FINGERTIPS

Module 2

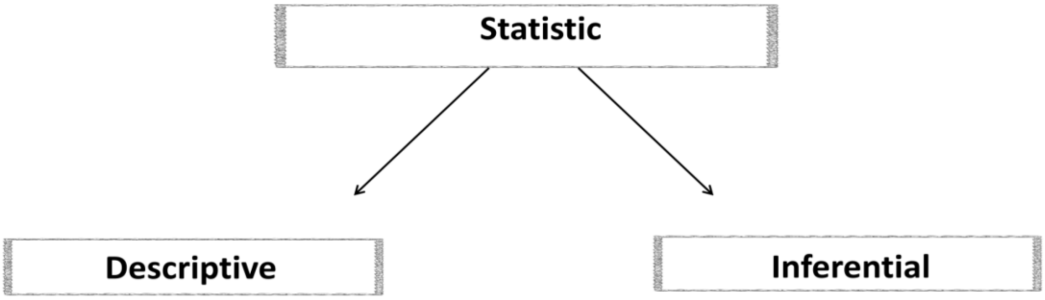
Statistical Fundamental-I

What is Statistics-

Statistics is a branch of Mathematics dealing with Data Collection, Organization, Analysis, Interpretation and Presentation.

Statistics - As defined by the American Statistical Association (ASA)- is the science of learning from data and of measuring, controlling and communication uncertainty.

Types of Statistics



Descriptive Statistics

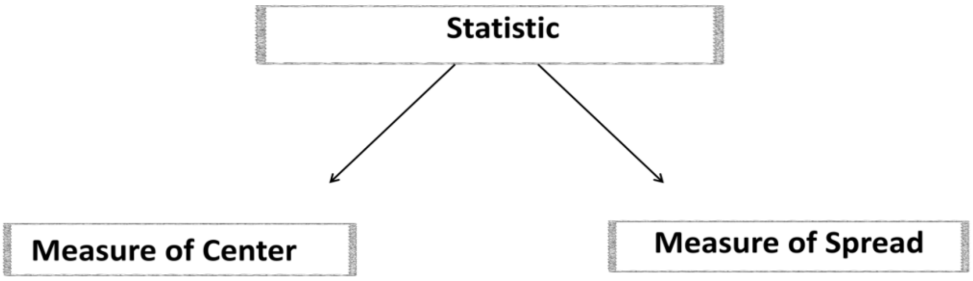
It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

Inferential Statistics

Inferential Statistics is used to draw inferences beyond the immediate data available. We can answer the following questions with the help of inferential statistics: making inferences about the population from the sample.

Descriptive Statistics	Inferential Statistics
Descriptive statistics work with smaller data. There is no need for sampling and the entire population data is available.	Inferential statistics work with large data set. Analyzing entire population based on sample parameter is a strength.
Process is simpler to do	Process is complex as we have to decide best sampling technique.
Descriptive statistics are likely to be 100% accurate because there is no assumption.	This is not 100% accurate. inferential statistics always make inference about larger population based on sample.
Find results are shown in form of charts, tables and graphs.	Find result in probability score.
Tool-Measure of central tendency (mean, median, mode) spread of data (range standard deviation)	Tool-hypothesis test, analysis of variance.
Organize, analyze and present data in meaningful manner.	Compress, test and predicts future outcome.

