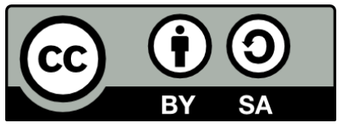
STA 131: INTRODUCTION TO STATISTICAL INFERENCE 1

(2 Credits)

Published by the Centre for Open and Distance Learning,
University of Ilorin, Nigeria

E-mail: codl@unilorin.edu.ng
Website: https://codl.unilorin.edu.ng



Course Development Team

Subject Matter Expert

Rasheed Gbenga JIMOH, Ph.D.
Department of Computer Science
University of Ilorin, Nigeria

Instructional Designers

Olawale Koledafe
Center for Open and Distance (CODL)
University of Ilorin, Nigeria

Miss. Damilola Adesodun
Department of Educational Technology,
University of Ilorin, Nigeria

Mr Olanrewaju O. Ismail
Department of Educational Technology,
University of Ilorin, Nigeria

Hassan Selim Olarewaju
Department of Educational Technology,
University of Ilorin, Nigeria

Language Editors

Bankole Ogechi Ijeoma
Center for Open and Distance (CODL)
University of Ilorin, Nigeria

From the Vice Chancellor

Courseware development for instructional use by the Centre for Open and Distance Learning (CODL) has been achieved through the dedication of authors and the team involved in quality assurance based on the core values of the University of Ilorin. The availability, relevance and use of the courseware cannot be timelier than now that the whole world has to bring online education to the front burner. A necessary equipping for addressing some of the weaknesses of regular classroom teaching and learning has thus been achieved in this effort.

This basic course material is available in different electronic modes to ease access and use for the students. They are available on the University's website for download to students and others who have interest in learning from the contents. This is UNILORIN CODL's way of extending knowledge and promoting skills acquisition as open source to those who are interested. As expected, graduates of the University of Ilorin are equipped with requisite skills and competencies for excellence in life. That same expectation applies to all users of these learning materials.

Needless to say, that availability and delivery of the courseware to achieve expected CODL goals are of essence. Ultimate attention is paid to quality and excellence in these complementary processes of teaching and learning. Students are confident that they have the best available to them in every sense.

It is hoped that students will make the best use of these valuable course materials.

Professor S. A. Abdulkareem
Vice Chancellor

Foreword

Courseware remains the nerve centre of Open and Distance Learning. Whereas some institutions and tutors depend entirely on Open Educational Resources (OER), CODL at the University of Ilorin considers it necessary to develop its own materials. Rich as OERs are and widely as they are deployed for supporting online education, adding to them in content and quality by individuals and institutions guarantees progress. Doing it in-house as we have done at the University of Ilorin has brought the best out of the Course Development Team across Faculties in the University. Credit must be given to the team for prompt completion and delivery of assigned tasks in spite of their very busy schedules.The development of the courseware is similar in many ways to the experience of a pregnant woman eagerly looking forward to the D-day when she will put to bed. It is customary that families waiting for the arrival of a new baby usually do so with high hopes. This is the apt description of the eagerness of the University of Ilorin in seeing that the centre for open and distance learning [CODL] takes off.

The Vice-Chancellor, Prof. Sulyman Age Abdulkareem, deserves every accolade for committing huge financial and material resources to the centre. This commitment, no doubt, boosted the efforts of the team. Careful attention to quality standards, ODL compliance and UNILORIN CODL House Style brought the best out from the course development team. Responses to quality assurance with respect to writing, subject matter content, language and instructional design by authors, reviewers, editors and designers, though painstaking, have yielded the course materials now made available primarily to CODL students as open resources.

Aiming at a parity of standards and esteem with regular university programmes is usually an expectation from students on open and distance education programmes. The reason being that stakeholders hold the view that graduates of face-to-face teaching and learning are superior to those exposed to online education. CODL has the dual-mode mandate. This implies a combination of face-to-face with open and distance education. It is in the light of this that our centre has developed its courseware to combine the strength of both modes to bring out the best from the students. CODL students, other categories of students of the University of Ilorin and similar institutions will find the courseware to be their most dependable companion for the acquisition of knowledge, skills and competences in their respective courses and programmes.

Activities, assessments, assignments, exercises, reports, discussions and projects amongst others at various points in the courseware are targeted at achieving the objectives of teaching and learning. The courseware is interactive and directly points the attention of students and users to key issues helpful to their particular learning. Students' understanding has been viewed as a necessary ingredient at every point. Each course has also been broken into modules and their component units in sequential order.

At this juncture, I must commend past directors of this great centre for their painstaking efforts at ensuring that it sees the light of the day. Prof. M. O. Yusuf, Prof. A. A. Fajonyomi and Prof. H. O. Owolabi shall always be remembered for doing their best during their respective tenures. May God continually be pleased with them, Aameen.

Bashiru, A. Omipidan
Director, CODL

INTRODUCTION

I welcome you to Internet Technology I, a second-semester course. Internet Technology I is a two (2) unit course that provides a general introduction to Internet Technology, covering a brief history of the Internet, how it grew from its humble origins into the worldwide network that is available today, identifying the most popular Internet services such as information retrieval, WWW and communication services.

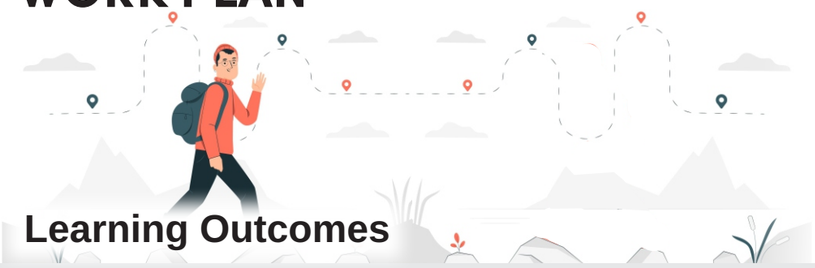
The relationship between the Internet and the World Wide Web is discussed. The course provides you with comprehensive knowledge on the concepts of Internet Technology, which include its internet architecture and internet protocol. The two most important protocols that allow networks to communicate with one another and exchange information, that is the TCP (Transmission Control Protocol) and IP (Internet Protocol), are also discussed.

Also, the functions of each layer at the TCP/IP networking model are covered. The brief history of HTML, XML, XHTML and DHTML is addressed. The course also covers in depth, HTML5, CSS and Javascript. The course also discusses the concept of a markup language and how to create web pages using HTML5 elements, CSS and Javascript. the course also discusses other equally important topics like WYSIWYG, Test Editors, including notepad, notepad++ and others.

Course Goal

In your voyage through this course, you will gain familiarity and some level of confidence in identifying the appropriate basic statistical analysis for various data types resulting from a research undertaking. Participants will learn to perform the statistical analysis with the aid of micro-con

WORK PLAN



Learning Outcomes

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

- Differentiate between various kinds of data;
- Apply the relevant analytical or graphical tools require for different data type;
- Differentiate different measures of location;
- Recognize different measures of spread;
- Interpret simple statistical concepts and
- Analyze data and draw conclusions for policy making.

Week 01 → Week 02 → Week 03

Course Guide

Module 1

Unit 1: Nature of Statistics

Unit 2: Definitions of terms in Statistics

Unit 3: Levels of Measurement

Unit 4: Types of Sampling

Unit 5: Study Questions

Module 2

Unit 1: Frequency Distribution

Unit 2: Graphical Presentation

Unit 3: Working Example

Module 3

Unit 1: Measure of Central tendency

Unit 2: Mean

Unit 3: Median

Unit 4: Mode

Unit 5: Midrange

Unit 6: Geometric Mean

Unit 7: Harmonic Mean

Module 4

Unit 1: Measure of Spread

Unit 2: Range

Unit 3: Variance

Unit 4: Standard Deviation

Course Infomation

The goal of this course is to give you the basic knowledge on Data Mining and Data Warehousing and introduce them to its application in Computer Science.

Pre-requisite | **CSC 319** Introduction to Digital Design and Microprocessors

Module 5

Unit 1: Measure of Position

Unit 2: Standard Score

Unit 3: Percentiles, Deciles and Quartiles

Unit 4: Study Questions

Unit 5: Assignment

Module 6

Unit 1: Measure of Shapes

Unit 2: Bell shape

Unit 3: Measure of Skewness

Unit 4: Measure of Kurtosis

Module 7

Unit 1: Regression and Correlation

Unit 2: Simple linear Regression Model

Unit 3: Correlation

Course Requirements

Requirements for success

The CODL Programme is designed for learners who are absent from the lecturer in time and space. Therefore, you should refer to your Student Handbook, available on the website and in hard copy form, to get information on the procedure of distance/e-learning. You can contact the CODL helpdesk which is available 24/7 for every of your enquiry.

Visit CODL virtual classroom on http://codllms.unilorin.edu.ng. Then, log in with your credentials and click on CSC 427. Download and read through the unit of instruction for each week before the scheduled time of interaction with the course tutor/facilitator. You should also download and watch the relevant video and listen to the podcast so that you will understand and follow the course facilitator.

At the scheduled time, you are expected to log in to the classroom for interaction.

Self-assessment component of the courseware is available as exercises to help you learn and master the content you have gone through.

You are to answer the Tutor Marked Assignment (TMA) for each unit and submit for assessment.

Embedded Support Devices

Support menus for guide and references

Throughout your interaction with this course material, you will notice some set of icons used for easier navigation of this course materials. We advise that you familiarize yourself with each of these icons as they will help you in no small ways in achieving success and easy completion of this course. Find in the table below, the complete icon set and their meaning.

 **Introduction**	 **Learning Outcomes**	 **Main Content**

 **Summary**	 **Tutor Marked Assignment**	 **Self Assessment**
 **Web Resources**	 **Downloadable Resources**	 **Discuss with Colleagues**
 **References**	 **Futher Reading**	 **Self Exploration**

Grading and Assessment

TMA

CA

Exam

Total

01 | **Picture:** Statistics

Photo: Pixels.com

Module 1

02 | **Picture:** Statistics

Photo: Pixels.com

UNIT 1

Nature of Statistics

Introduction

In this unit I will explain what variable is, the important of the course under study, definitions of terms, Data, Different scales of measurement and diferent sampling methods.

Learning Outcomes

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

1. The definition of some statistical terms
2. what data is, types and source of data
3. scales of measurement
4. different sampling methods

Main Content

Introduction

It's understandable many of you don't really have the basic understanding of statistics and here we'll go through the bit of what statistics is all about. Statistics play an essential role in all stages of a research project- from design up to analysis and interpretation.

