct connection at all between highly correlated variables. When

this is so, the correction is termed spurious or nonsense correlation. It can arise in two ways: a). There may be an indirect connection; b) There may be a series of coincidences.
The statistical techniques that measure the degree/strength of linear relationship in a bivariate normal distribution is called? If the calculated 'r; is close to -1, it indicates?
The correlation that is symbolized with 'r' is?

Self-Assessment Exercises

1. A scientist was interested in finding out the acute effect of neem leaf dust (mg/l) on African catfish during the 4 days experimental period and obtained the following result.

Concentration of neem leaf dust (mg/l)	0.00	1.00	2.00	3.00	4.00	5.00	6.00
Cumulative mortality (%)	0.00	30.00	40.00	56.70	63.60	76.70	90.00

Use the data to calculate correlation coefficient

2. The following data of dissolved oxygen values was recorded for two stations of a River. Use the Spearman's correlation to determine any relationship between the values obtained for the two stations.

Station 1	Station 2
7.4	10.4
7.6	10.8
7.9	11.1
7.2	10.2
7.4	10.3
7.1	10.2
7.4	10.7
7.2	10.5
7.8	10.8
7.7	11.2
7.8	10.6
8.3	11.4
7.4	8.6

2.4: Summary

The methods of comparing non dependable simple variables in both parametric and non-parametric instances were discussed. The Unit also discussed the methods of determining significant relationship between two variables that are not necessary dependent.

2.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp
- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp
- Helmut F. van Emden.(2008). *Statistics for Terrified Biologists.* Blackwell Publishing Limited. USA.

Simple linear correlation
https://www.ndsu.edu/faculty/horsley/Corr_revised.pdf

SIMPLE LINEAR CORRELATION Simple linear correlation ...
https://www.ndsu.edu/faculty/horsley/corr.pdf

Pearson Correlation and Linear Regression
http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/
https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/

https://www.youtube.com/watch?v=wHatBwHLrnA

https://www.youtube.com/watch?v=aztcS-3MwH0

2.6: Possible Answers to Self-Assessment Exercises

1.

X	Y	XY	X^2	Y^2
0	0	0	0	0
1	30	30	1	900
2	40	80	4	1600
3	56.7	170.1	9	3214.89
4	63.6	254.4	16	4044.96
5	76.7	383.5	25	5882.89
6	90	540	36	8100
21	**356.7**	**1458**	**91**	**23742.74**

$$\frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n})(\sum Y^2 - \frac{(\sum Y)^2}{n})}}$$

$$r = \frac{1458 - \frac{21 \, x \, 356.7}{7}}{\sqrt{(91 - \frac{(21)^2}{7})(23742.74 - \frac{(356.7)^2}{7})}}$$

$$= \frac{1458 - \frac{7490.7}{7}}{\sqrt{(91 - \frac{441}{7})(23742.74 - \frac{127234.89}{7})}}$$

$$\frac{1458 - 1070.1}{\sqrt{(91 - 63)(23742.74 - 18176.41)}}$$

$$\frac{387.9}{\sqrt{(28)(5566.33)}}$$

$$\frac{387.9}{\sqrt{155857.24}}$$

$$\frac{387.9}{394.79}$$

r =0.9825 this implies that it strongly positively correlated

$r^2 = 0.9825^2$

$r^2 = 0.9654$

interpretation = 96.54% i.e. you are 96.54% sure that the relationship exist ($p<0.05$)

2.

Station 1	Station 2	Rank 1	Rank 2	d	dt^2
7.4(4)	10.4(5)	**5.5**	5	0.5	0.25
7.6(8)	**10.8(9)**	8	9.5	-1.5	2.25
7.9	11.1(11)	12	11	1	1
7.2(3)	**10.2(2)**	**2.5**	**2.5**	0	0

7.4(5)	10.3(4)	**5.5**	4	1.5	2.25
7.1(1)	**10.2(3)**	1	**2.5**	-1.5	2.25
7.4(6)	10.7(8)	**5.5**	8	-2.5	6.25
7.2(2)	10.5(6)	**2.5**	6	-3.5	12.25
7.8(10)	**10.8(10)**	**10.5**	9.5	1	1
7.7	11.2(12)	9	12	-3	9
7.8(11)	10.6(7)	**10.5**	7	3.5	12.25
8.3	11.4(13)	13	13	0	0
7.4(7)	8.6(1)	**5.5**	1	4.5	20.25
					69

$r_s = 1 - \left(\frac{6\sum dt^2}{n^3 - n}\right)$

$r_s = 1 - \left(\frac{6(69)}{13^3 - 13}\right)$

$= 1 - \left(\frac{390}{2197 \quad -13}\right)$

$= 1 - \left(\frac{414}{2184}\right)$

$= 1 - 0.1896$

$= 0.810$

positively correlated

$r^2 = 0.810 \times 0.810 = 0.656 = 65.6\% = p>0.05$

Unit 3: Non-Parametric Tests

3.1: Introduction

There are five advantages that non parametric methods have over parametric methods.

a. They can be used to test population when the variable is not normally distributed.
b. They can be used when the data are nominal or ordinal.
c. Can be used to test hypothesis that do not involve population parameters.
d. In most cases, computation is easier than in parametric.
e. They are easier to understand.

The disadvantages include;

a. They are less sensitive than in parametric i.e., larger differences are needed before the null hypothesis can be rejected.
b. They tend to use less information than the parametric tests.
c. They are less efficient than their parametric counterparts when the assumptions of the parametric are met.

3.2: Intended Learning Outcomes

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

- understand the methods and application of sign test.
- Understand the methods and applications of Wilcoxon Signed Rank Test.
- Understand the methods and applications of Mann-Whitney test
- Understand the methods and applications of Kruskal Wallis Rank test

3.3: Non-Parametric tests

3.3.1: The sign test

The simplest non-parametric test is the sign test for single samples. It is used to test the value of a median for a specific sample. In using Sign test, you:

a. Hypothesize the specific value for the median of a population.
b. Select a sample of data and compare each value with the conjectured median.
c. Assign plus sign if the data value is above the conjectured median.
d. Assign minus sign if the data value is below the conjecture median.
e. And zero (0) if it is the same as the conjecture median.
f. Compare the number of plus and minus signs and ignore the zeros
g. If the null hypothesis (*Ho*) is true, the number of plus signs should be approximately equal to the number of minus signs.
h. But if the *Ho* is not true, there will be disproportionate number of plus or minus signs.

3.3.2: Wilcoxon Signed Rank test

This is a paired sample testing by ranks that represents the parametric counterpart of the paired sample t-test. Both tests can be used to solve the same kind of problem but the paired sample t-test is more powerful than the non-parametric test. Wilcoxon signed rank test is used when the samples are not drawn from normal distribution. In that all observations are pulled in ascending order of magnitude ignoring the signs and zero values. The non-zero values are assigned the ranks 1 to n.

