SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad

Module 7 : Descriptive Statistics

Lecture 17A : Introduction - I

CONCEPTS COVERED

1. Types of data – classify data as categorical or numerical data.

2. Visual Representation of numerical data and interpretation shape of the distribution

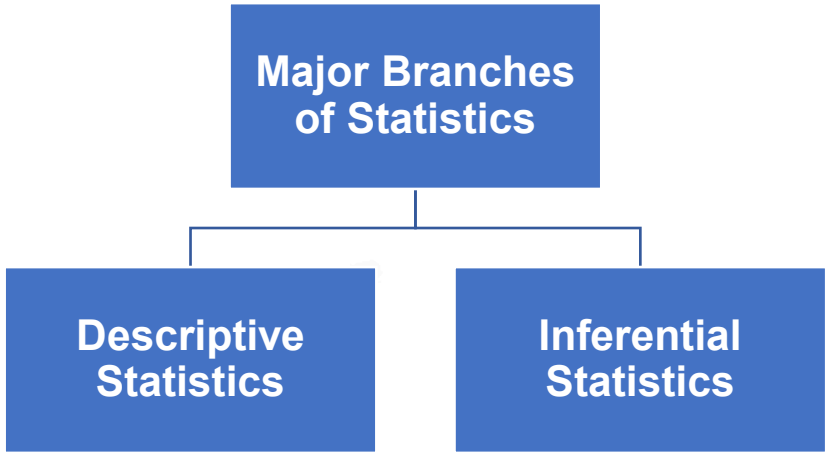
3. Compute and Interpret numerical Summaries of Data

- Compute and Interpret measures of central tendency: Mean, Median, Mode
- Compute and Interpret measures of dispersion: Range

What is Statistics?

Definition

Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.



Major Branches of Statistics

1. Description

***Descriptive statistics is the part of statistics concerned with the description and summarization of data*.**

2. Inference

The part of statistics concerned with drawing conclusions from data is called inferential statistics.

- **To draw a conclusion from the data, we must consider the possibility of chance – introduction to probability.**

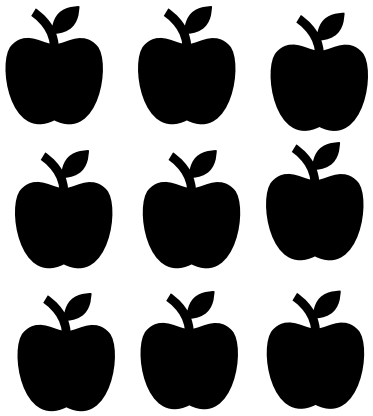
Population and Sample

Definition

- **The total collection of all the elements we are interested in is called a population.**

Definition

- **A subgroup of the population that will be studied in detail is called a sample.**



Population

Sample

Purpose of Statistical Analysis

- If the purpose of the analysis is to examine and explore information for its own intrinsic interest only, the study is descriptive.
- If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the analysis is inferential.
- A descriptive study may be performed on a sample or population.
- When an inference is made about the population based on the sample's information, then the study become inferential.

What is data ?

To learn something, we need information

Definition

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

Variables and Cases

- A variable that “varies,” and in formal definition, it is a characteristic or attribute that varies across all units.
- Case (Observation): A unit from which data are collected
- Rows represent cases: The same attribute is recorded for each case.
- Columns represent variables: The same type of value for each case is recorded for each variable.