The course will provide you basic understanding on the quantitative methods of research. The course takes an insight both research methodology and statistical procedures. Topics we are going to cover overview of the research process, sampling methods, exploratory data analysis and introduction to descriptive and inferential statistics. We'll discuss and interact about research methodology and statistical procedures through the use of lecture and the great interest you'll find in the course as we going deep bit by bit in the course. Individual and group workshops are part of the training curriculum which will serve to reinforce the learning gained from the lectures. We are going to make use of a statistical software for better and faster calculation and delivery of outputs. We are going to make use of output presentations to ensure that the concepts and the techniques are correctly understood.

Definitions of terms in Statistics

3 mins

SAQ 1&8

Statistics is the science dealing with the development of scientific procedures of

(I) Collection
(ii) Organisation
(iii) analysis and
(iv) Interpretation of results of analysis of statistical data

Statistics

This is the collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.

Variable

Is a characteristic or attribute that can assume different values.

Random Variable

A variable whose values are determined by chance. It is real valued function, whose values can be taken within a define range.

Population

All subjects possessing a common characteristic that is being studied.

Sample

A fractional part, subgroup, representative or subset of the population.

Parameter

Characteristic or measure obtained from a population.

Statistic (not to be confused with Statistics)

Characteristic or measure obtained from a sample.

Descriptive Statistics

This is the collection, organization, summarization, and presentation of data. In most cases, descriptive statistics are used to examine or explore one variable at a time. However, the relationship between two variables can also be described as with correlation and regression.

Inferential Statistics

Generalizing from samples to populations using probabilities. Performing hypothesis testing, determining relationships between variables, and making predictions. Inferential statistics can help us decide if a difference or relationship can be considered real or just a chance fluctuation.

It is important for you to know that population includes all objects of interest whereas the sample is only a portion of the population. Parameter is associated with population and statistic with sample. Parameters are usually denoted using Greek letters (μ, σ) while statistic is usually denoted using Roman letters (x, s).

There are several reasons why we don't work with populations. They are usually large, and it is often impossible to get data for every object we are studying. Sampling does not usually occur without cost, and the more items surveyed, the larger the cost.

We compute statistics and use them to estimate parameters. The computation is the first part of the statistics course (Descriptive Statistics) and the estimation is the second part (Inferential Statistics)

Qualitative Variables

This are variables which assume non-numerical values. For example: Colour, Sex, Marital Status, Number Plates, Matriculation Number and so on so on

Quantitative Variables

This are variables which assume numerical values. For example: Age, Height, Temperature, Weight, Score in a test, IQ,and

Most of the research you will encounter is of the quantitative type and that is what we will be dealing with in this unit. In such research, we rely on measuring variables and comparing groups on those variables, or examining the strength of the relationship between two or more variables. The belief here is that objectivity in the data collection process is paramount. Whoever was repeating this study or using the same instruments and methods would get approximately the same numbers.

However,you should know that another branch of research uses more qualitative approaches. These approaches employ more subjective approaches and frequently use interviews, focus groups, or single case designs, that lack objective measurement or have restricted generalisability. However, these methods are becoming more widely used these days as analysis methods improve and people search for better ways of gathering data about a problem. Focus groups recruit six to eight participants into a group in which the researcher has a structured set of questions which direct the discussion in the group to the research question. Usually, the whole discussion has to be tape-recorded or video-recorded and all interactions transcribed. The researchers then have to go back over the transcripts and extract the information they need. This can be a quite subjective, laborious, and costly process, but with a standardised set of guidelines, specific training, and greater familiarity with the technique, the considerable richness of these methods has been able to be tapped.

Discrete Variables

This are variables which assume a finite or countable number of possible values and usually obtained by counting. There are a finite or countable number of choices available with discrete data. You can't have 2.63 people in the room.

Example: The number of children in a family, The number of accidents in a day, The number of students in class, and so on.

Continuous Variables

This are variables which assume an infinite number of possible values and usually obtained by measurement. With the said above you should know a boundary is depending on the number of decimal places. For example: 64 is really anything 63.5 <= x < 64.5. Take for an example there are two decimal places, then 64.03 is really anything 63.025 <= x < 63.035. Boundaries always have one more decimal place than the data and end in a 5.

For example: Weight, Heigth, Lenght, width, Score in test, Time,Temperature and so on

SAQ 2,3&4

Data

2 mins

With the past knowlegde you have accumulated, the word data shouldnt be new to you any longer. Data is the collection of raw fact and the plural of the Latin word datum. In pratice the word data is used both as singular and plural in English. We shall use it in both senses. Some of the time, we shall talk of a set of data. By a data (set of data) we mean a set of information about certain individual, objects, or persons. Its cogent you know and note that data is the statistician's name for information. Take for an example,the set of information, containing such items as name, sex date of birth, height, weight, religion, nationality of each member of a class of students offering STA131 is a set of data.

Numerical Data: A data is numerical if the variable of interest has numerical values. For example your age, weight, height, and so on. However, If the values are non-numerical as in the case of sex, religion, nationality or social-class. Then some random variable may be defined on the set of values to extract the relevant information (of interest) in numerical data form. For example on sex (Male=0, Female=1), but such are just for classification and no arithemetic operation can be perform on them.

Statistical Data: A data is said to be a statistical data if it is numerical. A qualitative (categorical) data is made into as statistical data by defining an appropriate random variable X on its set of value. Examples of non-statistical on data that is converted as statistical data are: sex (Male=0, Female=1), Religion (Christianity=0, Islam=1, Others= 2).

Some Sources of Data

Some sources of collecting data are:

(I) Questionnaire by post
(ii) Direct through questionnaire
(iii) Official records
(iv) Business activity records
(v) Census
(vi) Statistical survey
(Vii) Planned experiment

Type of Data

You need to note that there are two types of data available for statistical activities. These are;

(I) **Primary data:** A data which an investigator collected for a purpose of enquiry into a particular problem and used for the same purpose enquiry. This is done through statistical survey or direct observation into the problem.

(ii) **Secondary data:** A data which is not originated by the investigator, but which the investigator obtains from someone else's records is called secondary data. Examples are data collected from CBN on price of commodities or on inflation rates, FRSC on number of accident, Police station on crimes committed over time.

SAQ 5,6&7

Levels of Measurement or Scales of Measurement

1 min

You need take note as well that we have four levels of measurement: Nominal, Ordinal, Interval, and Ratio. These go from lowest level to highest level. Data is classified according to the highest level which it fits. Each additional level adds something the previous level didn't have.

Nominal Level

This is for identification. It classifies data into mutually exclusive, all-inclusive categories in which no order or ranking can be imposed on the data. Nominal is the lowest level. Only names are meaningful here. Nominal variables classify data into categories. This process involves labelling categories and then counting frequencies of occurrence. Examples of this scale are: Colour, Number Plate, Your name (John or Hammed), Marital status (Single Or Married), Occupation, and so on

Ordinal Level

This has an additional property of ordering. It classifies data into categories that can be ranked. Differences between the ranks do not exist. Ordinal adds an order to the names. Ordinal variables order (or rank) data in terms of degree. Ordinal variables do not establish the numeric difference between data points. They indicate only that one data point is ranked higher or lower than another. Examples of this scale are: Levels in the University (100, 200, 300, 400,…), Grades a student can obtain in the University (1st class, 2nd class Upper, 2nd class lower, 3rd class, Pass),

Interval Level

This also has an additional property of "No Zero". It classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless. Interval adds meaningful differences. Interval variables score data. Thus the order of data is known as well as the precise numeric distance between data points. Examples are temperature, Intelligence quotient (IQ), Score in a test, and so on.

Ratio Level

This has an additional property of "True Zero". It classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure. Ratio adds a zero so that ratios are meaningful. Examples are height, age, weight and so on.

SAQ 9

Types of Sampling

1 min

There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

Random Sampling

Sampling in which the data is collected using chance methods or random numbers. Random sampling is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occurring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling.

Systematic Sampling

Sampling in which data is obtained when you select every kth object. Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every kth element is taken. This is similar to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.

Convenience Sampling

Sampling in which readily available data is used. You can easy carry out convenience sampling but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first set of people the surveyor runs into. It is require you know that convenient sampling is not advisable to use because the surveyor may as well be bias when selecting.

Stratified Sampling

Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques. Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. Take for an instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

Cluster Sampling

Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected. Cluster sampling is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

Summary

- Statistics is the collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.
- Data (set of data) we mean a set of information about certain individual, objects, or persons.
- There are four levels of measurement: Nominal, Ordinal, Interval, and Ratio

Self-Assessment Questions

SAQ

1. What is Variable?
2. What is data?
3. What are different ways of collecting data?
4. Distinguish between Primary data and Secondary data?
5. Explain with examples different classes of scales of measurement?
6. Distinguish between ratio and interval scales using examples?
7. Distinguish between Nominal and ordinal scales using examples?
8. Distinguish between Categorical variable and Numerical variable?
9. What are the different methods of sampling?

Tutor Marked Assessment

Question 1: What did descriptive statistics consist of

- the subjects under study
- group of subjects selected from a population
- a statistician trying to make inferences from samples to a population
- the collection, organization, summarization and presentation of data
- probability through games of chance like gambling and football

Question 2: A difference between a sample and a population can be explained as

- a sample is a group of subjects selected from the population whereas a population consists of all the subjects
- a sample is a group of all the subjects from the population whereas a population consists of a selected part of the entire population
- a sample is small whereas a population is large
- a sample is a group of subjects selected whereas a population is a selected part of the entire entity
- a sample and a population can be the same

Question 3: Why should you study Statistics?

- For you to be able to read and understand the various statistical information in their own areas of studies and have adequate knowledge about the statistical vocabulary, symbols, concepts and procedures
- To help you in writing comprehensive models and statistical models especially the normal, binomial and Poisson distributions
- For you to carryout experiments and trials in their own fields of studies and argue all cases concerning the final year project correctly
- To be able to make meaningful statistical decisions in finance and marketing especially in the current global economic downturn
- To specialize and build you in decision and risk taking in the capital markets which is often confronted with uncertainty

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA

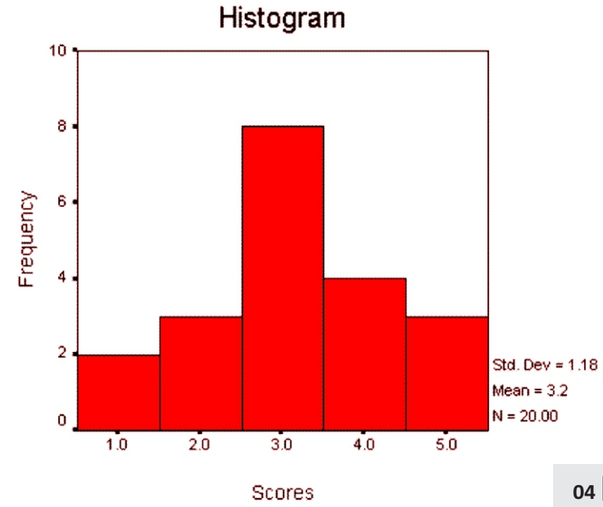
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.

