STATISTICAL METHODS BCH 2 nd sem UNIT -1

Statistics is a branch of mathematics that deals with the collection, review, and analysis of data. It is known for drawing the conclusions of data with the use of quantified models. Statistical analysis is a process of collecting and evaluating data and summarizing it into mathematical form.

Statistics can be defined as the study of the collection, analysis, interpretation, presentation, and organization of data. In simple words, it is a mathematical tool that is used to collect and summarize data.

Uncertainty and fluctuation in different fields and parameters can be determined only through statistical analysis. These uncertainties are determined by the probability that plays a very important role in statistics.

What is Statistics?

In simple words statistics is the study and manipulation of given data. It deals with the analysis and computation of given numerical data. Let us take into consideration some more definitions of statistics given by different authors here:

The Merriam-Webster dictionary defines the term statistics as “The particular data or facts and conditions of a people within a state - especially the values that can be expressed in numbers or in any other tabular or classified way”.

According to Sir Arthur Lyon Bowley, statistics is defined as “Numerical statements of facts or values in any department of inquiry placed in specific relation to each other”.

Statistics Examples

Some real-life examples of statistics are given below:

- To find the mean of the marks obtained by each student in a class of 40 students, the average value is the statistics of the marks obtained.
- Suppose you need to find the number of employed citizens in a city. If the city has a population of 10 lakh people, we will take a sample of 1000 people. Based on this, we can prepare the data, which is the statistic.

Basics of Statistics

Statistics consist of the measure of central tendency and the measure of dispersion. These central tendencies are actually the mean, median, and mode and dispersions comprise variance and standard deviation.

Mean is defined as the average of all the given data. Median is the central value when the given data is arranged in order. The mode determines the most frequent observations in the given data.

Variation can be defined as the measure of spread out of the collection of data. Standard deviation is defined as the measure of the dispersion of data from the mean and the square of the standard deviation is also equal to the variance.

Mathematical Statistics

Mathematical statistics is the usage of Mathematics to Statistics. The most common application of Mathematical statistics is the collection and analysis of facts about a country: its economy, and, military, population, number of employed citizens, GDP growth, etc. Mathematical techniques like mathematical analysis, linear algebra, stochastic analysis, differential equation, and measure-theoretic probability theory are used for different analytics.

Since probability uses statistics, Mathematical Statistics is an application of Probability theory.

For analyzing the data, two methods are used:

1. Descriptive Statistics: It is used to synopsize (or summarize) the data and their properties.
2. Inferential Statistics: It is used to get a conclusion from the data.

In descriptive statistics, the data or collection of data is described in the form of a summary. And the inferential stats are used to explain the descriptive one. Both of these types are used on a large scale.

There is one more type of statistics, in which descriptive statistics are transitioned into inferential stats.

Scope of Statistics

Statistics can be used in many major fields such as psychology, geology, sociology, weather forecasting, probability, and much more. The main purpose of statistics is to learn by analysis of data, it focuses on applications, and hence, it is distinctively considered as a mathematical science.

Methods in Statistics

The statistical process involves collecting, summarizing, analyzing, and interpreting variable numerical data. Some methods of statistics are given below.

- Data collection
- Data summarization

- Statistical analysis

What is Data in Statistics?

Data can be defined as a collection of facts, such as numbers, words, measurements, observations, quantities etc.

Types of Data

1. **Qualitative data-** it is a form of descriptive data.
    - Example- She can write fast, He is tall.
2. **Quantitative data-** it is in the form of numerical information.
    - Example- An elephant has four legs.

Types of quantitative data

1. **Discrete data-** it has a fixed value that can be counted
2. **Continuous data-** it has no fixed value but has a range that can be measured.

Collecting and Summarizing Data

Data:

A collection of observations, facts about an object is known as Data. Data can be in numbers or in statement/descriptive form.

For example,

The statement "How many legs does this table have?". Here, the counted (or collected) value of legs is known as data.

Data Organization of the collected data is required in order to be processed. Information can be provided by processing the data.

Description of Data

There are various ways to describe the data:

Mode:

Mode is the value that occurs very often in the list. It can be said that there is no mode value if no number is repeated in the list.

Median:

Median is the middle value of the list. Median divides the list into two halves.

Mean:

A mean is an average of all the numbers in the list. It can be calculated by adding up all the numbers and then dividing the sum by the number of values in the list.

Range:
The range is the difference between the largest and the smallest numbers.