Wilcoxon (T^I) = m (n+1) ─ T

= (mn + m) - T

Where m= number of ranks with less frequent sign

T= sum of ranks with less frequent sign

n= total number of sample

Test for significance

Note if either T or T^I is less than or equal to critical t, reject H_0 (Null hypothesis) . To accept H_0 the critical t must be less than both T and T^I at degree of freedom (df) = n.

3.3.3: Mann-Whitney test

This is a non-parametric analogue test of the parametric unpaired t-test. As in all non-parametric tests, the actual measurements are not employed, instead the ranks of the measurements are used.

$\mathbf{U^1 = n_1 n_2 - U}$

$\mathbf{U = n_1 n_2 + n_1((n_1 + 1))/2 - R_1}$

Where; n_1 and n_2 = total number of observations in group 1 and 2 respectively.

R1 and R2= sum of ranks of group 1 and 2 respectively.

Test of Significance

If U and U^1 calculated are both less than U critical, H_0 is accepted at df= $U_{\alpha(2)}\, n_1 n_2$. In checking the accuracy of ranking.

$\mathbf{R_1 + R_2 = (N\,(N + 1))/2 = (N^2 + N)/2}$

Where N= total number of observations

3.3.4: Kruskal-Wallis Rank test

This is often called an analysis of variance by ranks. It is a non-parametric test that can be used in a situation where the single factor is not applicable (when underlying assumptions are not met) i.e., ni is not the same. The procedure of ranking is similar to Mann-Whitney test.

$$\mathbf{H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_1^{\,2}}{n_1} - 3\,(N + 1)}$$

Where n = number of observations in a sample (group i)

N = $\sum_{i=1}^{k}$ (total number of observations in all i= 1)

R_1 = sum of ranks of ni observation in group i

KRUSKAL WALLIS TEST WITH TIED RANKS

$$Hc = \frac{H}{C}$$

$$\mathbf{C = correction\ factor = 1 - \frac{\sum T}{N^3 - N}}$$

$$\mathbf{\sum T = \sum_{i=1}^{m} (t_1^{\,3} - t_1)}$$

Where;

T=number of ties per groupdf =K—1

Note: if K is greater than 3 treatments, the H table cannot be used, it is therefore, recommended that chi=square at df = k-1 be used to verify Ho.

How many advantages does exist of non-parametric test methods over the parametric methods?

The group of tests that can be used when data are nominal or ordinal is?

What is parametric test known to compare?

Self-Assessment Exercises

1. The effect of drugs on the zone of inhibition of bacteria sample is as follows;

X1	X2
54.2	80.3
60.4	99.9
80.5	50.5
49.5	75.5
33.2	60.2
35.5	105.1
20.3	25.4
29.1	19.5
40.8	30.1
33.2	3.4

Calculate the relationship and Prove that Ho should either be accepted or rejected.

2. Consider the height (m) of males and students in Biostatistics II class as provided below, ca the difference between them and verify the accuracy of the result.

Male	Female
7.6	6.9
7.4	6.8
7.3	6.6
7.2	6.5
7.1	6.4
7.0	-
6.7	-

3.4: Summary

This unit dealt with four non-parametric tests, commonly used in Biostatistics. The methods and principles of four non-parametric tests were considered in this Unit

3.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp

- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp

Non-parametric Test (Definition, Methods, Merits ... - Byju's
https://byjus.com/maths/non-parametric-test/#:~:text=Non%2Dparametric%20tests%20are%20experiments,not%20have%20any%20underlying%20population.

Nonparametric Tests - Overview, Reasons to Use, Types
https://corporatefinanceinstitute.com/resources/knowledge/other/nonparametric-tests/

Nonparametric Statistics: Overview - Investopedia
https://www.investopedia.com/terms/n/nonparametric-statistics.asp
https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/parametric-and-non-parametric-data/

https://www.youtube.com/watch?v=ftnOBcXtBEQ

https://www.statisticshowto.com/mann-whitney-u-test/

https://www.youtube.com/watch?v=fEobVCV2TJE

https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/kruskal-wallis/

https://www.youtube.com/watch?v=q1D4Di1KWLc

3.6: Possible Answers to Self-Assessment Exercises

1.

X1	X2	*X1 –X2*	Ranks of di	***Signed Rank***
54.2	**80.3**	*-26.1*	5	***-5***
60.4	**99.9**	*-39.5*	9	***-9***
80.5	**50.5**	*30*	**8**	*+8*
49.5	**75.5**	*-26*	4	***-4***
33.2	**60.2**	*-27*	6	***-6***
35.5	**105.1**	*-69.6*	10	***-10***
20.3	**25.4**	*-5.1*	1	***-1***
29.1	**19.5**	*9.6*	**2**	*+2*

40.8	**30.1**	*10.7*	**3**	*+3*
33.2	**3.4**	*29.8*	**7**	*+7*

Wilcoxon (T^I) = 4(10+1) ─ 20
Wilcoxon (T^I) = 4(11) ─ 20
Wilcoxon (T^I) = 44 ─ 20
Wilcoxon (T^I) = 24
= 40+4 -20
= 44 – 20
=24
Then check the Wilcoxon table for significance difference

2.

Male	Female	Rank male	Rank female
7.6	**6.9**	12	6
7.4	**6.8**	11	5
7.3	**6.6**	10	3
7.2	**6.5**	9	2
7.1	**6.4**	8	1
7.0	-	7	
6.7	-	4	
n1=7	n2=5	R1=61	R2=17

Prove to either accept or reject the null hypothesis and verify the accuracy of the result.

$$U = n_1 n_2 + n_1((n_1 + 1))/2 - \mathbf{R_1}$$
$$U = 7x5 + 7\ ((7 + 1))/2 - 61$$
$$U = 35 + 7\ ((8))/2 - 61$$
$$U = 35 + ((56))/2 - 61$$
$$U = 35 + 28 - 61$$
$$U = 63 - 61$$
$$U = 2$$
$$U^1 = n_1 n_2 - U$$
$$U^1 = 7x5 - 2$$
$$35 - 2$$
$$U^1 = 33$$
$$\mathbf{R_1 + R_2} = (\mathbf{N}\ (\mathbf{N} + \mathbf{1}))/\mathbf{2}$$
$$61+17 = (12\ (12 + 1))/2$$
$$78 = (12\ (13))/2$$
$$78 = 156/2$$
$$78 = 78$$

Unit 4.0: Ecological Statistics

Unit Structure

4.1: Introduction

Biological communities vary in the number of species they contain and the knowledge of this number is important in understanding the structure of the community.

4.2: Intended Learning Outcomes

By the end of this unit, students should be able to:
- understand the methods and application of species richness.
- describe the methods and applications of diversity index.
- understand the methods and applications of species evenness
- describe the methods and applications of species dominance

4.3: Ecological Indices

4.3.1: Species Richness

This is the number of species in a community. The total number of species (S) in a community has been used in most studies as an index of species richness. This is wrong since S depends on the sample size, the higher the species richness). However, the following indices are independent of the sample size, they are based on the relationship S and the total number of individuals observed (N), which increase with increasing sample size.
Margalef's index (d) $= (S - 1)/(\text{In } (N))$
Where S = total number of species,
N = total number of individuals

In is the natural logarithm

Menhinicks index (D) $= S/VN$

4.3.2: Diversity index

Species diversity includes both specie richness and evenness. Communities with large number of species that are evenly distributed are the most diverse and communities with few species that are dominated by one species are the least diverse.