Sample data:

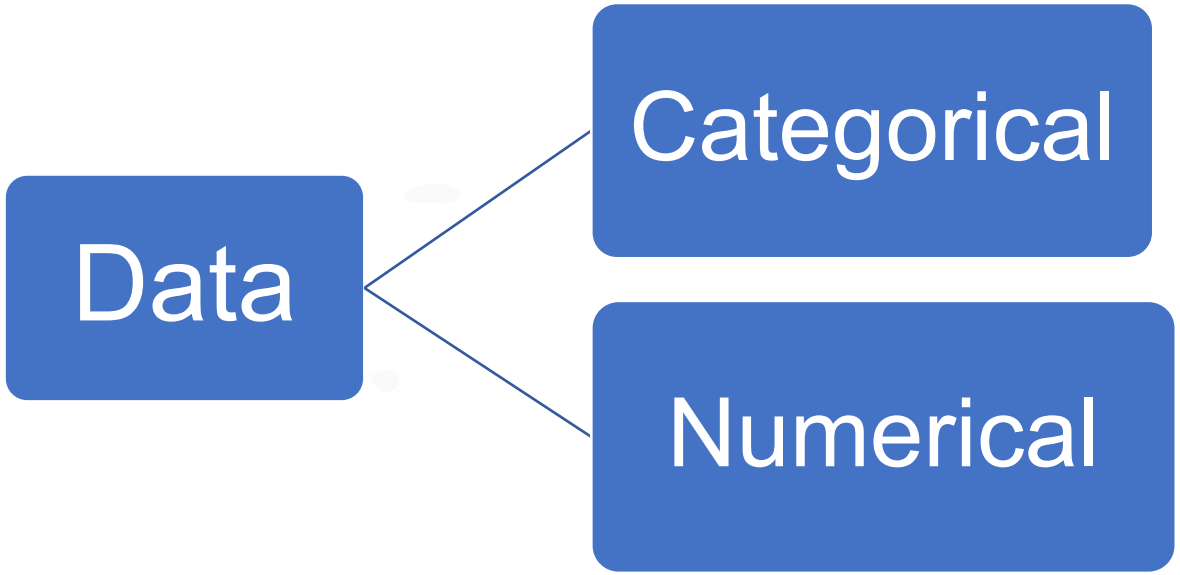
Variable

Sl No	Student Name	English Marks	Science Marks
1	Radha Krishna	93	85
2	Sai Raj	89	91

Case/Observation

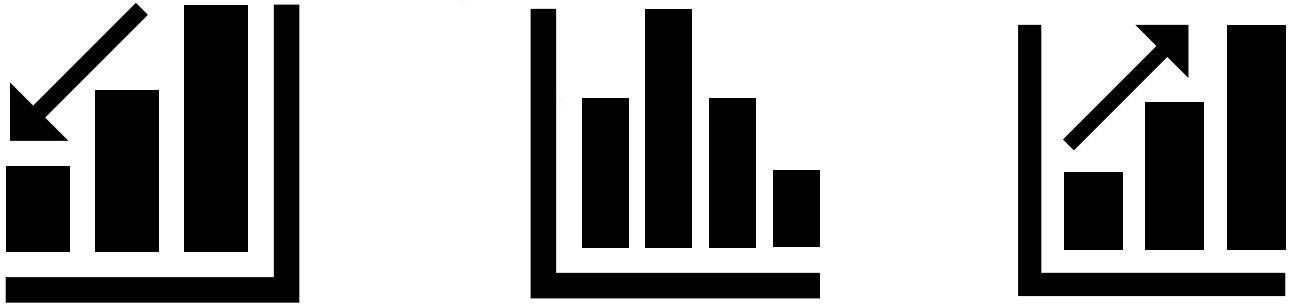
- Each variable must have its own column
- Each observation must have its own row

Classification of Data
Categorical and Numerical



Visual Representation

- Visual representation of numerical data is a crucial aspect of data analysis, and it provides insights into the distribution and characteristics of the data.
- Histogram is one of the standard methods for visualizing the numerical data.