Types of Statistics

Being a broad term, there exist different models of statistics:

Mean

A mean is an average of two or more numerals. Mean can be computed using Mathematical mean or Geometric mean. The mathematical mean shows how well the commodity performs over the period whereas the geometric mean shows the result of the investment of the same commodity over the same period.

Regression Analysis

It is a statistical process that determines the relationship between variables. It is the process of understanding how the value of a dependent variable changes when any of the independent variables is changed. For example, the price of the property fluctuates due to the particular industry or sector.

Skewness

Skewness is the measure of the distortion from the standard distribution in a set of data. A curve is said to be skewed if it is shifted to the left or to the right. If the curve is extended towards the right side, it is known as the positive skewed and if the curve is extended towards the left side, it is known as the left-skewed.

Kurtosis

Kurtosis is the measure of the tailedness in the frequency distribution. Data set may have heavy-tails or light-tails.

Variance

Variance in statistics is the measure of the data span. It is used to compare the performances of stocks over a period of time.

Representation of Data in Statistics

There are various ways to represent data. For example- graphs, charts and tables. The general representation of statistical data is done with the help of:

- Bar Graph
- Pie Chart
- Line Graph
- Pictograph

- Histogram
- Frequency Distribution

Bar Graph:

It is the rectangular bar representation of data. The bars can be horizontal or vertical. The length of the bar is proportional to the value that it represents. It represents data in the form of rectangular bars having length according to the values that they represent.

There are three types of bar graphs:

a. Vertical Bar Graph
b. Horizontal Bar Graph
c. Double bar Graph

A double bar graph is used to represent the two sets of data in the same graph.

Pie Chart:

It is also known as the Circle Graph as it uses sectors of the circle to represent the data. This graph is represented in the form of a circle which is divided into a various number of sectors where each sector represents a portion of the whole division.

A line graph is represented by the straight line which connects the data points. It is represented by a series of data points called markers. Usually, a line graph is used to represent the change of the data over the period of time.

Pictograph:
It is the representation of the frequency of data using the symbols or pictures. A symbol can represent one or more numbers of data. It represents data with the help of pictures.

Venn Diagrams:
It is the pictorial representation which contains a box along with circles. The box represents the Sample Space and the circles represent the events. There can be three types of Venn diagrams:

a. Two or more than two separate circles (When there is no common data)
b. Overlapping Circles (When some of the data is common)
c. Circle within a circle (When the outer circle is the superset of the inner circle)

Histogram:

It consists of rectangles Whose area is proportional to the frequency of a variable and whose width is equal to the class intervals.

Frequency Distribution:

The frequency of a value is represented by “f”. In a frequency table, specific data and values are arranged in ascending order of the given magnitude with their corresponding available frequencies.

Applications of Statistics

Information around the world can be determined mathematically through Statistics. There are various fields in which statistics are used:

1. **Mathematics:** Statistical methods like dispersion and probability are used to get more exact information.
2. **Business:** Various statistical tools are used to make quick decisions regarding the quality of the product, preferences of the customers, the target of the market etc.

3. **Economics:** Economics is totally dependent on statistics because statistical methods are used to calculate the various aspects like employment, inflation of the country. Exports and imports can be analysed through statistics.
4. **Medical:** Using statistics, the effectiveness of any drug can be analysed. A drug can be prescribed only after analysing it through statistics.
5. **Quality Testing:** Statistics samples are used to test the quality of all the products a Company produces.
6. **Astronomy:** Statistical methods help scientists to measure the size, distance, etc. of the objects in the universe.
7. **Banking:** Banks have several accounts to deposit customers' money. At the same time, Banks have loan accounts as well to lend the money to the customers in order to earn more profit from it. For this purpose, a statistical approach is used to compare deposits and the requesting loans.
8. **Science:** Statistical methods are used in all fields of science.
9. **Weather Forecasting:** Statistical concepts are used to compare the previous weather with the current weather so as to predict the upcoming weather.

There are various other fields in which statistics is used. Statistics have a number of applications in various fields in Mathematics as well as in real life. Some of the major uses of statistics are given below:

- Applied statistics, theoretical statistics, and mathematical statistics
- Machine learning and data mining
- Statistical computing
- Statistics is effectively applied to the mathematics of the arts and sciences
- Used for environmental and geographical studies
- Used in the prediction of weather
-

Data Collection Methods

To analyze and make decisions about a certain business, sales, etc., data will be collected. This collected data will help in making some conclusions about the performance of a particular business. Thus, data collection is essential to analyze the performance of a business unit, solving a problem and making assumptions about specific things when required. Before going into the methods of data collection, let us understand what data collection is and how it helps in various fields.