03 | **Picture:** Frequency distribution

Photo: Wikipedia.com

Module 2

Frequency Distributions & Graphs



04 | **Picture:** Frequency distribution

Photo: Wikipedia.com

UNIT 1

SAQ 1&2

Frequency Distribution

3 mins

Introduction

In this unit I will explain how to describe data already collected or available, and how to contruct frequency distribution tables (categorical, ungrouped and grouped type).and how to contruct the table.

Learning Outcomes

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

1. The definition of some terms for contructing frequency distribution
2. Guildlines in contructing classes
3. Creating a grouped frequency distribution
4. Graphical representation of frequency distribution tables

Main Content

Frequency Distribution

3 mins

SAQ 1&2

- **Raw Data**
  Data collected in original form.
- **Frequency**
  The number of times a certain value or class of values occurs.
- **Frequency Distribution**
  The organization of raw data in table form with classes and frequencies.
- **Categorical Frequency Distribution**
  A frequency distribution in which the data is only nominal or ordinal.
- **Ungrouped Frequency Distribution**
  A frequency distribution of numerical data. The raw data is not grouped.
- **Grouped Frequency Distribution**
  A frequency distribution where several numbers are grouped into one class.
- **Class Limits**
  Separate one class in a grouped frequency distribution from another. The limits could actually appear in the data and have gaps between the upper limit of one class and the lower limit of the next.
- **Class Boundaries**
  Separate one class in a grouped frequency distribution from another. The boundaries have one more decimal place than the raw data and therefore do not appear in the data. There is no gap between the upper boundary of one class and the lower boundary of the next class. The lower class boundary is found by subtracting 0.5 units from the lower class limit and the upper class boundary is found by adding 0.5 units to the upper class limit.
- **Class Width**
  The difference between the upper and lower boundaries of any class. The class width is also the difference between the lower limits of two consecutive classes or the upper limits of two consecutive classes. It is not the difference between the upper and lower limits of the same class.
- **Class Mark (Midpoint)**
  The number in the middle of the class. It is found by adding the upper and lower limits and dividing by two. It can also be found by adding the upper and lower boundaries and dividing by two.
- **Cumulative Frequency**
  The number of values less than the upper class boundary for the current class. This is a running total of the frequencies.

- **Relative Frequency**
  The frequency divided by the total frequency. This gives the percent of values falling in that class.
- **Cumulative Relative Frequency (Relative Cumulative Frequency)**

  The running total of the relative frequencies or the cumulative frequency divided by the total frequency. Gives the percent of the values which are less than the upper class boundary. You should know that in this form of distribution refers to groups of values.

Guidelines for classes

(1) There should be between 5 and 20 classes.

(2) The class width should be an odd number. This will guarantee you that the class midpoints are integers instead of decimals.

(3) The classes must be mutually exclusive. You should beware that no data value can fall into two different classes

(4) The classes must be all inclusive or exhaustive. This means that you must include all data values.

(5) The classes must be continuous. You should be careful as well that there's no gaps in a frequency distribution. Classes that have no values in them must be included (unless it's the first or last class which are dropped).

(6) You must make classes be equal in width. The exception here is the first or last class. It is possible to have an "below ..." or "... and above" class. This is often used with ages.

Creating a Grouped Frequency Distribution

1. You should find the largest and smallest values
2. You should compute the Range = Maximum - Minimum
3. Select the number of classes desired. As we said above in the guidlines for classes (no1) that class is usually between 5 and 20.

4. After selecting your number of classes,you should find the class width by dividing the range by the number of classes and rounding up. There are two things to be careful of here. You must round up, not off. Normally 3.2 would round to be 3, but in rounding up, it becomes 4. If the range divided by the number of classes gives an integer value (no remainder), then you can either add one to the number of classes or add one to the class width. Sometimes you're locked into a certain number of classes because of the instructions.

5. Pick a suitable starting point less than or equal to the minimum value. You will be able to cover: "the class width times the number of classes" values. You need to cover one more value than the range. Follow this rule and you'll be okay: The starting point plus the number of classes times the class width must be greater than the maximum value. Your starting point is the lower limit of the first class. Continue to add the class width to this lower limit to get the rest of the lower limits.

6. For you to get your upper limit of the first class, subtract one from the lower limit of the second class. Then continue to add the class width to this upper limit to find the rest of the upper limits.

7. Find the boundaries by subtracting 0.5 units from the lower limits and adding 0.5 units from the upper limits. The boundaries are also half-way between the upper limit of one class and the lower limit of the next class. Depending on what you're trying to accomplish, it may not be necessary to find the boundaries.

8. Tally the data.

9. You should find the frequencies.

10. Find the cumulative frequencies. Depending on what you're trying to accomplish, it may not be necessary to find the cumulative frequencies.

11. If necessary, find the relative frequencies and/or relative cumulative frequencies.

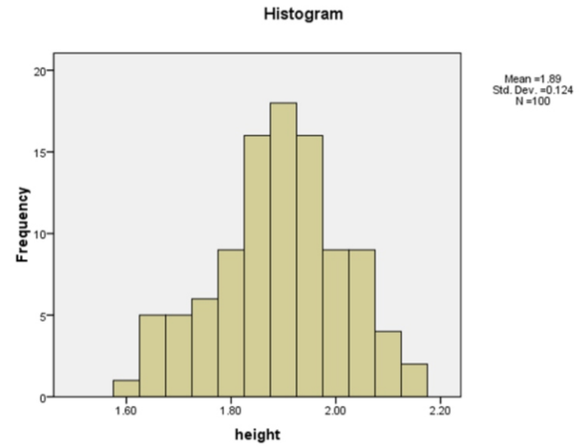
GRAPHICAL PRESENTATION

2 mins

Graphical presentation of Grouped data

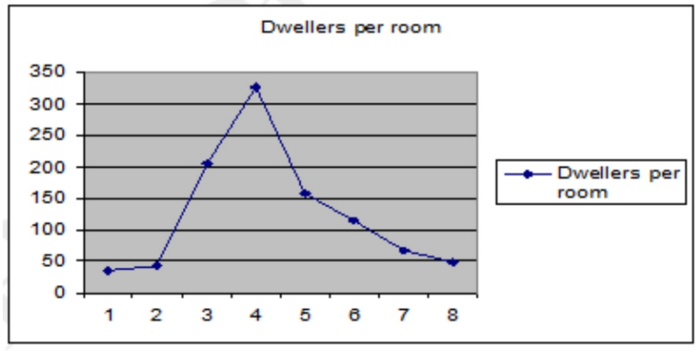
Histogram

This is a graph you enables you to display data by using vertical bars of various height to represent frequencies. The horizontal axis can be either the class boundaries, the class marks, or the class limits.



Frequency Polygon

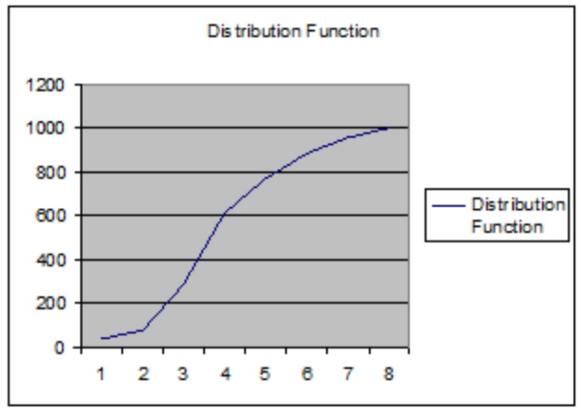
A line graph. The frequency is placed along the vertical axis and the class midpoints are placed along the horizontal axis. These points are connected with lines.



Ogive

A frequency polygon of the cumulative frequency or the relative cumulative frequency. The vertical axis the cumulative frequency or relative cumulative frequency. The horizontal axis is the class boundaries. The graph always starts at zero at the lowest class boundary and will end up at the total frequency (for a cumulative frequency) or 1.00 (for a relative cumulative frequency).

The chart of the cumulative distribution called Ogive



Graphical presentation of ungrouped data

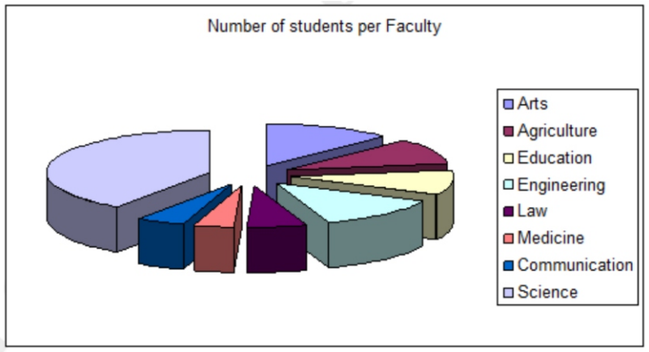
- **Pareto Chart**

This is a bar graph used for qualitative data with the bars arranged according to frequency and order of important or necessity.

- **Pie Chart**

Graphical depiction of data as slices of a pie. The frequency given will allow you to determines the size of the slice. The number of degrees in any slice is the relative frequency multiply by 360 degrees.

The chart given below is called Pie chart



Bar chart

This is a graph that enables you to use bar to represent data with or without space in between bars.

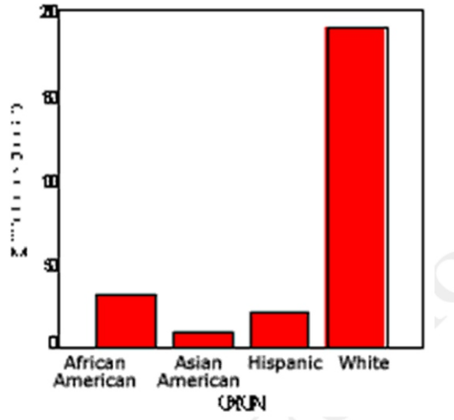


Figure 5: Ethnic category is a qualitative or categorical variable. You can see that most of the U.S. population is "White."