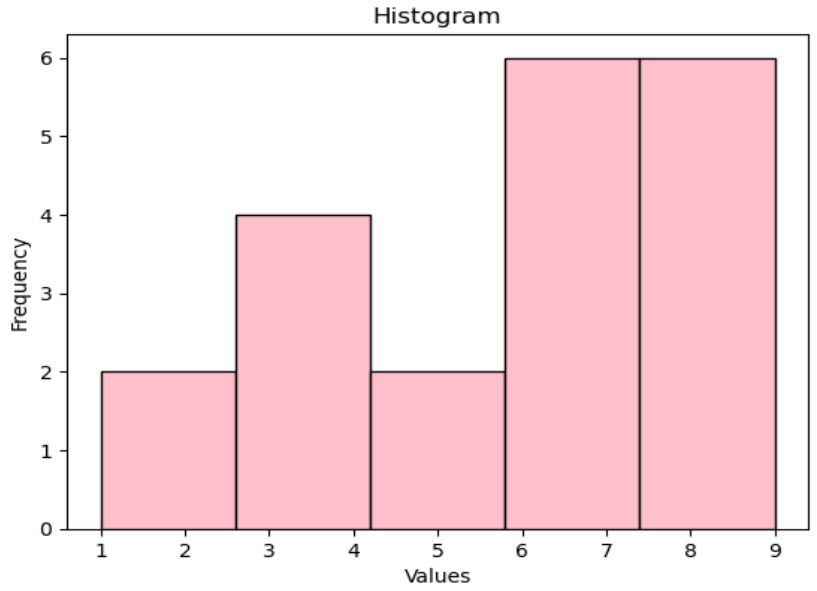


Histogram

A histogram is a graphical representation of the distribution of a dataset.

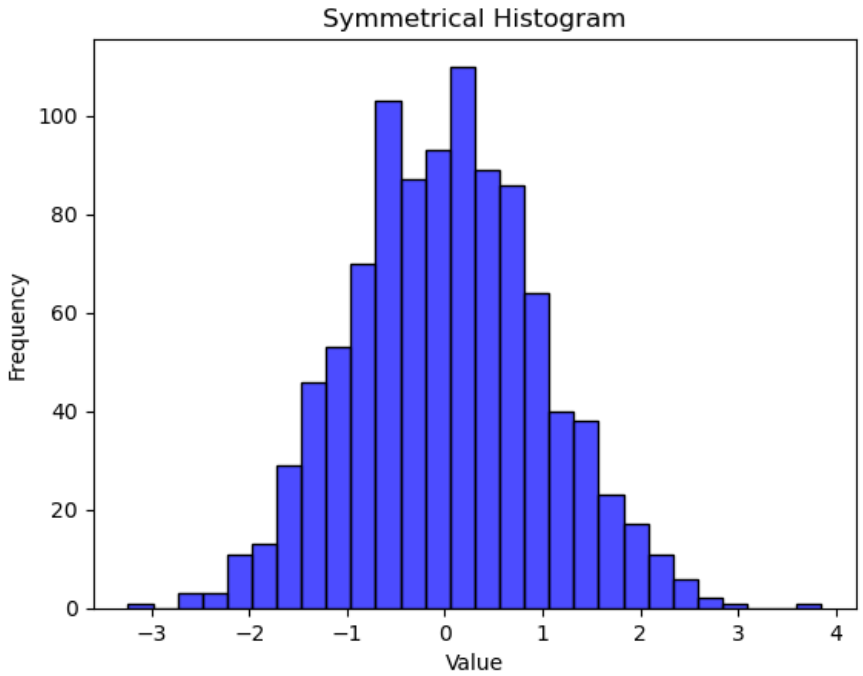
It consists of bars, where each bar represents a range of values (called a bin), and the height of the bar corresponds to the frequency of observations within that bin.