Type of Descriptive Statistics



Basic Terminologies

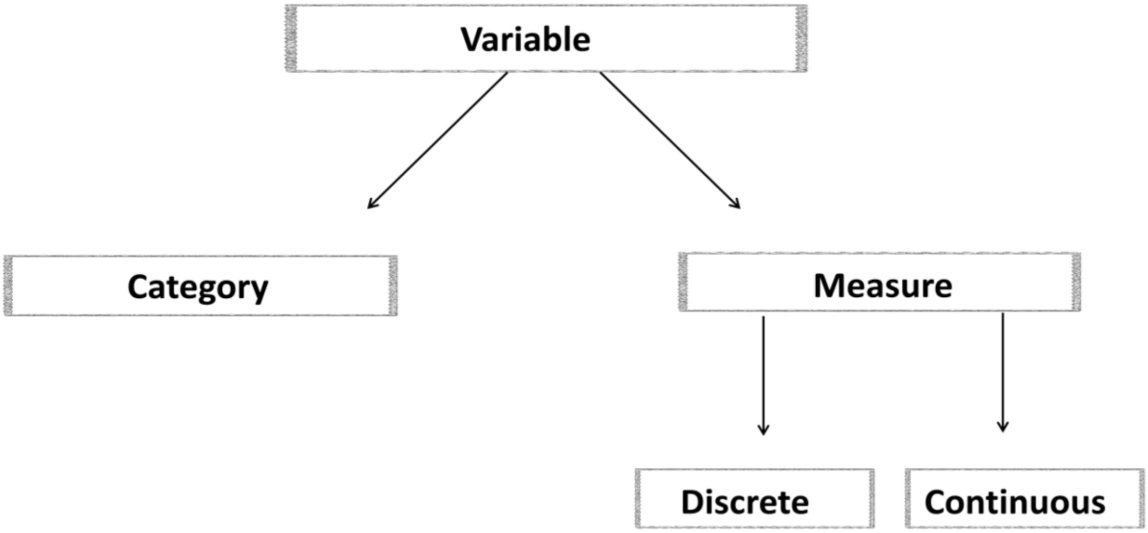
- Population
- Sample
- Variable

Population is the set of sources from which data has to be collected. Population in statistics include all members of a define group that we are studying or collecting information on for data driven decisions.

A sample is a subset of the population

What is Variable-

- A variable is named value that changes
- Simply something that varies like
- Your weight
- Environment Temperature
- Your Marks



Categorical Variable

Categorical variables have values that describe a ‘quality’ or ‘characteristic’ of a data unit, like ‘what type’ ‘or which category’.

Categorical variables can be put into categories. Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value.

Measure Variable

Measure variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore, numeric variables are quantitative variables.

Discrete Variable

A discrete variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value.

Examples of discrete variables include-

- The number of registered cars, number of business locations
- And number of children in a family
- All of which measured as whole units (i.e. 1, 2, 3 cars).

Continuous Variable

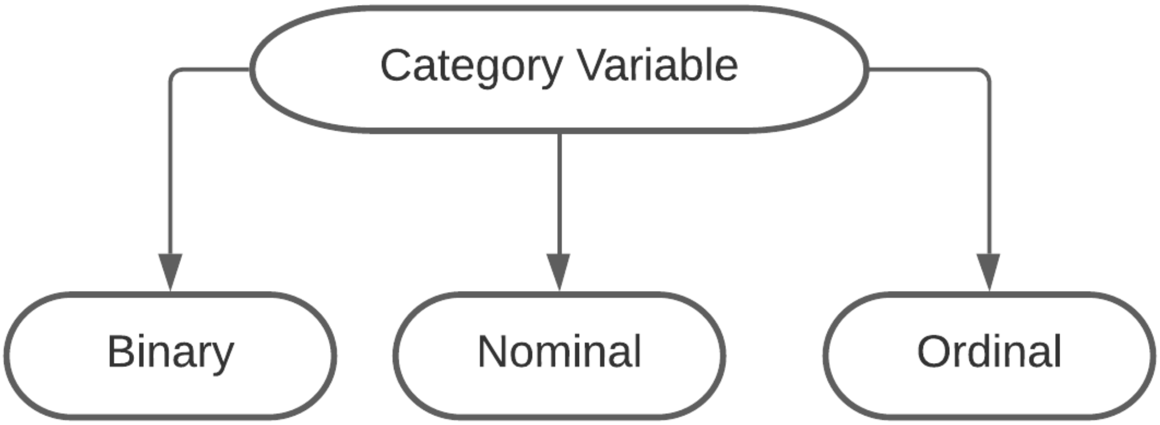
- Height
- Age
- temperature.