Shannon-Wiener diversity index (H): Peet (1974) terms those indices heterogeneity indices because they take into account both evenness and species richness to produce a single value. The most widely used measure of diversity are the information theory indices. Shannon and Wiener independently derived the function which has become known as Shannon-Wiener index of diversity.

$$H = \frac{N\text{Log } N - \sum_{i=1}^{S} F_1 \log F_1}{N}$$

CALCULATION OF SIGNIFICANCE DIFFERENCE

Hutcheson (1970) provides methods of calculating 't' to test for significant difference between samples.

$$t = \frac{H_1 - H_2}{\sqrt{H_1 S^2 - H_2 S^2}}$$

Where S^2H =Variance of $H = \frac{\sum F_1 \log^2 F_1 - \frac{(\sum F_1 \log F_1)^2}{N}}{N^2}$

$$F_1 \log^2 F_1 = \frac{(F_1 \log F_1)^2}{F_1}$$

Where S^2H =Variance of $H = \frac{\frac{(F_1 \log F_1)^2}{F_1} - \frac{(\sum F_1 \log F_1)^2}{N}}{N^2}$

4.3.3: Species Evenness

The relative abundance of species is also important. For instance, two communities may both contain the same number of species but one community may be dominated by one specie while the other community may contain large numbers of all species**.** The relative abundance of rare and common species is called evenness. Communities dominated by one or a few species have a low evenness while those that have a more distribution of species have high evenness. The ratio of the observed diversity (H) to the maximum diversity (H_{max}) is taken as a measure of evenness (E).

$E = H/Hmax = H/(\text{Log } S)$

4.3.4: Species Dominance

This is the best known of the second group of heterogeneity indices referred to as measure of **dominance**. They are weighed towards the

abundance of the commonest species. If a community with high diversity was randomly sampled twice, there is a good chance that the two sample will contain different species. However, if a low-density community were sampled twice, it is likely that both of the samples will contain many of the same species. Simpson (1949) derived a formula based on the expected outcome of two random samples.

$D = \sum_{0=1}^{s} \frac{ni(ni-1)}{N(N-1)}$

Where ni= the number of individuals in the ith species

N= the total number of individuals

The best known of the second group of heterogeneity indices is referred to as?

Shannon and Wiener independently derived the function which has become known as?

Self-Assessment Exercises

1. The following data were recorded of the abundance of Bacteria in three stations of Nile River. Using the data provided, determine all the ecological statistics in each station.

Species	Station 1	Station 2	Station 3
E. coli	21	19	20
Salmonella typhi	8	23	0
Shigella sp	18	7	23
Pseudomonas aureus	17	16	0
Streptococcus sp.	20	24	23
Staphylococcus sp	10	0	12
Klebsilla sp.	14	13	14

4.4: Summary

The Unit dealt with some ecological statistical tools where Five ecological indices were considered and treated.

4.5: References/Further Readings/Web Sources

- Ogbeibu, A.E. (2005). *Biostatistics: A practical Approach to Research and Data Handling*; First Edition, Mindex Publishing Company Limited, 263pp.
- Agarwal, B.L (2006). *Basic Statistics*, Revised Fourth Edition, Newage International Publishers, 763pp

- Omalu, ICJ and Arimoro, FO (2010). *Research Methods and Analyses*, Logicgate Publishers Ilorin, Minna, Kaduna, Nigeria, 276pp
- Murray, R.S. (1972). *Theory and Problems of Statistics*, McGraw-Hill Book Company, 340pp

Ecological statistics – CiteSeerX https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.9550&rep=rep1&type=pdf#:~:text=Statistical%20ecology%20is%20the%20subfield,abundance%20and%20distribution%20of%20organisms.

Environmental and Ecological Statistics | Home - Springer
https://www.springer.com/journal/10651

https://www.youtube.com/watch?v=GEsGTzOedXw

https://www.youtube.com/watch?v=w9TvlB4hf7k

https://www.youtube.com/watch?v=ghhZClDRK_g

https://www.youtube.com/watch?v=OBfpdM9SJIc

4.6: Possible Answers to Self-Assessment Exercises

1.

Species	Station 1	Station 2	Station 3
E. coli	21	19	20
Salmonella typhi	8	23	0
Shigella sp	18	7	23
Pseudomonas aureus	17	16	0
Streptococcus sp.	20	24	23
Staphylococcus sp	10	0	12
Klebsilla sp.	14	13	14
S =	**7**	**6**	**5**
N =	**108**	**102**	**92**

Margalef's index (d) $= \frac{S-1}{\text{In (N)}}$

Station 1

Margalef's index (d) $= \frac{7-1}{\text{In (108)}}$

$d = \frac{6}{4.682}$

$d = 1.2815$

Station 2

Margalef's index (d) $= \frac{6-1}{\text{In }(102)}$

d $= \frac{5}{4.6240}$

d =1.0811

Station 3

Margalef's index (d) $= \frac{5-1}{\text{In }(92)}$

d $= \frac{4}{4.5218}$

d =0.8446

Menhinicks index (D) $= \frac{S}{\sqrt{N}}$

Station 1

D $= \frac{7}{\sqrt{108}}$

D $= \frac{7}{10.39}$

D = 0.6737

Station 2

D $= \frac{6}{\sqrt{102}}$

D $= \frac{6}{10.09}$

D = 0.5946

Station 3

D $= \frac{5}{\sqrt{98}}$

D $= \frac{5}{9.8995}$

D = 0.5051

H= $\frac{N\text{Log } N-\sum_{i=1}^{S} F_1 \log F_1}{N}$

Station 1

Species	Fi	Log Fi	FiLogFi
E. coli	21	1.3222	27.7662
Salmonella typhi	8	0.9030	7.2240
Shigella sp	18	1.2552	22.5936
Pseudomonas aureus	17	1.2304	20.9168
Streptococcus sp.	20	1.3010	26.0200
Staphylococcus sp	10	**1.0000**	10.0000
Klebsilla sp.	14	1.1461	16.0454
	N=108		
	S=7		**130.5660**

H= $\frac{N\text{Log } N-\sum_{i=1}^{S} F_1 \log F_1}{N}$

H= $\frac{108\text{Log}108-130.5660}{108}$

H= $\frac{108(2.0334)-130.5660}{108}$

H= $\frac{219.6072-130.5660}{108}$

$H = \frac{89.0412}{108}$

$H = 0.8244$

Station 2

Species	Fi	LogFi	FiLogFi
E. coli	19	1.2787	24.2953
Salmonella typhi	23	1.3617	31.3191
Shigella sp	7	0.8450	5.9150
Pseudomonas aureus	16	1.2041	19.2656
Streptococcus sp.	24	1.3802	33.1248
Staphylococcus sp	**0**	0	0
Klebsilla sp.	13	1.1139	14.4807
	N=102		**128.4005**
	S=6		