Data = [1, 2, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9]



Symmetric Distribution

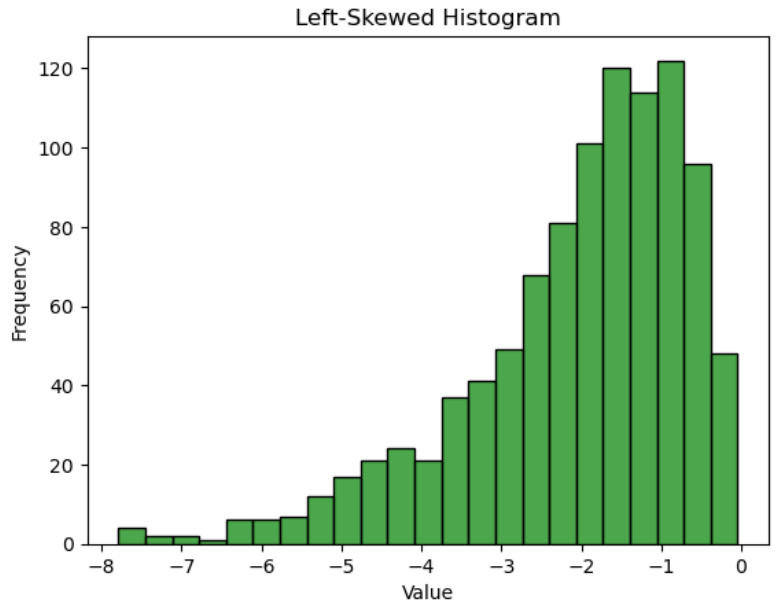
If the histogram is roughly symmetrical, it suggests that the data is evenly distributed around the mean.

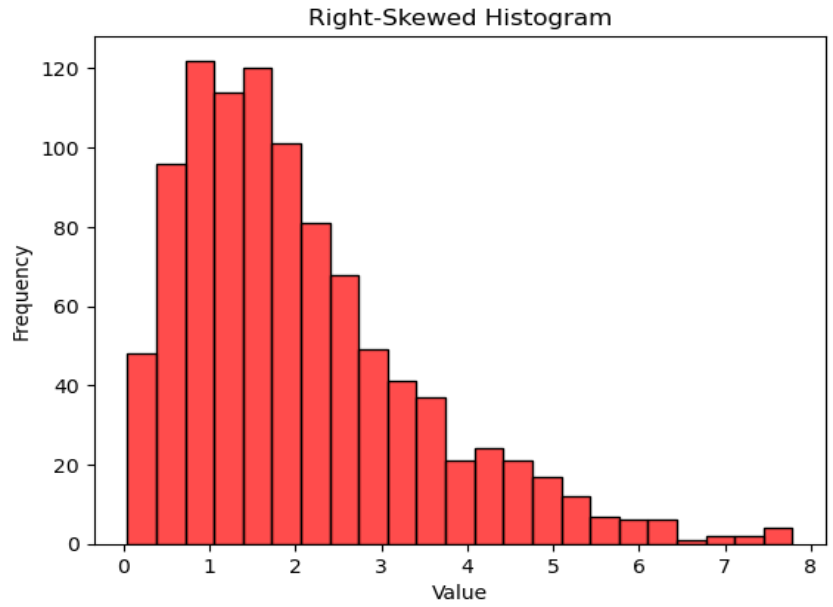


Skewed Distribution

If the histogram is skewed to the right (positively skewed), it indicates that the data has a tail on the right side.