Categorical Variable

- Categorical variables have values that describe a 'quality' or 'characteristic ‘of a data unit, like 'what type' or 'which category'.

- Categorical variables than can be put into categories. For example, the category “Toothpaste Brands” might contain the variables Colgate and Aquafresh.

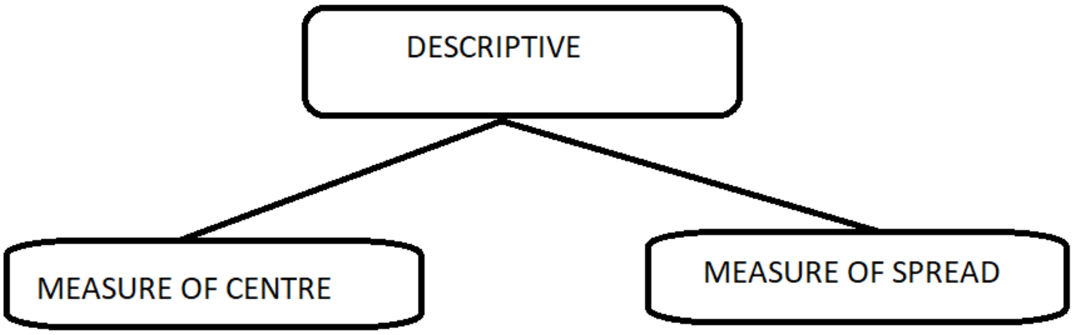
- Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value.



Type of Variable	Data Represent	Example
Binary Variables	Yes/No outcomes	Head/ tails in coin flip Win/Lose in match
Nominal Variables	Group with no rank or order Between them	• Species names • Colors • Brands
Ordinal Variables	Group with that are ranked in a specific order.	Finishing place in a race rating scale response in a survey

Type of Variable	Data Represent	Example
Independed Variables	Variables you manipulate in order to affect the outcome of an experiment.	Sales is independent from Profit.
Depended Variables	Variables you manipulate in order to affect the outcome of an experiment.	Profit depends on Sales.

Control Variables	Variables you manipulate in order to affect the outcome of an experiment.	Any measurement of plant health and growth: in this case, plant height and wilting



Measure of Centre/ Measure of central tendency -

- Mean
- Mode
- Median

Mean-

It is simply the average of all the data (salary) values. Add all the numbers then divide by the amount of numbers

16000+15000+10000+12000+8000+18000=79000/6

The Mean is = 13166

Mean real life example-

- What is mean of your last year expense?
- What is the average salary of your employee?
- What is the mean of your graduation score?
- What is mean of your bank accounts?

Merits of Mean

- It can be easily calculated; and can be easily understood. It is the reason that it is the most used measure of central tendency.
- As every item is taken in calculation, it is affected by every item.
- Fluctuations are minimum for this measure of central tendency when repeated samples are taken from one and the same population.
- It can further be subjected to algebraic treatment unlike other measures i.e. mode and media
- It does not depend upon any position.

Demerits of Mean

- It cannot be located graphically
- Its value will be effective only if the frequency is normally distributed. Otherwise, if case skewness is more, the results become ineffective
- Sometimes it gives impossible or laughable conclusions, e.g., if there are 60,50, and 12 students in three classes then average number of students is 60+50+42/3 = 50.67, which is impossible as students can’t be in fractions.
- A single item can bring big change in the result. For example, id there are three terms 4, 7, 10; X is 7 in this case. If we add a new term 95, the new X is 4+7+10+95/4 = 116/4 = 29. This is a big change as compared to the size of first three terms.