What is Data Collection?

In Statistics, data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends. Once the data is collected, it is necessary to undergo the data organization process.

The main sources of the data collections methods are “Data”. Data can be classified into two types, namely primary data and secondary data. The primary importance of data collection in any research or business process is that it helps to determine many important things about the company, particularly the performance. So, the data collection process plays an important role in all the streams. Depending on the type of data, the data collection method is divided into two categories namely,

- Primary Data Collection methods
- Secondary Data Collection methods

In this article, the different types of data collection methods and their advantages and limitations are explained.

Primary Data Collection Methods

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations. The primary data collection method is further classified into two types. They are

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Let us discuss the different methods performed to collect the data under these two data collection methods.

Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures. This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data. They are

Observation Method

Observation method is used when the study relates to behavioural science. This method is planned systematically. It is subject to many controls and checks. The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

Interview Method

The method of collecting data in terms of verbal responses. It is achieved in two ways, such as

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple
- Should follow a logical sequence
- Provide adequate space for answers
- Avoid technical terms
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

Schedules

This method is similar to the questionnaire method with a slight difference. The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

Secondary Data Collection Methods

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

Published data are available in various resources including

- Government publications
- Public records

- Historical and statistical documents
- Business documents
- Technical and trade journals

Unpublished data includes

- Diaries
- Letters
- Unpublished biographies, etc.

Visit BYJU'S -The Learning App for Maths related articles and also watch personalized videos to learn with ease.

Frequently Asked Questions – FAQs

What are the 4 methods of data collection?

The 4 methods of data collection are:
Observation method
Interview method
Questionnaire method
Schedules

What is data collection and its types?

Data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to estimate the outcome of the situation. The data collection methods enable you to conclude an answer to the relevant question. Some of the data collection types include surveys, delphi technique, focus groups, interviews and so on.

What are the primary data collection methods?

As we know, the primary data collection is expensive and time consuming. The primary data collection methods are:
Observation method
Interview method
Questionnaire method
Schedules
Surveys

What are data collection tools?

The devices or instruments used to collect the data are called data collection tools. The tools are questionnaires on papers or system based (virtual form) interviews, checklists, interviews, etc.

What are quantitative data collection methods?

Quantitative data collection methods are a part of the primary data, i.e. a type of information that is obtained directly from the first-hand source through experiments, surveys, or observations.

1.2 MEASUIRES OF CENTRAL TENDENCY

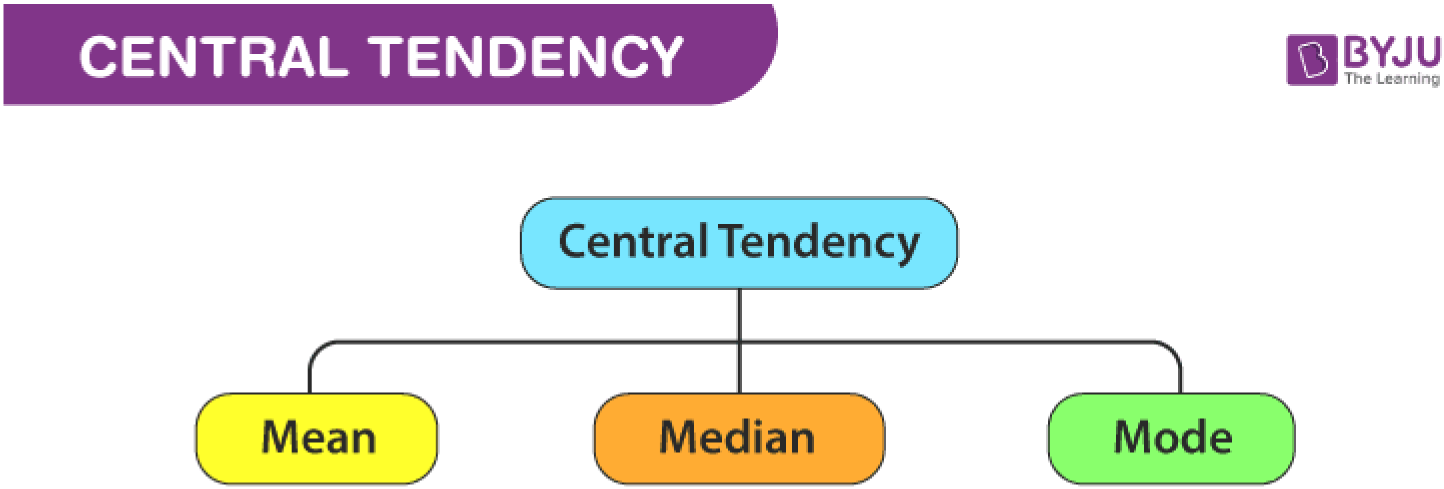
The following are the five measures of average or central tendency that are in common use : (i) Arithmetic average or arithmetic mean or simple mean (ii) Median (iii) Mode (iv) Geometric mean (v) Harmonic mean Arithmetic mean, Geometric mean and Harmonic means are usually called Mathematical averages while Mode and Median are called Positional averages.