Its imperative that you know that bar charts can be used quite effectively with quantitative data as well but some problems occur, the next shows a bar chart of the sex partners data.

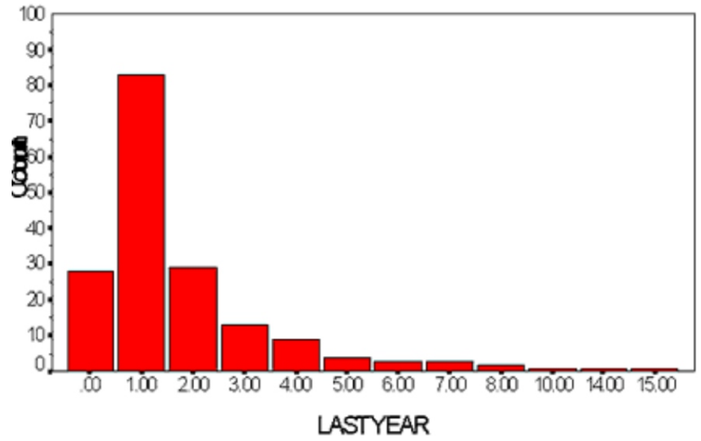


Figure 6: A bar chart of the "number of sex partners last year" variable.

Pictograph

A graph that uses pictures to represent data.

Stem and Leaf Plot

A data plot which uses part of the data value as the stem and the rest of the data value (the leaf) to form groups or classes. You can make use of this to sort data quickly.

Consider the following scores in an examination.
12,14,16,17,19,23,20,20,21,23,34,35,37,33,33,34,35,36,35,34,38,45,43,46,42,44,40,41
50,53,54,55

Stem	Leaf
1	2 4 6 7 9
2	0 0 1 3
3	3 3 4 4 4 5 5 5 7 8
4	0 1 2 3 4 5 6
5	0 3 4 5

Working Examples

1 min

Categorical data:

You should consider the following blood group of 25 participants in a training course at Macheal Imodu Training Institute.

A, A, B, B, A,

AB, A, A, O, AB,

B, A, B, AB, AB,

A, A, B, AB, B,

B, A, O, O, AB

You should then construct a frequency distribution table and obtain the relative and as well as the cumulative frequencies

Solution:

Group	Tally	Frequency	Relative frequency	Cumulative frequency	Proportion
A	~~1111~~ 1111	9	9/25	9	0.36
B	~~1111~~ 11	7	7/25	16	0.28
O	111	3	3/25	19	0.12
AB	~~1111~~ 1	6	6/25	25	0.24

Ungrouped data

Take for an example a man kept count of the number of gsm text messages he received each day over a period of 100 consecutive days. The frequency distribution of the observations is shown below.

No. of text messages received per day	0	1	2	3	4	5
Frequency	45	30	17	4	3	1
Relative Frequency	45/100	30/100	17/100	4/100	3/100	1/100
Cumulative Frequency	45	75	92	96	99	100
Proportion	0.45	0.30	0.17	0.04	0.03	0.01

Grouped data

Eighty Randomly selected light bulbs were tested to determine their lifetimes in hours. This frequency distribution was obtained.

Class boundaries	Tally	Frequency	Relative frequency	Cumulative frequency
52.5 – 63.5	~~1111~~ 1	6	6/80	6
63.5 – 74.5	~~1111 1111~~ 11	12	12/80	18
74.5 – 85.5	~~1111 1111 1111 1111 1111~~	25	25/80	43
85.5 – 96.5	~~1111 1111 1111~~ 111	18	18/80	61
96.5 – 107.5	~~1111 1111~~ 1111	14	14/80	75
107.5 – 118.5	~~1111~~	5	5/80	80
Total		80	1	

Study questions

1 min

Given the following data on the height of 100 students in Basic Statistics class.

1.83 1.90 2.05 1.98 1.86 1.80 1.85 2.15 1.97 2.02
2.02 1.93 1.88 1.65 1.98 1.89 2.04 1.91 1.84 1.84
1.95 1.88 1.92 1.64 1.89 2.06 1.65 1.91 1.84 2.04
1.86 1.79 1.73 1.72 1.91 2.12 1.90 1.64 2.01 2.04
1.75 1.78 1.70 1.92 1.85 1.69 1.88 1.95 2.00 2.03
1.99 1.95 1.87 1.83 1.89 2.07 1.97 1.88 1.93 1.60
2.10 1.92 1.97 1.79 1.75 1.95 2.00 1.93 1.87 1.69
1.96 1.89 1.89 1.93 1.98 1.65 1.81 1.74 1.79 1.77
1.86 1.79 1.99 1.85 2.04 1.75 1.80 1.83 1.65 1.87
1.69 1.96 2.09 1.78 1.95 1.96 1.94 1.90 1.83 1.88

(i).Construct a frequency Distribution table using the class intervals 1.59-1.69, 1.70-1.80, 1.81-1.91..., also indicate the relative, cumulative frequencies as well as proportion columns on it.

Summary

- **Categorical Frequency Distribution** is a frequency distribution in which the data is only nominal or ordinal.
- **Ungrouped Frequency Distribution** is a frequency distribution of numerical data. The raw data is not grouped.
- **Grouped Frequency Distribution** is a frequency distribution where several numbers are grouped into one class.

Self-Assessment Questions

SAQ

1. What is categorical frequency distribution?
2. What is ungrouped frequency distribution?
3. What is grouped frequency distribution?
4. What are the graphical representation of data for ungrouped and categorical data?
5. What are the graphical representation of data for grouped data?

Tutor Marked Assessment

- **Question 1:** Assumming you went out with a questionnaire to collect information on height, weight and date of birth from all your class mates. The data you obtained this way can be called

  (a)universal data

  (b)secondary data

  (c)primary data

  (d)categorical data

  (e)empirical data

- **Question 2:** Define a sample

  (a)a quarter of the class

  (b)a cup of water from the ocean

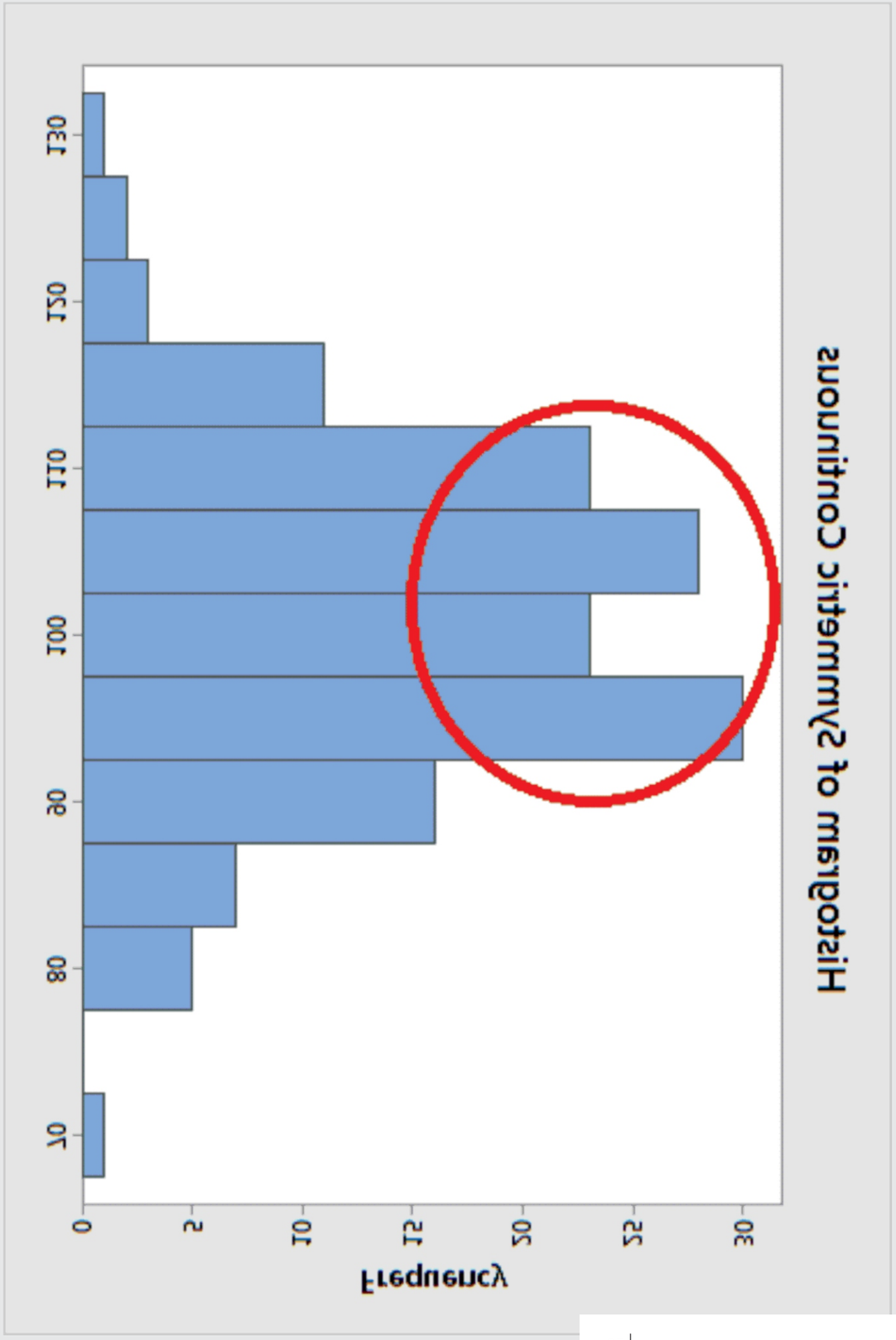
  (c)all list of all the students in STA131 in 2008/2009

  (d)a portion of the population of interest

  the entire list of members of the population of interest

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Lioyd R. Jaisingh (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.



05 | **Picture:** Measure of central tendency

Photo: wikipedia.com

Module 3

Measures of Central Tendency

06 | **Picture:** Measure of location

Photo: Pixels.com

UNIT 1

Measures of Location

Introduction

In this unit I will explain the best way to reduce a set of data and still retain part of the information is to summarize the set with a single value. But how can you calculate a number that is representative of an entire list of numbers. To achieve this, we need to study the following: Mean, Median, Mode, Geometric mean, and Harmonic mean under both grouped and ungrouped data situation.

Learning Outcomes

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

1. Measures of central tendency
2. How to calculate mean
3. How to calculate Median
4. How to calculate Mode
5. How to calculate Geometric mean
6. How to calculate Harmonic mean.