$H = \frac{N\text{Log } N - \sum_{i=1}^{S} F_1 \log F_1}{N}$

$H = \frac{102\text{Log}102 - 128.4005}{102}$

$H = \frac{102(2.0086) - 128.4005}{102}$

$H = \frac{204.8772 - 128.4005}{102}$

$H = \frac{76.4767}{102}$

$H = 0.7497$

Station 3

Species	Fi	LogFi	FiLogFi
E. coli	20		
Salmonella typhi	0		
Shigella sp	23		
Pseudomonas aureus	0		
Streptococcus sp.	23		
Staphylococcus sp	12		
Klebsilla sp.	14		
S	**5**		
N	**92**		

H = 0.6848

Station 1

Species	Fi	Log Fi	FiLogFi	$(F_1 \log F_1)^2$	$(F_1 \log$ /Fi
E. coli	21	1.3222	27.7662	770.9618	**36.7124**
Salmonella typhi	8	0.9030	7.2240	52.1861	**6.5232**
Shigella sp	18	1.2552	22.5936	510.4700	**28.3594**
Pseudomonas aureus	17	1.2304	20.9168	437.5125	**25.7360**
Streptococcus sp.	20	1.3010	26.0200	677.0404	**33.8520**
Staphylococcus sp	10	**1.0000**	10.0000	100.0000	**10.000**

Klebsilla sp.	14	1.1461	16.0454	257.4548	**18.3896**
	N=108				
	S=7		**130.5660**		

$F_1\log^2F_1 = 159.5726$

Where S^2H =Variance of $H = \frac{\sum F_1\log^2F_1 - \frac{(\sum F_1\log F_1)^2}{N}}{N^2}$

Where S^2H =Variance of $H = \frac{159.5726 - \frac{(130.5660)^2}{108}}{108^2}$

S^2H =Variance of $H = \frac{159.5726 - \frac{17047.480}{108}}{11664}$

S^2H =Variance of $H = \frac{159.5726 - 157.847}{11664}$

S^2H =Variance of $H = \frac{1.7255}{11664}$

S^2H =Variance of $H = 0.000147$

Station 2

Species	Fi	LogFi	FiLogFi	$(F_1\log F_1)^2$	$(F_1\log F_1)^2/$ Fi
E. coli	19	1.2787	24.2953	590.2616	31.0664
Salmonella typhii	23	1.3617	31.3191	980.8860	42.6472
Shigella sp	7	0.8450	5.9150	34.9872	4.9981
Pseudomonas aureus	16	1.2041	19.2656	371.1633	23.1977
Streptococcus sp.	24	1.3802	33.1248	1097.2523	45.7188
Staphylococcus sp	**0**	0	0	0.0000	0
Klebsilla sp.	13	1.1139	14.4807	209.6907	16.1300
	N=102		**128.4005**		**163.7582**
	S=6				

$F_1\log^2F_1 = 163.7582$

Where S^2H =Variance of $H = \frac{\sum F_1\log^2F_1 - \frac{(\sum F_1\log F_1)^2}{N}}{N^2}$

Where S^2H =Variance of $H = \frac{163.7582 - \frac{(128.4005)^2}{102}}{102^2}$

Where S^2H =Variance of $H = \frac{163.7582 - \frac{16486.6884}{102}}{10404}$

Where S^2H =Variance of $H = \frac{163.7582 - 161.6342}{10404}$

Where S^2H =Variance of H $= \frac{2.1240}{10404}$

S^2H =Variance of H $= 0.00020415$

$t = \frac{H_1 - H_2}{\sqrt{H_1S^2 - H_2S^2}}$

Where S^2H =Variance of H $= \frac{\sum F_1 \log^2 F_1 - \frac{(\sum F_1 \log F_1)^2}{N}}{N^2}$

$F_1\log^2 F_1 = \frac{(F_1 \log F_1)^2}{F_1}$

$t = \frac{H_1 - H_2}{\sqrt{H_1S^2 - H_2S^2}}$

$t = \frac{0.8244 - 0.7497}{\sqrt{0.000147 - 0.00020415}}$

$t = \frac{0.0747}{\sqrt{-0.00005715}}$

$t = \frac{0.0747}{0.007559}$

$t = 9.8822$

Degree of freedom are calculated using the equation

$df = \frac{(S^2H_1 + S^2H_2)^2}{\frac{(S^2H_1)^2}{N_1} + \frac{(S^2H_2)^2}{N_2}}$

$df = \frac{(S^2H_1 + S^2H_2)^2}{\frac{(S^2H_1)^2}{N_1} + \frac{(S^2H_2)^2}{N_2}}$

$df = \frac{(0.000147 + 0.00020415)^2}{\frac{(0.000147)^2}{108} + \frac{(0.00020415)^2}{102}}$

$df = \frac{(0.00035115)^2}{\frac{0.0000021609}{108} + \frac{0.00000004168}{102}}$

$df = \frac{0.0000001233}{0.000000020008 + 0.0000000004086}$

$df = \frac{0.0000001233}{0.0000000204166}$

df = 6.03920 = 6

Evenness

Station 1

$E = \frac{0.8244}{\text{Log } 7}$

$E = \frac{0.8244}{0.8450}$

E = 0.9756

Station 2

$E = \frac{H}{Hmax} = \frac{H}{\text{Log } S}$

$E = \frac{0.7497}{\text{Log } 6}$

$E = \frac{0.7497}{0.7781}$

E = 0.9635

Station 3 ???

D

Station 1

Species	ni	ni-1	ni(ni-1)
E. coli	21	20	420
Salmonella typhi	8	7	56
Shigella sp	18	17	306
Pseudomonas aureus	17	16	272
Streptococcus sp.	20	19	380
Staphylococcus sp	10	**9**	90
Klebsilla sp.	14	13	182
	N=108		
	S=7		**1706**

$D = \sum_{0=1}^{s} \frac{1706}{108(108-1)}$

$D = \frac{1706}{108(107)}$

$D = \frac{1706}{11556}$

D = 0.1476

Unit 5: Computer approach to Biostatistics

Unit Structure
5.1: Introduction
5.2: Intended Learning Outcomes
5.3: Statistical Software
5.4: Summary
5.5: References/Further Readings/Web Sources
5.6: Possible Answers to Self-Assessment Exercises

5.1: Introduction

The absence of adequate tools has been, for many years, an obstacle to the development of Multivariate and Multidimensional analyses as they were studied only in a theoretical context. Multidimensional analysis may be defined as a group of techniques that have the aim to visualize, classify and interpret the data. It tries to underline the latent structure of the data, removing the redundant information. Multidimensional Statistical analysis includes: Principal Component Analysis, Correspondence Analysis, Discriminant Analysis, Canonical Correlation Analysis and Cluster Analysis.
Simple Correspondence analysis is one of the most known tools for qualitative data. It studies the relationships between the modalities of two qualitative variables. Multiple Correspondence Analysis is used when there are more than 2 qualitative variables where the Simple Correspondence Analysis is not possible and the relationships between the characters are studied. Discriminant analysis is used to verify if the prior classification is confirmed after using the explicative variables. I.e., it classifies a new observation in one of the groups. Cluster analysis is a group of techniques that have the aim to classify observations or individuals in clusters. The observations in each cluster must be similar and the clusters must be well separated.