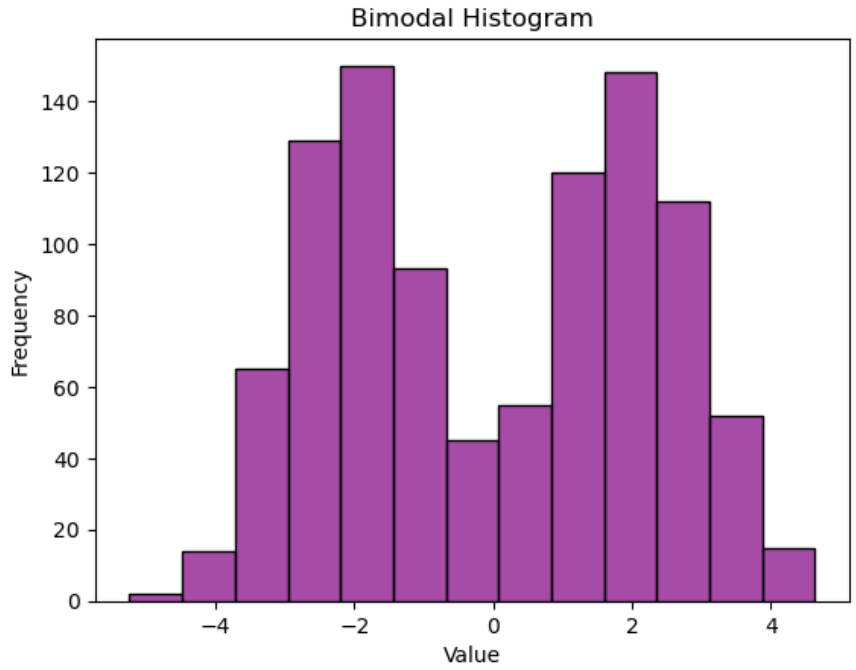
Similarly, if it's skewed to the left (negatively skewed), there's a tail on the left side.





Bimodal Distribution

If there are two distinct peaks, it suggests that the data may have two underlying subgroups.



Descriptive Measures

Most commonly used descriptive measures can be categorized as

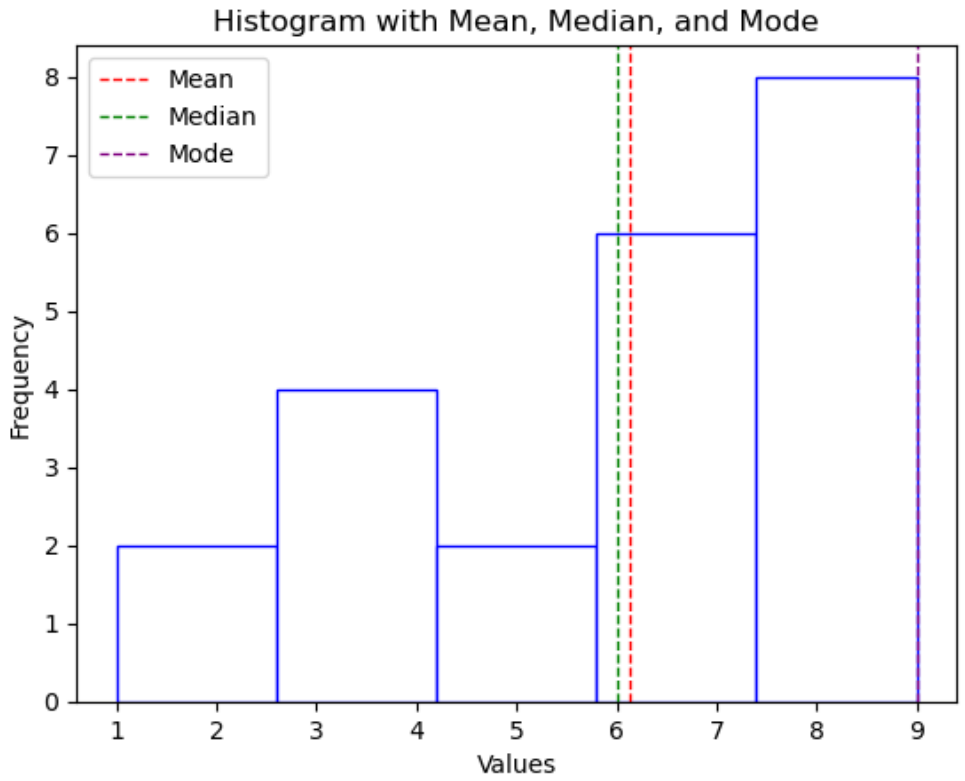
Measures of central tendency: These are measures that indicate the most typical value or center of a data set.