Main Content

Measures of Central Tendency

4 mins

SAQ 1&2

Average could mean one of four things. The arithmetic mean, the median, midrange, mode, Geometric mean or Harmonic mean. For this reason, it is better to specify which average you're talking about.

Mean

SAQ 1&2&3

This is what people usually intend when they say "average"

Population Mean: $\mu = \frac{\sum x}{N}$

Sample Mean: $\bar{x} = \frac{\sum x}{n}$

Frequency Distribution: $\bar{x} = \frac{\sum xf}{\sum f}$

The mean of a frequency distribution is also the weighted mean.

You can use the mean to compute other statistics (such as the variance) but you cannot use it to compute Open ended grouped frequency distributions. It is often not appropriate for skewed distributions such as salary information.

Median

You need know that data must be ranked (sorted in ascending order) first. The median is the number in the middle.

For you to find the depth of your median, there are several formulas you could be used, the one that we will use is:
median = 0.5 * (n + 1)

Raw Data

The median is the number in the "depth of the median" position. If the sample size is even, the depth of the median will be a decimal -- you need to find the midpoint between the numbers on either side of the depth of the median.

Ungrouped Frequency Distribution

First thing you should do is to find the cumulative frequencies for the data. The first value with a cumulative frequency greater than depth of the median is your median. If the depth of the median is exactly 0.5 more than the cumulative frequency of the previous class, then the median is the midpoint between the two classes.

Grouped Frequency Distribution

This is the tough one.

Since the data is grouped, you have lost all original information. Some textbooks have you simply take the midpoint of the class. This is an over-simplification which isn't the true value (but much easier to do). The correct process is to interpolate.

Find out what proportion of the distance into the median class the median by dividing the sample size by 2, subtracting the cumulative frequency of the previous class, and then dividing all that by the frequency of the median class.

Multiply this proportion by the class width and add it to the lower boundary of the median class.

$$Median = L_1 + \left(\frac{\sum f/2 - fcm}{fm} \right) c$$

where $_1L$ is the lower bound of the Median class, fcm is the cumulative frequency before the median class, fm is the frequency of the median class, c is the class size or class length.

The Median is the center number and is good for skewed distributions because it is resistant to change.

Mode

SAQ 1&3

The mode is the most frequent data value. There may be no mode if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).

For grouped frequency distributions, the modal class is the class with the largest frequency.

$$Mode = L_1 + \left(\frac{d_1}{d_1 + d_2}\right)c$$

where $_1L$ is the lower bound of the Modal class, $_1d$ is the excessive frequency below the modal class. $_2d$ is the excessive frequency above the modal class, c is the class size or class length.

The Mode is used to describe the most typical case. You can use mode with nominal data whereas the others can't. The mode may or may not exist and there may be more than one value for the mode.

Midrange

The midrange is simply the midpoint between the highest and lowest values.

(Max + Min)/2

The Midrange is not used very often. It is a very rough estimate of the average and is greatly affected by extreme values (even more so than the mean).

Property	Mean	Median	Mode	Midrange
Always Exists	No	Yes	No	Yes
Uses all data values	Yes	No	No	No
Affected by extreme values	Yes	No	No	Yes

Geomentric Mean

SAQ 4&6

The Geometric Mean (GM) for ungrouped data is defined as

$$GM = \sqrt[n]{x_1 * x_2 * x_3 * \ldots * x_n}$$

But for grouped data, it is defined as

$$GM = \sqrt[n]{x_1^{f_1} * x_2^{f_2} * x_3^{f_3} * \ldots * x_n^{f_n}}$$

Harmonic Mean

SAQ 4&6

The Harmonic Mean (HM) for n positive observation is defined as

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

For grouped data Harmonic Mean is

$$HM = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \cdots + \frac{f_n}{x_n}} = \frac{n}{\sum \frac{f_i}{x_i}}$$

Summary

- Average could mean one of four things. The arithmetic mean, the median, midrange, mode, Geometric mean or Harmonic mean.
- The mode is the most frequent data value.
- The median is the middle number.
- The midrange is simply the midpoint between the highest and lowest values.

Tutor Marked Assessment

- **Question 1:** Eighty Randomly selected light bulbs were tested to determine their lifetimes in hours. This frequency distribution was obtained.

Class boundaries	Frequency	Mid-Class	f.x	$f.x^2$
52.5 – 63.5	6	58.0	348	20184
63.5 – 74.5	12	69.0	828	57132
74.5 – 85.5	25	80.0	2000	160000
85.5 – 96.5	18	91.0	1638	149058
96.5 – 107.5	14	102.0	1428	145656
107.5 – 118.5	5	113.0	575	64975
$\sum_{i=1}^{6}$	80	513	6817	597005

You should calculate average lifetime of the light bulbs using the mean, median and the mode.

- **Question 2:** several packs of coco beans were planted to examine the rate of germination of the beans per pack after ten days of planting. The frequency distribution of the number of germinations per pack from 100 randomly selected packs after 10 days is as follows

Germination (x)	0	1	2	3	4	5	6	7	8	9	10	Total
Frequency (f)	11	13	9	11	9	7	12	9	10	7	2	100
f.x	0	13	18	33	36	35	72	63	80	63	20	433
$f.x^2$	0	13	36	99	144	175	432	441	640	567	200	2747

(1) Estimate the average number of germination per pack.

(2) Estimate the average number of beans that failed to germinate per pack, if each pack containing 10 beans

Question 3: The frequency distribution of the discrete random variable X is given in Table herein as follows

X	Frequency F	Xf	$X^2.f$	$X^3.f$
0	1	0	0	0
1	7	7	7	7
2	15	30	60	120
3	10	30	90	270
4	6	24	96	384
5	1	5	25	125
$\sum_{i=1}^{5}$	40	96	278	906

Calculate the mean, median, and the mode of X in that order

Question 4: Calculate the Geometric mean and harmonic mean of the following 8 observations:

67938542

Question 5: Given the random variables $X_1X_2,\ldots,,X_n$ where n=10, $\sum_{i=1}^{10} X = 50_i$, $\prod_{i=1}^{10} X_i = 3870720$ and $\sum_{i=1}^{10}\left(\frac{1}{X_i}\right) = 2.42619$. Calculate the mean, Geometric mean and the harmonic mean of X respectively.

References

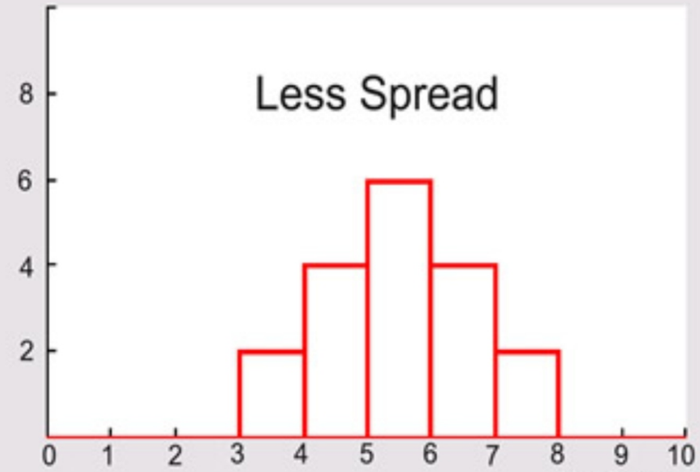
- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Lioyd R. Jaisingh (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- oss, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.

Module 4
Measures of Dispersion

MEASURE OF DISPERSION

07 | **Picture:** Measure of dispersion

Photo: Wikipedia.com



08 | **Picture:** Measure of spread

Photo: Wikipedia.com

UNIT 1

Measures of Spread

Introduction

In this unit I will explain the best way to reduce a set of data and still retain part of the information is to summarize the set with a single value. Data description is not complete until the spread, variability is also known. To achieve this, we examine the following: Range, Variance, Standard deviation and Coefficient of Variation.

Learning Outcomes

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

1. Measures of dispersion or spread
2. How to obtain range
3. How to calculate Variance
4. How to calculate standard deviation
5. How to calculate Coefficient of variation.

Main Content

Range

1 min

SAQ 1&2

The range is the simplest measure of variation to find. It is simply the highest value minus the lowest value.

RANGE = MAXIMUM - MINIMUM

I should also let you know that since the range only uses the largest and smallest values, it is greatly affected by extreme values, that is - it is not resistant to change.

Variance

1 min

SAQ 1,2&3

“Average Deviation"

The range only involves the smallest and largest numbers, and it would be desirable to have a statistic which involved all of the data values.

The first attempt you might make at this is something they might call the average deviation from the mean and define it as:

$$Ave.\ Dev = \frac{\sum (x - \mu)}{N}$$

I should inform you further that the problem with this is that summation is always zero. So, the average deviation will always be zero. That will make you to know already why the average deviation is never used.

Population Variance

So, to keep it from being zero, we squared deviation from the mean and called it "squared deviation from the mean". This "average squared deviation from the mean" we call it the variance.

For Ungrouped data

$$Population\ Variance = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Unbiased Estimate of the Population Variance

It might be a bit confusing that you would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$Sample\ Variance = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Standard Deviation

2 mins

SAQ 1,2&3

I should let you know that there is a problem with variances. You should recall that the deviations were squared and what that means is the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$Population\ Standard\ Deviation = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$Sample\ Standard\ Deviation = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Also,its very important you are aware that the sample standard deviation is not the unbiased estimator for the population standard deviation.

Let me tell you the interesting part here a calculator does not have a variance key on it. It does have a standard deviation key. You will have to square the standard deviation to find the variance.

Sum of Squares (shortcuts)

The sum of the squares of the deviations from the means is given a shortcut notation and several alternative formulas.

$$SS(x) = \sum (x - \bar{x})^2$$

A little algebraic simplification returns:

$$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n}$$

You maybe confused that what was wrong with the first formula? Okay! Let's solve the puzzle considering the following example - the last row are the totals for the columns

1. Total the data values: 23
2. Divide by the number of values to get the mean: 23/5 = 4.6
3. Subtract the mean from each value to get the numbers in the second column.
4. Square each number in the second column to get the values in the third column.
5. Total the numbers in the third column: 5.2
6. Divide this total by one less than the sample size to get the variance: 5.2 / 4 = 1.3

X	$x - \bar{x}$	$(x - \bar{x})^2$
4	4 - 4.6 = -0.6	$(-0.6)^2 = 0.36$
5	5 - 4.6 = 0.4	$(0.4)^2 = 0.16$
3	3 - 4.6 = -1.6	$(-1.6)^2 = 2.56$
6	6 - 4.6 = 1.4	$(1.4)^2 = 1.96$
5	5 - 4.6 = 0.4	$(0.4)^2 = 0.16$
23	0.00 (Always)	5.2

Not too bad, you think. But this can get pretty bad if the sample mean doesn't happen to be a "nice" rational number. Think about having a mean of 19/7 = 2.714285714285... Those subtractions get nasty, and when you square them, they're really bad. Another problem with the first formula is that it requires you to know the mean ahead of time. For a calculator, this would mean that you have to save all of the numbers that were entered.