5.2: Intended Learning Outcomes

By the end of this unit, students should be able to:
- understand the different software for biostatistical analyses

5.3: Statistical Software

Some of the software used for Multidimensional Data Analysis includes: Excel, SPAD, Xl-stat, SPSS, GraphPad Prism, PSPP. Other software not

only for multidimensional analysis are Matlab, Stata, Eviews, Gauss. They are not open source and some of them perform only some techniques of multidimensional analysis.

SPSS is a statistical package developed especially for the social sciences but it has wide applicability in the areas of biology and agriculture. It can be used to execute some statistical procedures such as summaries, custom tables, Anova (Analysis of variance), correlation and regression analyses, non-parametric tests time series, create charts (Bar, line, Pie, Area, Histogram, Scatter etc.) and lots more.
The SPSS is a comprehensive package that is command based but commands list can be generated from menus and interactive operations are possible.
The MINITAB statistical software provides a wide range of statistical analysis and graphical capabilities. Like the SPSS, it is also a comprehensive command based statistical package. MINITAB can be used for various types of statistical analysis ranging from editing and manipulating data, basic statistics, arithmetic, regression, Anova, non-parametric test, exploratory data analysis etc.

The statistical package developed especially for social sciences is?
The statistical software that provides a wide range of statistical analysis and graphical capabilities is?

Self-Assessment Exercises

1. What is SPSS?
2. Multiple Correspondence Analysis is used when there are more than……..qualitative variables

5.4: Summary

The different types of biostatistics software were mentioned and discussed. The unit gave insight into some biostatistics software

5.5: References/Further Readings/Web Sources

- Bailey, N.T.J. (1994). *Statistical Methods in Biology*. Third Edition. Cambridge University Press. United Kingdom.
- Bluman, A.G. (2004). *Elementary Statistics*. A Step by Step Approach. Fifth Edition. McGraw-Hill Companies Incorporated.

London.

- Daniel, W.W. (1995). *Biostatistics*: a foundation for Analysis in Health sciences. Sixth Edition. John Wiley and sons Incorporated. USA.
- Helmut F. van Emden.(2008). *Statistics for Terrified Biologists.* Blackwell Publishing Limited. USA.

Biostatistics: A Computing Approach - 1st Edition - Routledge
https://www.routledge.com/Biostatistics-A-Computing-Approach/Anderson/p/book/9781584888345#:~:text=Biostatistics%3A%20A%20Computing%20Approach%20focuses,that%20can%20facilitate%20such%20understanding.

Biostatistics: A Computing Approach (Chapman & Hall ...
https://www.amazon.com/Biostatistics-Computing-Approach-Chapman-Hall/dp/1584888342

https://www.xlstat.com/en/

https://www.youtube.com/watch?v=4wrtykLDdus

https://www.graphpad.com/series/getting-started/

https://www.youtube.com/watch?v=M0Sl-3eu974

5.6: Possible Answers to Self-Assessment Exercises

1. **SPSS** is a statistical package developed especially for the social sciences but it has wide applicability in the areas of biology and agriculture. It can be used to execute some statistical procedures such as summaries, custom tables, Anova (Analysis of variance), correlation and regression analyses, non-parametric tests time series, create charts (Bar, line, Pie, Area, Histogram, Scatter etc.) and lots more.

2. two

Glossary

- **Data set:** A collection of related, discrete items of data that may be accessed individually or collectively, or managed as a single, holistic entity. Data sets are generally organized into some formal structure, often in a tabular format
- nonparametric tests: A test that makes minimal assumptions about the distribution of the data or about certain parameters of a statistical model.
- Pearson's correlation coefficient (r): this is used for any bivariate populations which are normally distributed.
- regression to the mean: Tendency for a variable that has an extreme value on its first measurement to have a more typical value on its second measurement.
- Spearman's rank correlation coefficient = r_s: this is non parametric rank correlation
- Species richness: The number of species within a region. (A term commonly used as a measure of species diversity, but technically only one aspect of diversity.)
- Statistical computing: The collection and interpretation of data aimed at uncovering patterns and trends. It may be used in scenarios such as gathering research interpretations, statistical modeling or designing surveys and studies, and advanced business intelligence. R is a programming language that's highly compatible with statistical computing.
- The Shannon evenness index, abbreviated as SEI, provides information on area composition and richness. It covers the number of different land cover types (m) observed along the straight line and their relative abundances (P_i). It is calculated by dividing the Shannon diversity index by its maximum (h (m)). Therefore, it varies between 0 and 1 and is relatively easy to interpret.

End of the module Questions

1. The relationship between two simple variables that are dependent on each other is called?
2. A dependent variable (Y) is the variable in regression that cannot be?
3. When there are two variables under study the correlation is termed?
4. The variable that increases with the other variable decrease or vice versa is called?
5. In a toxicology experiment, hybrid catfish was exposed to acute concentrations of neem leaf powder as presented in the table below;

Cumulative Mortality (%)	0	20	38	48	60	72	82	90	98	100
Conc. of neem powder (mg/L)	0	2	4	6	8	10	12	14	16	18

From the table; a. Plot a scattergram b. Calculate the gradient c. Determine the equation of the relationship

6. In a toxicology experiment, hybrid catfish was exposed to acute concentrations of neem leaf powder as presented in the table below;

Cumulative Mortality (%)	0	20	38	48	60	72	82	90	98	100
Conc. of neem powder (mg/L)	0	2	4	6	8	10	12	14	16	18

From the table; a. Pearson correlation and b. Coefficient of determination

7. The following data was obtained from the colony counts of bacteria in two sampling points.

Station A ($x 10^6$cfu)	3.56	3.67	3.98	3.80	3.76	3.98	2.56	3.23	3.52	4.32	2.67	4.
Station B ($x 10^6$cfu)	3.23	3.76	3.09	3.02	3.42	3.23	3.67	3.24	3.34	2.17	3.11	3.

Determine the correlation between the sampling points

8. Consider the scores of 22 students in Biostatistics II class thought by one of the two lecturers A and B but took the same examination. **Lecturer A: A, A, A,B^-, B^-, B^+, C^-, C^-, C, D** *Lecturer B: A, A, A,*

B, B, B^+, B^-, C, C, C^+, C^-, D^+, D. Test the null hypothesis that the students performed equally well in the course under both lecturers and prove the accuracy of the result.

9. A hydrobiologist studying the effluent characteristics of flow station at different locations collected five effluent samples from three flow stations for the determination of total hydrocarbon (THC) concentrations (mg/l). The result of the analysis are given below.