Measures of dispersion: These measures indicate the variability of a dataset.

Measures of Central Tendency

- Mean
- Median
- Mode

Data = [1, 2, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9, 9, 9]



Mean: 6.136
Median: 6.0
Mode: 9

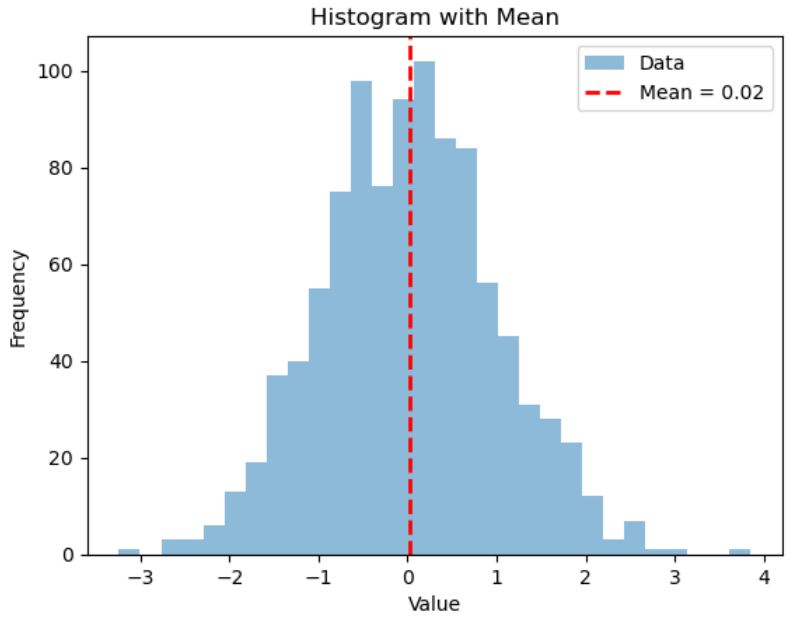
The Mean

The most commonly used measure of central tendency is the mean.

Definition

The mean of a data set is the sum of the observations divided by the number of observations

- **The mean is usually referred to as average.**
- **For discrete observations:**
- **Sample mean:** $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
- **Population mean:** $\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$



Example:

1. The winning scores in the U.S masters golf tournament in the years from 2004 to 2013 were as follows: 280,278,272,276,281,279,276,281,289,280. Find the sample mean of these scores.

Sample mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$

$= \frac{276 + 281 + 279 + 276 + 281 + 289 + 280 + 280 + 278 + 272}{10}$

$= 279.2$

Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- Let $y_i = x_i * c$ where c is a constant then $\bar{y} = \bar{x} * c$

Example : the marks of students : 88,74,86,67,90,49

The mean of above marks is 75.67.

Let us suppose that you have decided to add 5 marks as bonus marks to each student, then the data becomes : 93, 79, 91, 72, 95 , 54

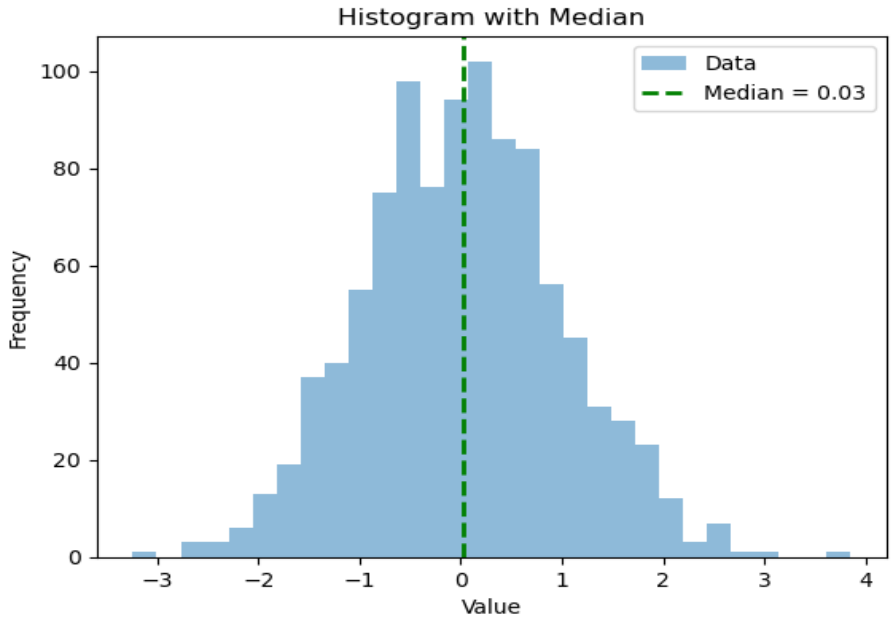
The mean of the new data set is 80.67 = 75.67 + 5