Now, let me introduce you to the shortcut formula (In statitics they say there's always a shortcut for everything...) The only things that you will require to find are the sum of the values and the sum of the values squared. Thus,in this formula there is no subtraction and no decimals or fractions until the end. The last row contains the sums of the columns, just like before.

(1) Record each number in the first column and the square of each number in the second column.

(2) Total the first column: 23

(3) Total the second column: 111

(4) Compute the sum of squares: 111 - 23*23/5 = 111 - 105.8 = 5.2

(5) Divide the sum of squares by one less than the sample size to get the variance = 5.2 / 4 = 1.3

X	x^2
4	16
5	25
3	9
6	36
5	25
23	111

Grouped Data

Population Variance= $Var(x) = \sigma^2 = \dfrac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n}$

Sample Variance= $Var(x) = s^2 = \dfrac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n-1}$

Population Standard deviation= $SD(x) = \sigma = \sqrt{\dfrac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n}}$

Sample Standard deviation= $SD(x) = s = \sqrt{\dfrac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n-1}}$

The short cut

Population Variance= $Var(x) = \sigma^2 = \dfrac{\sum_{i=1}^{n} fx_i^2 - n\bar{x}^2}{n} = \dfrac{\sum_{i=1}^{n} fx_i^2 - \dfrac{\left(\sum_{i=1}^{n} fx\right)^2}{n}}{n}$

Sample Variance= $Var(x) = s^2 = \dfrac{\sum_{i=1}^{n} fx_i^2 - n\bar{x}^2}{n-1} = \dfrac{\sum_{i=1}^{n} fx_i^2 - \dfrac{\left(\sum_{i=1}^{n} fx\right)^2}{n}}{n-1}$

Coefficient of Variation(CV) : You can used this for comparison. It's as well used to determine the most consistent

sets of data $CV = \dfrac{SD}{\bar{X}} \times 100$

Where SD is the standard deviation.

Summary

- Data description is not complete until the spread, variability is also known.
- The range is the simplest measure of variation to find. It is simply the highest value minus the lowest value.
- We squared deviation from the mean and called it "squared deviation from the mean". This "average squared deviation from the mean" we call it the variance.
- The square root of the variance is called the Standard deviation.

Self-Assessment Questions

SAQ

1. What is a range?
2. Define variance?
3. Define standard deviation?
4. Define Coefficient of Variation?

Tutor Marked Assessment

- **Question 1:** Eighty Randomly selected light bulbs were tested to determine their lifetimes in hours. This frequency distribution was obtained.

Class boundaries	Frequency	Mid-Class	f.x	$f.x^2$
52.5 – 63.5	6	58.0	348	20184
63.5 – 74.5	12	69.0	828	57132
74.5 – 85.5	25	80.0	2000	160000
85.5 – 96.5	18	91.0	1638	149058
96.5 – 107.5	14	102.0	1428	145656
107.5 – 118.5	5	113.0	575	64975
$\sum_{i=1}^{6}$	80	513	6817	597005

You should calculate the Variance, standard deviation and Coefficient of variation of the lifetime of the bulbs

Question 2: Given that $X_1, X_2, \ldots, X_n$ is a random sample of size n such that n=15,

$$\sum_{i=1}^{15} X_i = 120, \sum_{i=1}^{15} X_i^2 = 1240, \sum_{i=1}^{15} (X_i - \bar{X})^2 = 280, \sum_{i=1}^{15} (X_i - \bar{X})^3 = 0, and \sum_{i=1}^{15} (X_i - \bar{X})^4 = 9352$$

Calculate the Mean, Variance, standard deviation and Coefficient of variation..

Question 3: The frequency distribution of the discrete random variable X is given in Table herein as follows

X	Frequency f	fX	fX^2
0	1	0	0
1	7	7	7
2	15	30	60
3	10	30	90
4	6	24	96
5	1	5	25
$\sum_{i=1}^{5}$	40	96	278

Find the Mean, variance, Standard deviation and Coefficient of Variation of X

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Lioyd R. Jaisingh (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- oss, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.

Module 5

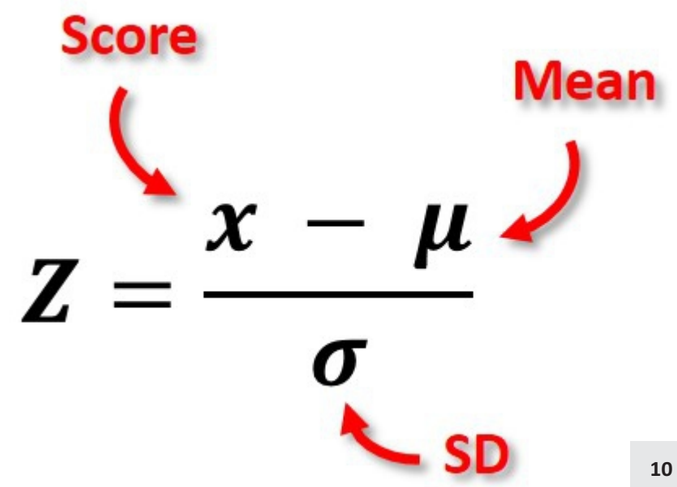
Measures of Position

MEASURES OF POSITION

FOURTH QUARTER- STATISTICS

09 | **Picture:** Measure of position

Photo: wikipedia.com



10 | **Picture:** z-score

Photo: Wikipedia.com

UNIT 1

Measures of Position

Introduction

In this unit I will explain the best way to reduce a set of data and still retain part of the information is to summarize the set with a single value. We shall look at measures of Position: Z-score, percentiles, decentiles and quarters.

Learning Outcomes

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

1. Measures of position
2. How to obtain the Z-score of a certain value
3. How to calculate the percentiles
4. How to calculate decentiles
5. How to calculate quartiles

Main Content

Standard Scores (z-scores)

1 min

SAQ 1&2

You should know that your standard score can be obtained by subtracting the mean and dividing the difference by the standard deviation. The symbol is z, which is why we call it a z-score.

$$z = \frac{x - \mu}{\sigma} \quad or \quad z = \frac{x - \bar{x}}{s}$$

Let me introduce you to the following, the mean of the standard scores is zero and the standard deviation is 1 and you should know the delightful feature of the standard score is that no matter what the original scale was, when the data is converted to its standard score, the mean is zero and the standard deviation is 1.

Percentiles, Deciles, Quartiles

1 min

SAQ 1&2

Percentiles (100 regions)

The kth percentile is the number which has k% of the values below it. The data must be ranked.

- Rank the data you're given.
- Then you should proceed to find k% (k /100) of the sample size, n.
- If this is an integer, add 0.5. If it isn't an integer round up.
- Find the number in this position. If your depth ends in 0.5, then take the midpoint between the two numbers.

Okay..it's neccessary to point out to you that sometimes it's easier to count from the high end rather than counting from the low end. Take this as an example,the 80th percentile is the number which has 80% below it and 20% above it. Rather than stressing yourself with counting 80% from the bottom, it's earlier for you to count 20% from the top.

Note: The 50th percentile is the median.

If you wish to find the percentile for a number (rather than locating the kth percentile), then let me take you through the stressful

- Take the number of values below the number
- Then add 0.5 to the number you selected in (1)
- Divide by the total number of values
- Convert it to a percent

Deciles (10 regions)

The percentiles divide the data into 100 equal regions. The deciles divide the data into 10 equal regions. You can use this instructions to find a percentile, except instead of dividing by 100 in step 2, divide by 10.

Quartiles (4 regions)

The quartiles divide the data into 4 equal regions. Instead of dividing by 100 in step 2, divide by 4.

You should note this and comprehend as well; The 2nd quartile is the same as the median. The 1st quartile is the 25th percentile, the 3rd quartile is the 75th percentile.

In most cases we often use the quartiles (much more so than the percentiles or deciles).

Interquartile Range (IQR)

The interquartile range is the difference between the third and first quartiles. That's it: Q3 - Q1

Semi Interquartile Range (SIQR)

The Semi interquartile range is the difference between the third and first quartiles divided by 2, that is (Q3 - Q1)/2.

Summary

- Standard score can be obtained by subtracting the mean from the value and
- dividing the difference by the standard deviation.
- The kth percentile is the number which has k% of the values below it.
- The percentiles divide the data into 100 equal regions.
- The deciles divide the data into 10 equal regions.
- The quartiles divide the data into 4 equal regions.

Self-Assessment Questions

1. What is standard score?
2. What is percentile?
3. What is Decile?
4. What is Quartile?
5. How do we obtain each of the above?

Tutor Marked Assessment

- **Q1. Given the following table:**

Class Interval	Frequency
0 – 4	4
5 – 9	10
10 – 14	14
15 – 19	22
20 – 24	29
25 – 29	4
30 – 34	3
35 – 39	2

Obtain the Mean, Median, Mode, Variance, Standard deviation, Q_1, Q_3, P_{24}, P_{67}, D_4, D_8, D_2 , IQR, SIQR

Q2. Given the following data on the shoe sizes of some students:

4, 8, 9, 7, 6, 7, 8, 5, 5, 7.

Obtain the Mean, Median, Mode, Variance, Standard deviation, Q_1, Q_3, P_{24}, P_{67}, D_4, D_8, D_2 , IQR and SIQR.

Q3. True or False Questions

(1) The mean of a set of data always divides the data set such that 50 percent of the values lie above the mean and 50 percent lie below the mean.

(2) The mode is a measure of variability.

(3) The median of a set of data values is that value that occurs the most.

(4) The mean is not equal to the median in a symmetrical distribution.

(5) Of the mean, the median, and the mode of a data set, the mean is most influenced by an outlying value in the data set.

(6) If the number of observations in a data set is odd, the median cannot be accurately found, but rather is approximated.

(7) A data set with more than one mode is said to be bimodal.

(8) The sum of the deviations from the mean for any data set is always 0.