Location A	Location B	Location C
14.6	8.4	6.9
12.1	5.0	7.3
9.6	5.5	5.8
8.2	6.6	4.1
10.2	6.3	5.1

10. A hydrobiologist obtained eight samples of water from each of four forest ponds and the pH of each water sample was measured. The results are given below (one of the samples from pond 3 was lost) H_0 = pH is the same in all four ponds. H_A = pH is not the same in all four ponds.

Pond 1	Pond 2	Pond 3	Pond 4
7.68	7.69	7.74	7.71
7.70	7.70	7.75	7.71
7.72	7.71	7.77	7.74
7.73	7.73	7.78	7.79
7.73	7.74	7.80	7.84
7.76	7.74	7.81	7.85
7.78	7.78	7.81	7.81
7.80	7.81	-	7.91

11. Use the appropriate non-parametric test to analyze the effect of drugs A and B on the zone of inhibition of *Salmonella typhi*

Drug A	54.2	60.4	80.5	49.5	33.2	35.5	20.3	29.1	40.8	33.2	28.9	33.2	32.1
Drug B	80.3	99.9	50.5	75.5	60.2	105.1	25.4	19.5	30.1	34.4	46.3	49.4	60.1

12. Determine and test for the accuracy of your result when evaluating the differences between the samples of heights (m) of male and female students in your department as

Male	7.6	7.4	7.3	7.2	7.1	7.0	6.7	8.2	5.6	5.8	8.9
Female	6.9	6.8	6.6	6.5	6.4	6.2	6.9	7.2	6.6	-	-

13. Using the appropriate non-parametric test, determine the possibility that the pH sampled in all the four ponds are the same

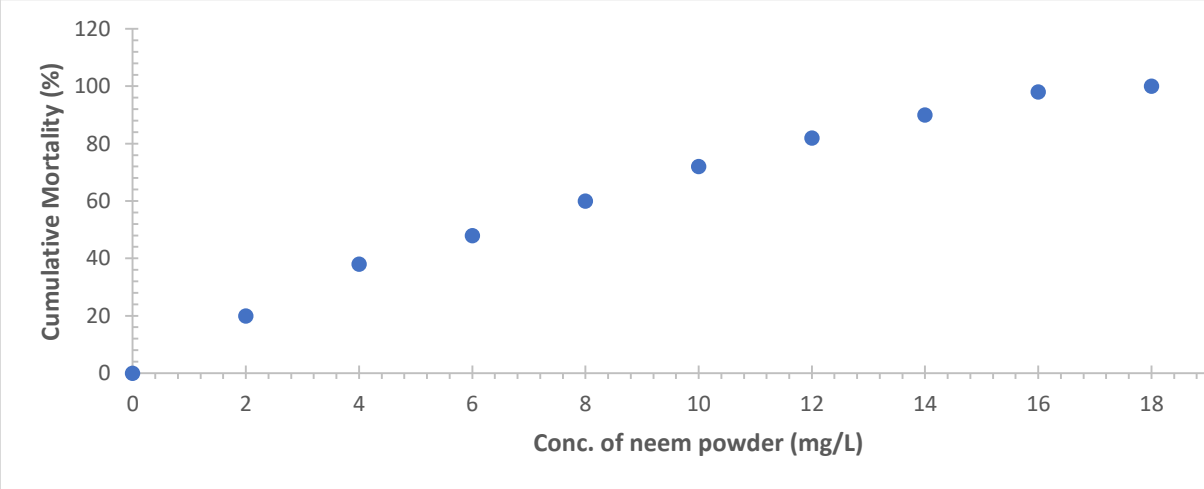
Pond A	7.68	7.70	7.72	7.73	7.73	7.76	7.78	7.80	7.81	8.23

Pond B	7.69	7.70	7.71	7.73	7.74	7.74	7.78	7.81	7.82	-
Pond C	7.74	7.75	7.77	7.78	7.80	7.81	7.81	-	-	-

14. Consider the scores of 22 students in Biostatistics II class thought by one of the two lecturers A and B but took the same examination. Lecturer A: A, A, A,B^-, B^-, B^+, C^-, C^-, C, D *Lecturer B: A, A, A, B, B, B^+, B^-, C, C, C^+, C^-, D^+, D.* Test the null hypothesis that the students performed equally well in the course under both lecturers and prove the accuracy of the result.
15. Which measurement of species diversity does take into consideration the number of individuals within a species or population?
16. What type of statistics can MINITAB be used for

Answers

1. Simple linear regression
2. Controlled/Manipulated
3. Simple
4. Negative simple relationship
5.



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

X	*Y*	*XY*	X^2
0	*0*	*0*	*0*
2	*20*	*40*	*4*
4	*38*	*152*	*16*
6	*48*	*288*	*36*
8	*60*	*480*	*64*
10	*72*	*720*	*100*
12	*82*	*984*	*144*
14	*90*	*1260*	*196*
16	*98*	*1568*	*256*
18	*100*	*1800*	*324*

$\sum 90 \quad \sum 608 \quad \sum 7292 \quad \sum 1140$

n = 10

$$\mathbf{b} = \frac{7292 - \frac{(90)(608)}{10}}{1140 - \frac{(90)^2}{10}}$$

$$\mathbf{b} = \frac{7292 - \frac{54720}{10}}{1140 - \frac{8100}{10}}$$

$$\mathbf{b} = \frac{7292 - 5472}{1140 - 810}$$

$$\mathbf{b} = \frac{1820}{330}$$

$$\mathbf{b} = 5.515$$

Y = a + bx

b=5.515

a = y – bx

mean y = 608/10 = 60.8

mean x = 90/10 = 9

a = 60.8 – 5.515 (9)

a= 60.8 – 49.636

a= 11.164

Y = 11.164 + 5.515x

6.

X	*Y*	*XY*	X^2
0	*0*	*0*	*0*
2	*20*	*40*	*4*
4	*38*	*152*	*16*
6	*48*	*288*	*36*
8	*60*	*480*	*64*
10	*72*	*720*	*100*
12	*82*	*984*	*144*
14	*90*	*1260*	*196*
16	*98*	*1568*	*256*
18	*100*	*1800*	*324*
$\sum 90$	$\sum 608$	$\sum 7292$	$\sum 1140$

n = 10

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

Y^2
0
400
1444
2304
3600
5184
6724
8100
9604
10000
$\sum 47360$

$$r = \frac{7292 - \frac{(90)(608)}{10}}{\sqrt{(1140 - \frac{(90)^2}{10})(47360 - \frac{(608)^2}{10})}}$$

$$r = \frac{7292 - \frac{54720}{10}}{\sqrt{(1140 - \frac{8100}{10})(47360 - \frac{369664}{10})}}$$

$$r = \frac{7292 - 5472}{\sqrt{(1140 - 810)(47360 - 36966.4)}}$$

$$r = \frac{1820}{\sqrt{(330)(10393.6)}}$$

$$r = \frac{1820}{\sqrt{3429888}}$$

$$r = \frac{1820}{1851.9957}$$

r = 0.9827

$r^2 = 0.9827^2$

$r^2 = 0.9657$

7.