Median-

It is the value in the middle when the data items are arranged in ascending order.
If you have two middle values then you should take average of both

9,3,1,8,3,6

1,3,3,6,8,9

The Median is 4.5

9,1,3,6,8

1,3,6,8,9

The Median is 6

Merits of Median

- It is simple to understand
- It does not require all the observations of the data for its determination
- It is not affected by the extreme values of a series
- It can be determined graphically which is shown a little later along with the quartiles etc.
- It can be determined easily in open and series without estimating the lowest or highest-class limits.

Demerits of Median

- It is not based upon all value of given data.
- It is not based on all the observations of the series.
- It is insensitive to some changes in data value
- It needs the arrangement of a series in ascending or descending order and more particularly in a frequency distribution
- It is very much affected by fluctuations in sampling.

Mode

It is the most frequently occurring value in a series of data in case of no repeating values, there would be no mode.

9,3,1,8,3,6

The Mode is 3

Merits of Mode

- It gives the most representative value of a series
- It is always present within the data
- It is capable of studying qualitative data as its determination depends on the frequencies rather than the values of the items
- It can be determined graphically either through a histogram, or through a frequency polygon.
- It is considered a reliable average for studying skewness of a distribution.
- It is not affected by the extreme values of a series. For example, let a series be as below.

10	12	15	14	15	16	17

- It is understood by a layman as it refers to a value that occurs for a maximum time, thus, when we talk of a modal size of shoes, a layman easily understands that it refers a size of shoes which demanded by the maximum number of customers.
- It is very much useful in the field of business, and commerce as it helps a businessman in taking a decision on the varieties of the goods, he should procure in large quantities to enhance his sales.

Demerits of Mode

- It is not based on all the observations of a series but on the concentration of frequencies of the items. If any non-modal value is left out of the series, or is added, the value of the mode is not altered.

- In case of a continuous and bimodal series, its determination becomes difficult, and lingering as it involves passing through a number of trials and use of interpolation formulae in those cases.
- It can't be easily determined graphically, if two, or more values of a series have the same highest frequency. Thus, for the following series, mode cannot be easily located through histogram.

Marks:	1-35	36-70	71-105	106-140	141-175
F:	35	25	40	20	40

- It cannot be determined from a series with unequal class intervals unless they are equalized on the assumption that the frequencies are evenly distributed, and such assumption may not also hold good
- In certain cases, it contradicts its very meaning and nature when certain value with lesser frequency is determined as the model values.

Mode real life example-

- I want to know the Age mode of our class?
- If we want to know about Modi's foreign visits?
- If mall owner wants to know his best seller product?
- Your restaurant visit will talk about your favorite restaurant/dish

What is ANALYTICS?

- Analytics provides us with meaningful information which may otherwise be hidden from us within large quantities of data
- Analytics uses data and math to answer business questions, discover relationships, predicts unknown outcomes and automate decisions.

ANALYTICS v/s ANALYSIS

Analysis	Analytics
We perform analysis on things that have already happened in past.	Analytics is working on future
The why? How? What? Of happened in past.	Analytics is utilizing machine learning, Statistics, algorithm, models to take better decisions and get better insight from data
What we earn last year	Analytics is defined as a process of Transforming data into action.

Types of Analytics

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

DESCRIPTIVE

- This can be termed as the simplest form of analytics. The mighty size of big data is beyond human comprehension and the first stage hence involves crunching the data into understandable chunks.
- The purpose of this analytics type is just to summarize the findings and understand what is going on.

DIAGNOSTIC

- Diagnostic analytics is used to determine why something happened is the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations.
- Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

REDICTIVE

- Predictive analytics is used to predict future outcomes. However, it is important to note that it cannot predict if an event will occur in the future; it merely forecasts what are the probabilities of the occurrence of the event.

- A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes.

PRESCRIPTIVE

- The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take. Prescriptive analysis is typically not just with one individual action, but is in fact a host of other actions.