Definition

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.

Measures of Central Tendency

The central tendency of the dataset can be found out using the three important measures namely mean, median and mode.



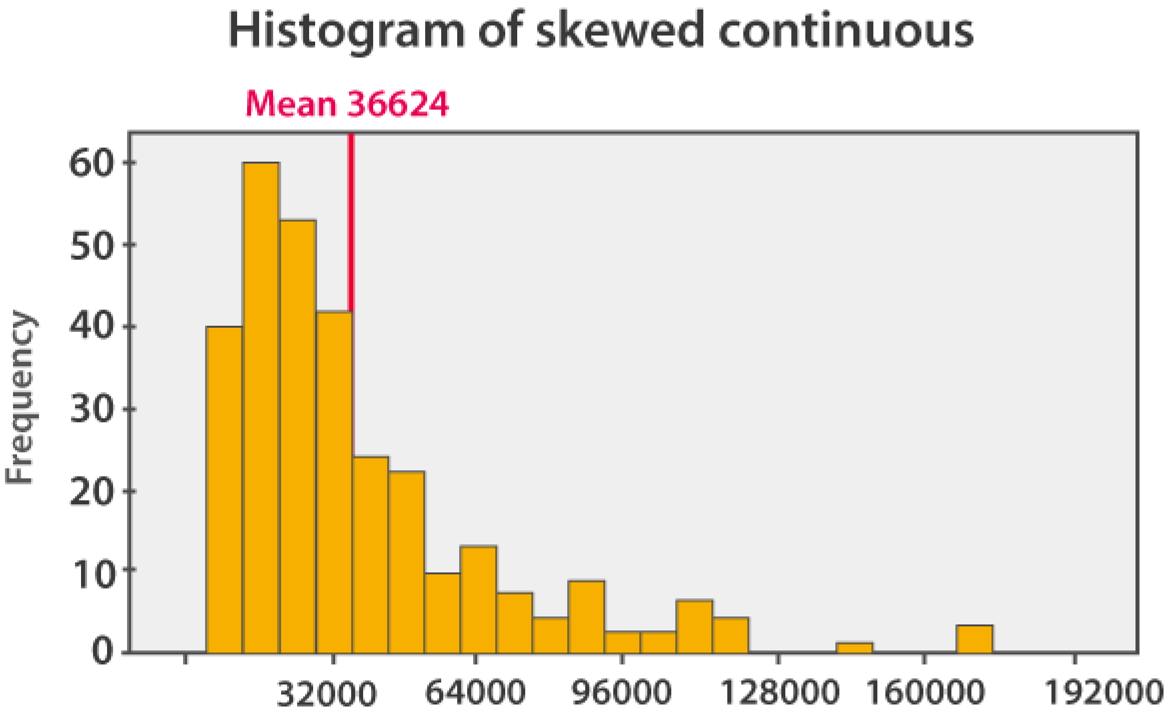
Mean

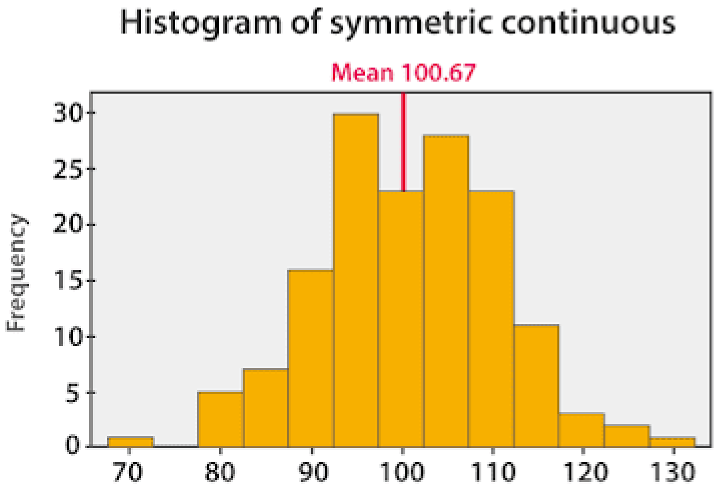
The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean. Some other measures of mean used to find the central tendency are as follows:

- Geometric Mean
- Harmonic Mean
- Weighted Mean

It is observed that if all the values in the dataset are the same, then all geometric, arithmetic and harmonic mean values are the same. If there is variability in the data, then the mean value differs. Calculating the mean value is completely easy. The formula to calculate the mean value is given as

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.



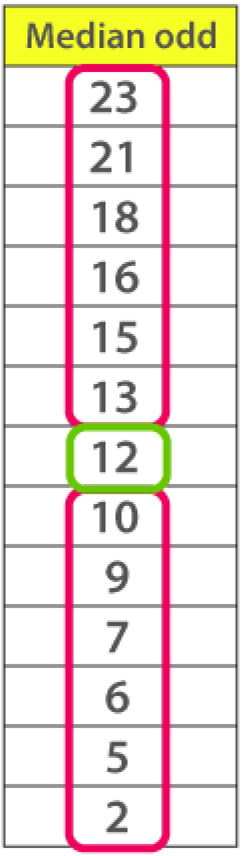


In symmetric data distribution, the mean value is located accurately at the centre. But in the skewed continuous data distribution, the extreme values in the extended tail pull the mean value away from the centre. So it is recommended that the mean can be used for the symmetric distributions.

Median

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.