Station 1	*Station 2*	*Rank 1*	*Rank 2*	*D*	dt^2
3.56	*3.23*	*6*	*7.5*	*-1.5*	*2.25*
3.67	*3.76*	*7*	*13*	*-6*	*36*
3.98	*3.09*	*10.5*	*3*	*7.5*	*56.25*
3.80	*3.02*	*9*	*2*	*7*	*49*
3.76	*3.42*	*8*	*11*	*-3*	*9*
3.98	*3.23*	*10.5*	*7.5*	*3*	*9*
2.56	*3.67*	*1*	*12*	*-11*	*121*
3.23	*3.24*	*4*	*9*	*-5*	*25*
3.52	*3.34*	*5*	*10*	*-5*	*25*
4.32	*2.17*	*13*	*1*	*12*	*144*
2.67	*3.11*	*2*	*4*	*-2*	*4*
4.21	*3.12*	*12*	*5*	*7*	*49*
3.12	*3.13*	*3*	*6*	*-3*	*9*
					538.5

$r_s = 1 - (\frac{6\sum dt^2}{n^3 - n})$

$r_s = 1 - (\frac{3231}{2184})$

$r_s = 1 - (\frac{6(538.5)}{13^3 - 13})$

$r_s = 1 - 1.4794$

$r_s = 1 - (\frac{3231}{2197 - 13}$

$r_s = -0.4794$

8

LA	LB	Rank LA	Rank LB
A1	**A6**	**3.5**	**3.5**
A2	**A5**	**3.5**	**3.5**
A3	**A4**	**3.5**	**3.5**
B-11	**B10**	12	9.5
B-12	**B9**	12	9.5
B+7	**B+8**	7.5	7.5
C-18	B-13	19	12
C-19	**C15**	19	16
C17	**C16**	16	16
D22	C+14	22.5	14
	C-20		19
	D+21		21
	D23		22.5
n1=10	n2=13	R1=118.5	R2=157.5

$U = n_1 n_2 + n_1((n_1 + 1))/2 - \mathbf{R_1}$

$U = 10x13 + 10\ ((10 + 1))/2 - 118.5$

$U = 130 + 10\ ((11))/2 - 118.5$

$U = 130 + 10x\ 5.5 - 118.5$

$U = 130 + 55 - 118.5$

$U = 185 - 118.5$

U = 66.5
$U^1 = n_1n_2 - U$
$U^1 = 10x13 - 66.5$
130 -66.5
$U^1 = 63.5$
$\mathbf{R_1+R_2 = (N\ (N+1))/2}$
$118.5+157.5 = (23\ (23+1))/2$
$276 = (23\ (24))/2$
$276 = 552/2$
276= 276

9

Location A	Location B	Location C	Rank A	Rank B	Rank C
14.6	8.4	6.9	15	11	8
12.1	5.0	7.3	14	2	9
9.6	5.5	5.8	12	4	5
8.2	6.6	4.1	10	7	1
10.2	6.3	5.1	13	6	3
			R1 = 64	R2= 30	R3= 26

$$\mathbf{H = \frac{12}{N(N+1)}\ \sum_{i=1}^{k} \frac{R_1^{\,2}}{n_1} - 3\ (N+1)}$$
$$\mathbf{H = \frac{12}{15(15+1)}\ \left(\sum_{i=1}^{k} \frac{64^2}{5} + \frac{30^2}{5} + \frac{26^2}{5}\right) - 3\ (15+1)}$$
$$\mathbf{H = \frac{12}{15(16)}\ \left(\sum_{i=1}^{k} \frac{4096}{5} + \frac{900}{5} + \frac{676}{5}\right) - 3\ (16)}$$
$$\mathbf{H = \frac{12}{240}\ (819.2 + 180 + 135.2) - 48}$$
$$\mathbf{H = 0.05\ (1134.4) - 48}$$
$$\mathbf{H = 56.72 - 48}$$
$$\mathbf{H = 8.72}$$

10

Pond 1	Pond 2	Pond 3	Pond 4	Rank 1	Rank 2	Rank 3	Rank4
7.68**1**	7.69**2**	7.74**1 4**	7.71**6**	1	2	**13.5***	**6***
7.70**3**	7.70**4**	7.75**1 6**	7.71**7**	**3.5***	**3.5***	16	**6***
7.72**8**	7.71**5**	7.77**1 8**	7.74**1 5**	8	**6***	18	**13.5***
7.73**9**	7.73**1 1**	7.78**2 1**	7.79**2 2**	**10***	**10***	20*	22
7.73**1 0**	7.74**1 2**	7.80**2 4**	7.84**2 9**	**10***	**13.5***	23.5*	29
7.76**1 7**	7.74**1 3**	7.81**2 6**	7.85	17	**13.5***	26.5*	30

7.781 **9**	7.782 **0**	7.812 **7**	7.812 **8**	20*	20*	26.5*	26.5*
7.802 **3**	7.812 **5**	-	7.91	23.5*	26.5*		31
n=8	n=8	n=7	n=8	R1=93	R2=95	R3=144	R4=164

$$\mathbf{H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_1^2}{n_1} - 3\,(N+1)}$$

$$\mathbf{H = \frac{12}{31(31+1)} \left(\sum_{i=1}^{k} \frac{93^2}{8} + \frac{95^2}{8} + \frac{144}{7} + \frac{164^2}{8}\right) - 3\,(31+1)}$$

$$\mathbf{H = \frac{12}{31(32)} \left(\sum_{i=1}^{k} \frac{8649}{8} + \frac{9025}{8} + \frac{20736}{7} + \frac{26896}{8}\right) - 3\,(32)}$$

$$\mathbf{H = \frac{12}{992} \left(1081.125 + 1128.125 + 2962.285 + 3362\right) - 96}$$

$$\mathbf{H = 0.01209 \ \ (8533.535) - 96}$$

$$\mathbf{H = 103.17 - 96}$$

$$\mathbf{H = 7.170}$$

Groups (M)	1	2	3	4	5	6	7
Tied Ranks	3.5 3.5	6 6	10 10	13.5 13.5	20 20	23.5 23.5	26.5 26.5
		6	10	13.5	20		26.5
				13.5			26.5
No. of tied Ranks	**2**	**3**	**3**	**4**	**3**	**2**	**4**

$$Hc = \frac{H}{C}$$

$$\mathbf{C = \text{correction factor} = 1 - \frac{\sum T}{N^3 - N}}$$

$$\mathbf{\sum T = \sum_{i=1}^{m} (t_1^{\,3} - t_1)}$$

Where;