Type of Analysis-

- **Qualitative**
- **Quantitative**

Qualitative Analysis

- It is mostly deals with generic data using text media
- It is also known as non-statistical analysis
- It deals with description
- Data can be observed but not measured
- Like – colors textures, smell, taste, beauty
- Qualitative = Quality

It is based on meaning expressed through words

Quantitative Analysis-

- It is the science of collecting and interpreting objects with numbers.
- It is also known as statistical analysis
- It is based on meaning derived from number.
- Data which can be measured.
- Like –length, height, area, weight, speed, time, temperature, cost age

- Quantitative= Quantity

Measure of spread/ dispersion

- As the name suggests, the measure of dispersion shows the scattering of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data.
- Characteristics of Measure of Dispersion
  - A measure of dispersion should be rigidly defined
  - It must be easy to Calculate and understand
  - Not affected much by the fluctuations of observations
  - Based on all observation

Types of Measure of Dispersion

Absolute measure of dispersion

The measures which express the scattering of observation in terms of distances i.e., range, quartile, deviation.

Relative measure of dispersion

The measures which express the variations in terms of the average of deviations of observation like mean deviation and standard deviation.

Range

Measure of spread/ dispersion

- As the name suggests, the measure of dispersion shows the scattering of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data.
- Characteristics of Measure of Dispersion
  - A measure of dispersion should be rigidly defined
  - It must be easy to Calculate and understand
  - Not affected much by the fluctuations of observations
  - Based on all observation

Types of Measure of Dispersion

Absolute measure of dispersion

The measures which express the scattering of observation in terms of distances i.e., range, quartile, deviation.

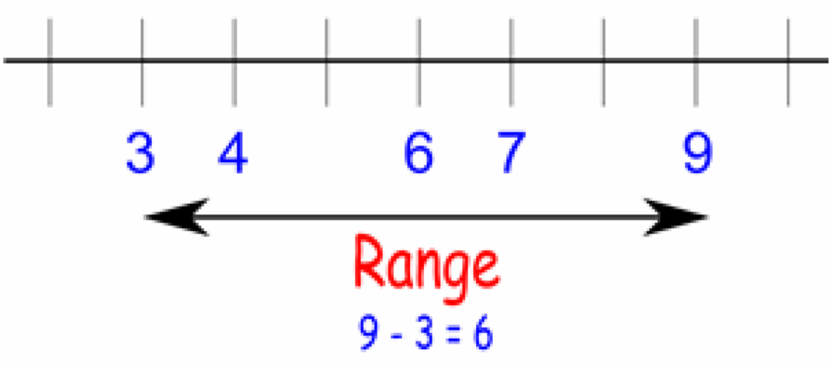
Relative measure of dispersion

The measures which express the variations in terms of the average of deviations of observation like mean deviation and standard deviation.

Range

- A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set.
- If X max and X min are the two extreme observations then

Range = X max – X min



Merits of Range

- It is the simplest
- Easy to calculate
- Easy to understand
- Independent of change of origin

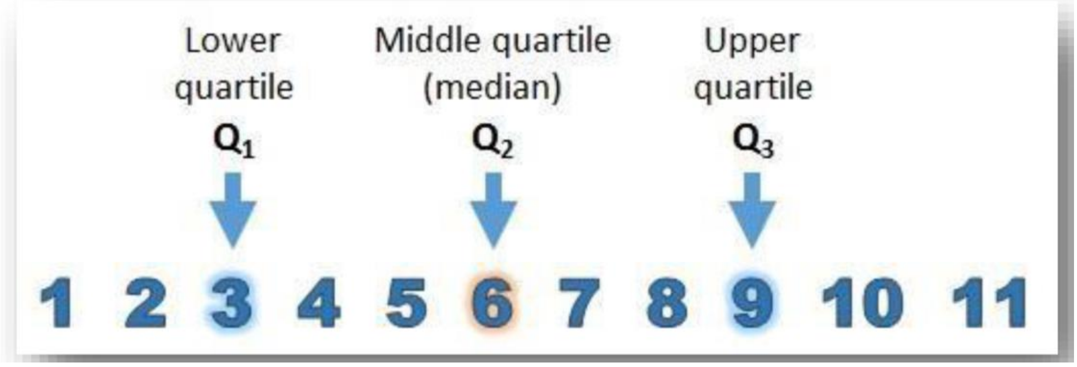
Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on the change of scale

Quartile-

The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The second quartile, (Q2) is the median of the data set. The third quartile, (Q3) is the middle number between the median and the largest number.