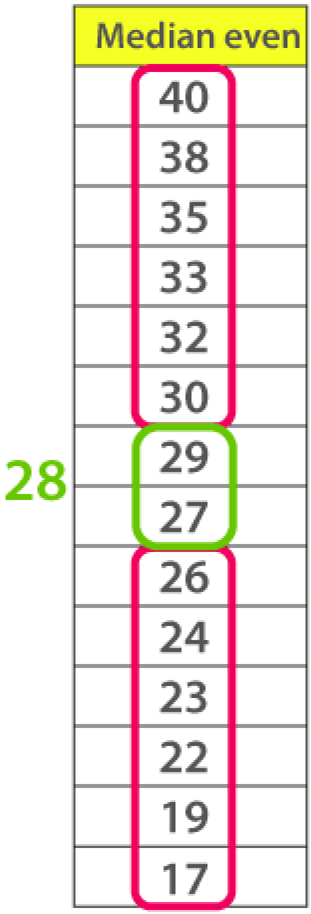
Consider the given dataset with the odd number of observations arranged in descending order – 23, 21, 18, 16, 15, 13, 12, 10, 9, 7, 6, 5, and 2



Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2

Here 12 is the middle or median number that has 6 values above it and 6 values below it.

Now, consider another example with an even number of observations that are arranged in descending order – 40, 38, 35, 33, 32, 30, 29, 27, 26, 24, 23, 22, 19, and 17



Median even
40
38
35
33
32
30
29
27
26
24
23
22
19
17

When you look at the given dataset, the two middle values obtained are 27 and 29.

Now, find out the mean value for these two numbers.

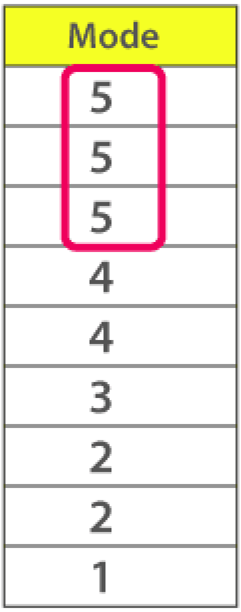
i.e.,(27+29)/2 =28

Therefore, the median for the given data distribution is 28.

Mode

The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all.

Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5



Mode
5
5
5
4
4
3
2
2
1

Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.

Based on the properties of the data, the measures of central tendency are selected.

- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency.
- If you have categorical data, the mode is the best choice to find the central tendency.