T=number of ties per group

df =K─1

$$\mathbf{\sum T = (2^3 - 2^1) + (3^3 - 3^1) + (3^3 - 3^1) + 4^3 - 4^1 + 3^3 - 3^1 + 2^3 - 2^1 + 4^3 - 4^1}$$

$$\mathbf{\sum T = 8 - 2 + 27 - 3 + 27 - 3 + 64 - 4 + 27 - 3 + 8 - 2 + 64 - 4}$$

$$\mathbf{\sum T = 6 + 24 + 24 + 60 + 24 + 6 + 60}$$

$$\mathbf{\sum T = 204}$$

$$\mathbf{C = \text{correction factor} = 1 - \frac{204}{31^3 - 31}}$$

$$\mathbf{C = \text{correction factor} = 1 - \frac{204}{29791 - 31}}$$

$$\mathbf{C = \text{correction factor} = 1 - \frac{204}{29760}}$$

$$\mathbf{C = \text{correction factor} = 1 - 0.00685}$$

$$\mathbf{C = 0.9932}$$

$$Hc = \frac{H}{C}$$

$$Hc = \frac{7.170}{0.9932}$$

Hc = 7.219

11

Drug A (X1)	Drug B (X2)	*X1 –X2 (di)*	Ranks of di	***Signed Rank***
54.2	*80.3*	*-26.1*	*8*	*-8*
60.4	*99.3*	*-38.9*	*12*	*-12*
80.5	*50.5*	*30*	*11*	*11*
49.5	*75.5*	*-26*	*7*	*-7*
33.2	*60.2*	*-27*	*9*	*-9*
35.5	*105.1*	*-69.6*	*13*	*-13*
20.3	*25.4*	*-5.1*	*2*	*-2*
29.1	*19.5*	*9.6*	*3*	*3*
40.8	*30.1*	*10.7*	*4*	*4*
33.2	*34.4*	*-1.2*	*1*	*-1*
28.9	*46.3*	*-17.4*	*6*	*-6*
33.2	*49.4*	*-16.2*	*5*	*-5*
32.1	*60.1*	*-28*	*10*	*-10*

Wilcoxon (T^I) = m (n+1) — T

where m= number of ranks with less frequent sign; ***T= sum of ranks with less frequent sign;*** *n= total number of sample*

$T^I = 3\ (13+1) — 18$ $T^I = 3\ (14) — 18$

$T^I = 42 — 18$

$T^I = 24$

12.

Male	*Female*	*Rank male*	*Rank female*
7.6	*6.9*	*18*	*10.5*
7.4	*6.8*	*17*	*9*
7.3	*6.6*	*16*	*6.5*
7.2	*6.5*	*14.5*	*5*
7.1	*6.4*	*13*	*4*
7.0	*6.2*	*12*	*3*
6.7	*6.9*	*8*	*10.5*
8.2	*7.2*	*19*	*14.5*
5.6	*6.6*	*1*	*6.5*
5.8	-	*2*	
8.9	-	*20*	
n1 = 11	***n2 = 9***	***R1 =140.5***	***R2 =69.5***

$U = n_1 n_2 + n_1 \frac{(n1+1)}{2} - R_1$

$U = (11)(9) + 11 \frac{(11+1)}{2} - 140.5$

$U = 99 + 11 \frac{12}{2} - 140.5$

$U = 99 + 11(6) - 140.5$

$U^1 = 99 - 24.5$

$U^1 = 74.5$

$R_1 + R_2 = \frac{N\,(N+1)}{2}$

U = 99 +66 – 140.5
U = 165 – 140.5
U = 24.5
$U^1 = n_1n_2 - U$
$U^1 = (11)(9) - (24.5)$

$$140.5+69.5 = \frac{20\,(21+1)}{2}$$
$$210 = \frac{20\,(21)}{2}$$
$$210 = \frac{420}{2}$$
$$210 = 210$$

13.

Pond A	*Pond B*	*Pond C*	*Rank A*	*Rank B*	*Rank C*
7.68	***7.69***	***7.74***	*1*	*2*	*11**
7.70	***7.70***	*7.75*	*3.5**	*3.5**	*13*
7.72	***7.71***	*7.77*	*6*	*5*	*15*
7.73	***7.73***	*7.78*	*8**	*8**	*17**
7.73	***7.74***	*7.80*	*8**	*11**	*19.5**
7.76	***7.74***	*7.81*	*14*	*11**	*22.5**
7.78	*7.78*	*7.81*	*17**	*17**	*22.5**
7.80	*7.81*	-	*19.5**	*22.5**	-
7.81	*7.82*	-	*22.5**	*25*	-
8.23	-	-	*26*	-	-
			R1= 125.5	*R2=105*	*R3=120.5*

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_1{}^2}{n_1} - 3\,(N+1)$$

Where n = number of observations in a sample (group i)
N = $\sum_{i=1}^{k}$ (total number of observations in all i= 1)
R_1 = sum of ranks of ni observation in group i

$$H = \frac{12}{26(26+1)} \sum_{i=1}^{k} \frac{125.5^2}{10} + \frac{105^2}{9} + \frac{120..5^2}{7} - 3\,(26+1)$$
$$H = \frac{12}{26(27)} \sum_{i=1}^{k} \frac{1575.25}{10} + \frac{11025}{9} + \frac{14520.25}{7} - 3\,(27)$$
$$H = \frac{12}{702} \sum_{i=1}^{k} 1575.525 + 1225 + 2074.3214 - 3\,(27)$$
$$H = \frac{12}{702} \sum_{i=1}^{k} 4874.345 - 81$$
$$H = 0.017094\ (4878.345) - 81$$
$$H = 83.3221 - 81$$
$$H = 2.3221$$

14.

LA	*LB*	*Rank LA*	*Rank LB*
A	***A***	***3.5***	***3.5***
A	***A***	***3.5***	***3.5***
A	***A***	***3.5***	***3.5***
B-	***B***	***12***	***9.5***
B-	***B***	***12***	***9.5***
B+	***B+***	***7.5***	***7.5***
C-	***B-***	***18***	***12***

C-	***C***	***18***	***15.5***
D	***C***	***21.5***	***15.5***
	C+		***14***
	C-		***18***
	D+		***20***
	D		***21.5***
n1=9	***n2=13***	***R1=99.5***	***R2=153.5***

$U = n_1 n_2 + n_1 \frac{(n1+1)}{2} - R_1$

$U = (9)(13) + 9 \frac{(9+1)}{2} - 99.5$

$U = 117 + 9 \frac{10}{2} - 99.5$

$U = 117 + 9(5) - 99.5$

$U = 117 + 45 - 99.5$

$U = 162 - 99.5$

$U = 62.5$

$U^1 = n_1 n_2 - U$

$U^1 = (9)(13) - (62.5)$

$U^1 = 117 - 62.5$

$U^1 = 54.5$

$R_1 + R_2 = \frac{N\,(N+1)}{2}$

$99.5 + 153.5 = \frac{22\,(22+1)}{2}$

$253 = \frac{22\,(23)}{2}$

$210 = \frac{506}{2}$

$253 = 253$

15.Species Richness

16.MINITAB can be used for various types of statistical analysis ranging from editing and manipulating data, basic statistics, arithmetic, regression, Anova, non-parametric test, exploratory data analysis etc