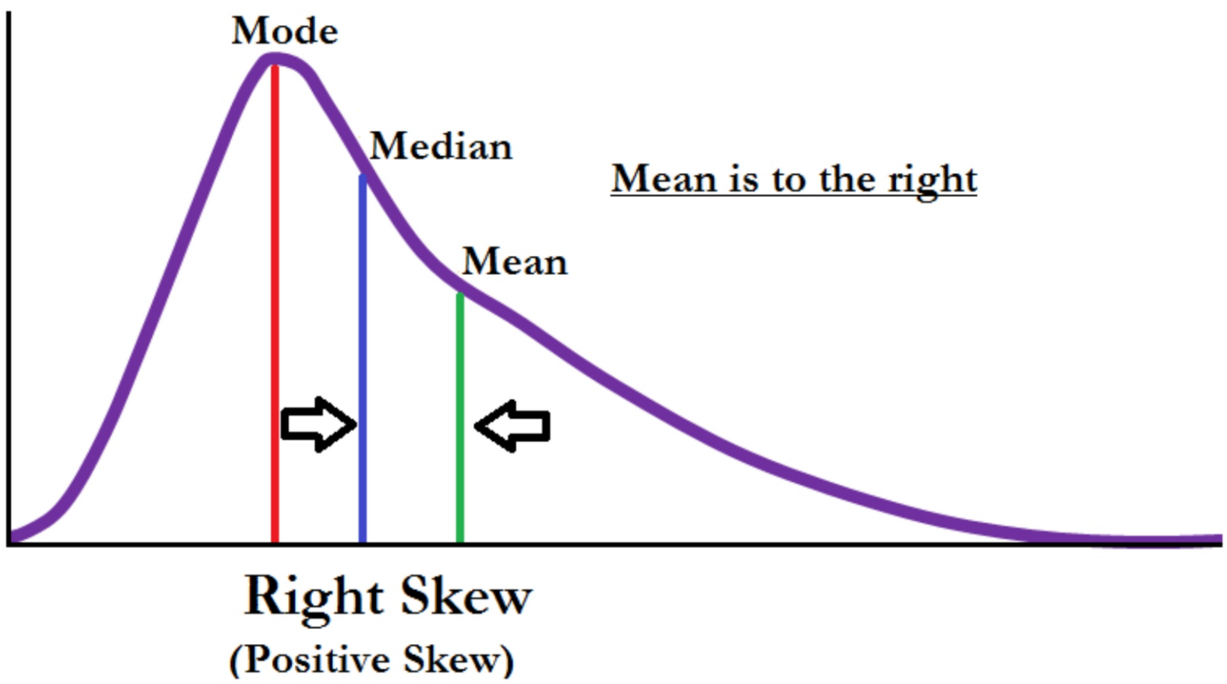
(9) For a negatively skewed distribution, the tail is to the right of the mean.

(10) For a positively skewed distribution, the mode is less than the median, and the median is less than the mean.

References

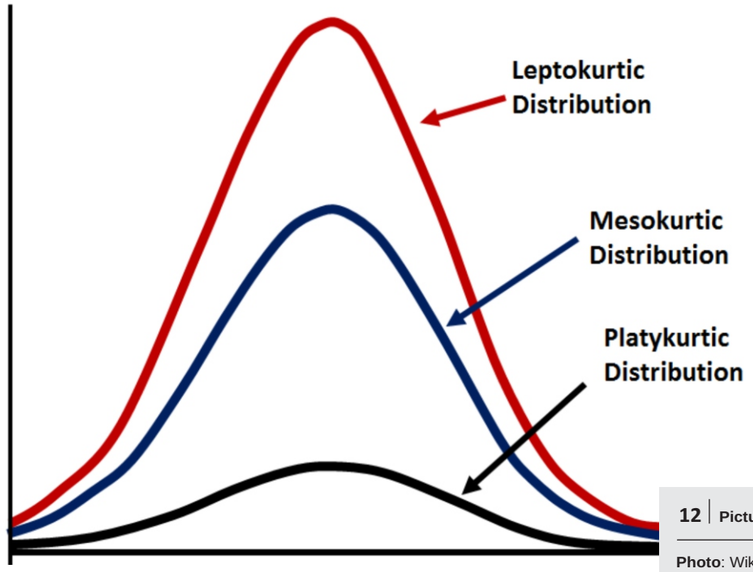
- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Lioyd R. Jaisingh (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- oss, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.

Module 6
Measures of Shape



11 | **Picture:** Measure of shapes

Photo: Wikipedia.com



12 | **Picture:** Measure of shape

Photo: Wikipedia.com

UNIT 1

Measures of Shape

Introduction

In this unit I will explain the best way to reduce a set of data and still retain part of the information is to summarize the set with a single value. We shall look at measures of shapes: Bell shape -Symmetric Distribution, Skewness, and Kurtosis.

Learning Outcomes

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

1. Measures of shape
2. How to obtain the skewness
3. How to calculate the kurtosis

Main Content

Bell Shaped

1 min

SAQ 1, 2,3,4, 5,6,7,&8

You should be inform that the bell shaped distribution is to be symmetric. It is the most assumed distribution in statistical analysis. Note its peculiar nature:

Mode=Median=Mean

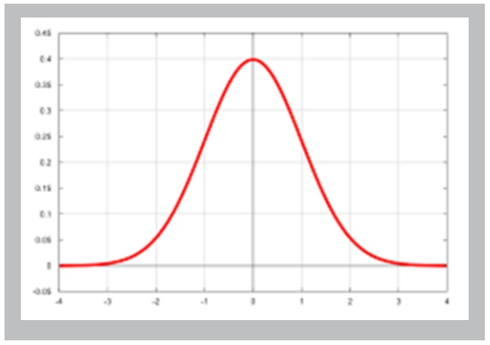


Figure 6.1: Symmetric Distribution

Skewness: is dimensionless

1 min

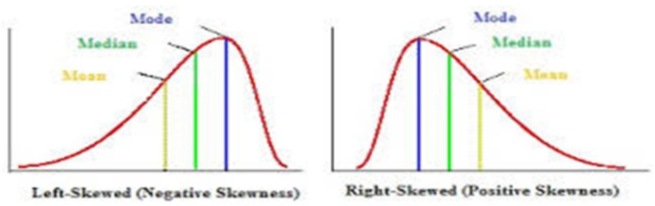


Figure 6.2

Pearsonian Coefficient of Skewness

$$PCS = \frac{3(meam - median)}{S\,\text{tan}\,dard \quad deviation}$$

PCS=0 iff mean=median

PCS>0 iff mean > median (right skewed or Positive Skewed)

PCS <0 iff mean < median (left skewed or Negative Skewed)

General Measure of Skewness (GMS):

Consider the rth central moment

$$m_r = \frac{1}{n}\sum_{i-1}^{n}(x_i - \bar{x})^r$$

$$m_2 = \frac{1}{n}\sum_{i-1}^{n}(x_i - \bar{x})^2 = S^2$$

$$m_3 = \frac{1}{n}\sum_{i-1}^{n}(x_i - \bar{x})^3$$

$$GMS = \frac{m_3}{(m_2)^{\frac{3}{2}}} = \frac{m_3}{(\sqrt{S^2})^3} = \frac{m_3}{(S^2)^{1.5}}$$

GMS is dimensionless

GMS > 0 means distribution is right skewed or Positive skewed

GMS < 0 means distribution is left skewed or Negative skewed

All depend on m3

Application

Compute the PCS and GMS for the following observations:
6,7,8,8,8,9,10,10,11 and 13.

	x	f	CM	fx	f(x-9)²	(f(x-9)³	f(x-9)⁴
	6	1	1	6	9	-27	81
	7	1	2	7	4	-8	16
	8	3	5	24	3	-3	3
	9	1	6	9	0	0	0
	10	2	8	20	2	2	2
	11	1	9	11	4	8	16
	13	1	10	13	16	64	256
Sum	64	10		90	38	36	374

Mean = 90/10=9

Median = 8.5

$$SD = \sqrt{\frac{38}{10-1}} = \sqrt{4.222} = 2.055$$

$$PCS = \frac{3(9-8.5)}{2.055} = 0.7299$$

GMS

$$GMS = \frac{m_3}{(S^2)^{1.5}}$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2 = \frac{38}{10} = 3.8$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{36}{10} = 3.6$$

$$\therefore GMS = \frac{3.6}{(3.8)^{1.5}} = 0.4860$$

SAQ 9

Measure of Kurtosis

This is departure of distribution from normality. Kurtosis is defined as

$$Kurtosis = \frac{m_4}{(m_2)^2} = \frac{m_4}{(S^2)^2}$$

Where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Application

	x	f	CM	fx	f(x-9)²	(f(x-9)³	f(x-9)⁴
	6	1	1	6	9	-27	81
	7	1	2	7	4	-8	16
	8	3	5	24	3	-3	3
	9	1	6	9	0	0	0
	10	2	8	20	2	2	2
	11	1	9	11	4	8	16
	13	1	10	13	16	64	256
Sum	64	10		90	38	36	374

$$Kurtosis = \frac{m_4}{(m_2)^2} = \frac{m_4}{(S^2)^2}$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 = \frac{374}{10} = 37.4$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2 = \frac{38}{10} = 3.8$$

$$\therefore Kurtosis = \frac{m_4}{(m_2)^2} = \frac{m_4}{(S^2)^2} = \frac{37.4}{(3.8)^2} = \frac{37.4}{14.44} = 2.59$$

Kurtosis is dimensionless.

If kurtosis < 3 means small value and it indicates a flatter than normal (platykurtic) distribution.

If kurtosis = 3 indicates a normal (Mesokurtic) distribution.

Also beware that if kurtosis > 3 indicates a narrow than normal (leptokurtic) distribution.

Summary

- You should be inform that the bell shaped distribution is to be symmetric in nature.
- Skewness is dimensionless.
- PCS>0 iff mean > median (right skewed or Positive Skewed)
- PCS <0 iff mean < median (left skewed or Negative Skewed)
- Kurtosis is departure of distribution from normality.
- If kurtosis < 3 means small value and it indicates a flatter than normal (platykurtic) distribution.
- If kurtosis = 3 indicates a normal (Mesokurtic) distribution.
- Also beware that if kurtosis > 3 indicates a narrow than normal (leptokurtic) distribution.