A quartile divides a sorted data set into 4 equal parts, so that each part represents ¼ of data set.



Merits of Quartile

- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile

- It uses half of the data
- Independent of change of origin

- The best measure of dispersion for open-end classification

Introduction to Variance

- Variance is the average squared deviation from the mean of a set of data.
- It is used to find the standard deviation.

Introduction to Standard Deviation

- Standard Deviation shows the variation in data.
- If the data is close together, the standard deviation will be small.
- If the data is spread out, the standard deviation will be large.
- Standard Deviation is often denoted by the lowercase Greek letter sigma.
- The bell curve is commonly seen in statistics as a tool to understand standard deviation.
- The following graph of a normal distribution represents a great deal of data in real life. The mean or average is represented by the Greek letter μ, in the center.

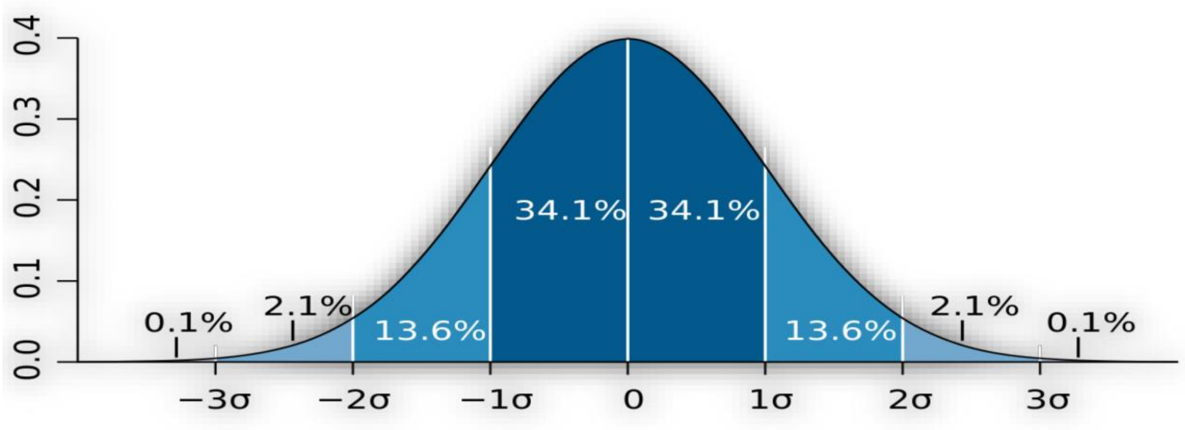
Merits of Standard Deviation

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations.
- Suitable for further mathematical treatment.
- Least affected by the fluctuation of the observations.
- The standard deviation is zero if all the observations are constant.

- Independent of the change of origin.

Demerits of the Standard Deviation

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale.



How to find Standard Deviation

- Both are the popular measures of how spread out the data points are from a center value mean
- For example, let’s find the standard deviation of the following data: 1,2,2,4,6

1. Calculate the mean of data: 15/5 = 3
2. Subtract the mean from each data value: -2, -1, -1, 1, 3
3. Square each of the new data value: 4,1,1,1,9
4. Sum these squared data values: 16
5. Divide this sum by (number of observations -1): 16 / (5-1) = 4
6. This number is Variance and Square root of this number is standard deviation: Sqrt (4) = 2

- For instance, standard deviations of price data are frequently used as a measure of volatility; While monitoring some industrial process, if process indicators go beyond design standards then it may be troublesome hence variance/standard deviation can be used in such cases.