The Median

- **Another frequently used measure of center is the median. Essentially, the median of a data set is the number that divides the bottom 50 % of the data from the top 50%.**

Definition

The median of a data set is the middle value in its ordered list.



Steps to obtain median

1.Arrange the data in increasing order. Let n be the total number of observations in the dataset.

2. If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e. $\frac{n+1}{2}$ observation

3. If the number of observations is even, then the median is the mean of the two middle observations in the ordered list, i.e. mean of $\frac{n}{2}$ and $\frac{n}{2}+1$ observation.

Examples of median

Scores = [75, 82, 90, 65, 88, 72, 91, 78, 85, 79]

Median = 80.5

Numbers = [11, 22, 15, 29, 33, 15, 17, 22, 19, 25, 27]

Median = 22

Outliers Effect

Example 1: 1 2, 12, 5, 7, 6, 7, 3

Sample Mean = 6 , Sample Median = 6

Example 2: 2, 117, 5, 7, 6, 7, 3

Sample Mean = 21 , Sample Median = 6

***The sample mean is sensitive to outliers , whereas the sample median is not sensitive to outliers*.**

Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ median = old\ median + c$
- Let $y_i = x_i * c$ where c is a constant then $new\ median = old\ median * c$

Mode

Another measure of central tendency is the sample mode

Definition

The mode of a dataset is its most frequently occurring value.

Steps to obtain mode

- If no value occurs more than once, then the dataset has no mode.
- Else, the value that occurs with the greatest frequency is a mode of the data set.

Example 1: 2 , 12 , 5 , 7 , 6 , 7 , 3

7 occurs twice, hence 7 is mode

Example 2: 2 , 105 , 5 , 7 , 6 , 3

No mode

Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ mode = old\ mode + c$
- Let $y_i = x_i * c$ where c is a constant then $new\ mode = old\ mode * c$

Relationship of Mean, Median, and Mode

- In statistics, for a moderately skewed distribution, there exists a relation between mean, median and mode.
- This mean median and mode relationship is known as the "empirical relationship" which is defined as Mode is equal to the difference between 3 times the median and 2 times the mean.

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

Measures of dispersion

- To measure the amount of variation, or spread , in a data set.
- Measures of dispersion are also known as measures of variation or spread.

Measures of dispersion

- Range
- Variance
- Standard Deviation
- Interquartile Range

Range

Definition

- **The range of a data set is the difference between its largest and smallest values.**

$$Range = Max - Min$$

Where Max and Min denote the maximum and minimum observations, respectively.

- **Range is sensitive to outliers.**

REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das

CONCLUSION

1. Types of data – identifying the data type
- categorical data or
- numerical data

2. Visual Representation of numerical data and interpreting the shape of the distribution
- Symmetric Distribution
- Skewed Distribution
- Bimodal Distribution

3. Computed and Interpreted the numerical Summaries of Data
- Compute and Interpret measures of central tendency: Mean, Median, Mode
- Compute and Interpret measures of dispersion: Range


THANK YOU