Self-Assessment Questions

SAQ

1. What is Skewness
2. What is Kurtosis
3. How do we measure skewness and kurtosis

Tutor Marked Assessment

- If the number of measurements in a data set is odd, the median is the value when the data set is ordered from the smallest value to the largest value.
- If the number of measurements in a data set is even, the median is the of the two values when the data set is ordered from the smallest value to the largest value.
- The (mean, median, mode) for a set of data is the value in the data set that occurs most frequently.
- Two measures of central tendency.
- For a symmetrical distribution, the mean, mode, and median are all (equal to, different from) one another.
- For a negatively skewed distribution, the mean is (smaller, greater) than the median and the mode.
- For a positively skewed distribution, the tail of the distribution is to the (right, left) of the distribution.
- For a positively skewed distribution, the median is (smaller, larger) than mean.
- Given that X1, X2, . . ., Xn is a random sample of size n such that n=15,

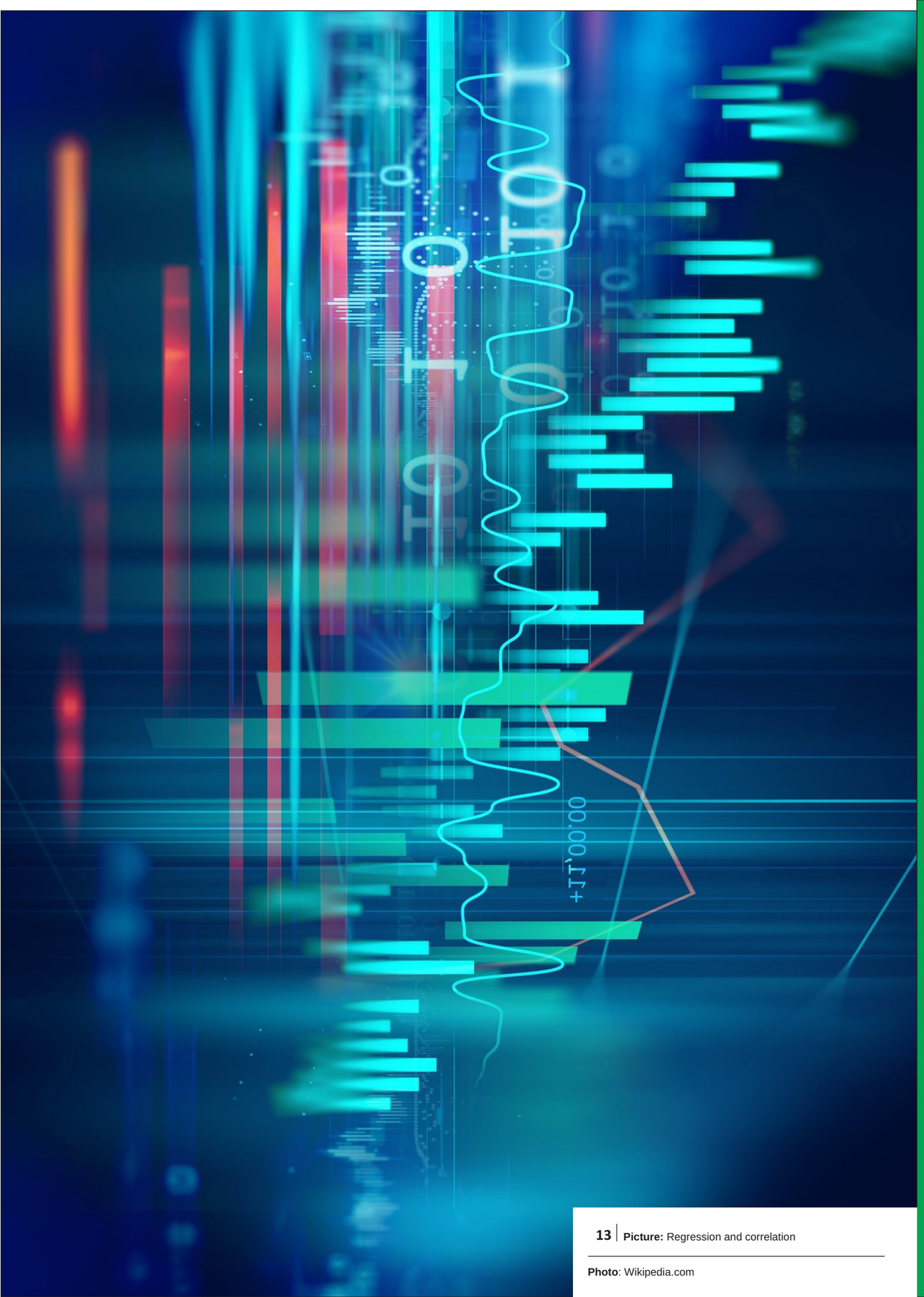
$$\sum_{i=1}^{15} X_i = 120, \sum_{i=1}^{15} X_i^2 = 1240, \sum_{i=1}^{15} (X_i - \bar{X})^2 = 280, \sum_{i=1}^{15} (X_i - \bar{X})^3 = 0, and \sum_{i=1}^{15} (X_i - \bar{X})^4 = 9352$$

Calculate the coefficients of kurtosis and Skewness.

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Lioyd R. Jaisingh (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- oss, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.

Unit 1 | Module 6 | STA 131
Measures of Shape



13 | **Picture:** Regression and correlation

Photo: Wikipedia.com

Module 7

Regression and Correlation

14 | **Picture:** Regression and correlation

Photo: Wikipedia.com

UNIT 1

Regression and Correlation

Introduction

In this unit I will explain the best way to study a pair of variables is to examine the linear relationship We shall look at simple linear regression and types correlation that exists.

Learning Outcomes

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

1. Simple linear regression
2. How to fit linear regression model
3. How to calculate correlation

Main Content

Simple Linear Regression Model

2 mins

SAQ 1&2

A simple linear regression model is one that has one dependent variable (Y) and one independent variable (X). It is defined as

$$Y_i = \alpha + \beta X_i + \sum_i$$

Where α is our intercept,

β is our slope

$\sum$ is the error term, which is normally distributed as mean zero and variance σ^2, that is $\sum \sim N(0,\sigma^2)$.

Using Least Square Estimation method, by minimizing the sum of error squares $\sum E^2 = S = \sum (Y_i - (\alpha + \beta X_i))^2$, the estimates of

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

and

$$\hat{\alpha} = \frac{\sum y - \hat{\beta}\sum X}{n} = \overline{Y} - \hat{\beta}\overline{X}$$

The fitted linear regression model is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

Correlation

2 mins

SAQ 2

You should relate that correlation is degree of linear relationship between two variables (X and Y). It takes values between 1 and -1 and It is denoted by r. We have different types of correlation.

(1) Positively correlated: When as one variable increases the other one also increases, then we say the two variables are positively correlated.

(2) Negatively correlated: When as one variable increases the other variable decreases., then the two variables are negatively correlated.

Pearson Correlation coefficient is defined as

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

$$-1 \le r \le 1$$

Rank correlation coefficient

$$r = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Where

$$d_i = \left(rank(X_i) - rank(Y_i)\right)$$

$$d_i = difference \ \ between \ \ \ their \ ranks$$

Applications

Consider the following data

X	1	2	3	4	5	6	7	8	9
Y	12	14	13	16	18	16	19	16	17

(i) Fit a Simple linear regression to the data?
(ii) Obtain the correlation coefficient using Pearson and Rank?

Solution

	x	y	xy	x^2	y^2
	1	12	12	1	144
	2	14	28	4	196
	3	13	39	9	169
	4	16	64	16	256
	5	18	90	25	324
	6	16	96	36	256
	7	19	133	49	361
	8	16	128	64	256
	9	17	153	81	289
Sum	45	141	743	285	2251

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{9(743) - (45)(141)}{9(285) - (45)^2} = 0.633$$

$$\hat{\alpha} = \frac{\sum y - \hat{\beta}\sum X}{n} = \frac{141 - (0.633)(45)}{9} = 12.5$$

So the required fitted linear regression model for the data is

$$\hat{Y}_i = 12.5 + 0.633 \ X_i$$

b) The correlation coefficient:

Pearson correlation

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

$$r = \frac{9(743) - (45)(141)}{\sqrt{\left(9(285) - (45)^2\right)\left(9(2251 - (141)^2\right)}} = 0.757$$

The Rank Correlation coefficient is

X_Rank	y	Y_Rank	d=Diff	d^2
1	12	1	0	0
2	14	3	-1	1
3	13	2	1	1
4	16	5	-1	1
5	18	8	-3	9
6	16	5	1	1
7	19	9	-2	4
8	16	5	3	9
9	17	7	2	4
			Sum	30

$$r = 1 - \frac{6(30)}{9(81 - 1)}$$

$$r = 1 - \frac{180}{720} = 1 - 0.25 = 0.75$$

Summary

- A simple linear regression model is one that has one dependent variable (Y) and one independent variable (X).
- Correlation is degree of linear relationship between two variables (X and Y). It takes values between 1 and -1 and It is denoted by r.
- Positively correlated: When as one variable increases the other one also increases, then we say the two variables are positively correlated.
- Negatively correlated: When as one variable increases the other variable decreases., then the two variables are negatively correlated.

Self-Assessment Questions

1. What is a simple linear regression?
2. How do we fit a simple linear regression model?
3. What is correlation?
4. How do we obtain the coefficient of correlation?

Tutor Marked Assessment

1. Given $\sum_{i=1}^{10} x_i = 96$, $\sum_{i=1}^{10} x_i^2 = 1060$, $\sum_{i=1}^{10} x_i y_i = 328$, $\sum_{i=1}^{10} y_i = 30$, $\sum_{i=1}^{10} y_i^2 = 108$, SSE = 6.439, n =10,

   Obtain the estimate of "**a**" for the fitted linear model $\hat{y}_i = \hat{a} + \hat{b}x_i$

   Obtain the Estimate of "**b**" for the fitted linear model $\hat{y}_i = \hat{a} + \hat{b}x_i$

   Obtain the Pearson correlation coefficient **r**

2. Given the random variables (X_i, Y_i), for i=1, 2, . . ., 10 such that

   $\sum_{i=1}^{10} X_i = 55, and \sum_{i=1}^{10} X_i^2 = 385$, $\sum_{i=1}^{10} Y_i = 110$, $\sum_{i=1}^{10} Y_i^2 = 1540, \sum_{i=1}^{10} X_i Y_i = 764$.

   Calculate the moment correlation coefficient between X and Y. And fif a linear model $y_i = a + bx_i$

References

- Oyejola, B. A. and Adebayo, S. B. (2004). Basic Statistics for Biology and Agriculture Students. OLAD Ilorin.
- Lioyd R. Jaisingh (2000). Statistics for the Utterly Confuse, McGraw-Hill, USA.
- oss, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientist. Elsevier Academic Press, USA
- Gupta, S. C. (2011). Fundamentals of Statistics. 6th Edition. Himalaya Publishing House. Delhi.