① Introduction to statistics :-

→ statistics is the science of collecting, organizing and analyzing the data.

→ Data is facts/pieces of information that can be measured, collected and analyzed.

Examples of data can be weights of students inthe class like { 60Kg, 65Kg, 70Kg -- }

→ Before understanding the role of statistics in real world, let us understand what actually data analysts and data scientists do.

For example, we have a house price dataset.

city	area	no. of rooms	price
Bangalore	1000	2	50lakhs
Mumbai	1500	3	80lakhs
Pune	1250	2	60 lakhs
Hyderabad	1700	4	65 lakhs

By using this dataset, data scientist can create a model using machine learning/deep learning where we can predict house price by providing city, area and number of rooms to the model.

Data analysts use this data to create visual reports which help people make meaningful decisions.

The statistics is involved in both jobs performed by data scientists and analysts. For data scientists, it helps in building the model and for data analysts, it helps in providing conclusions and meaningful summarization of data.

→ Applications of statistics :-

(i) Data exploration and summarization.
(ii) model building and validation.
(iii) statistical analysis
(iv) Hypothesis testing
(v) optimization and efficiency
(vi) creating reports.

(A) Types of statistics :-

(i) Descriptive statistics :-

Descriptive statistics involves methods for summarizing, organizing data to make it understandable. This describes the basic features of data in a study.

It mainly consists of :-

(a) Measure of central tendency
mean, median, mode

(b) Measure of dispersion
variance, standard deviation

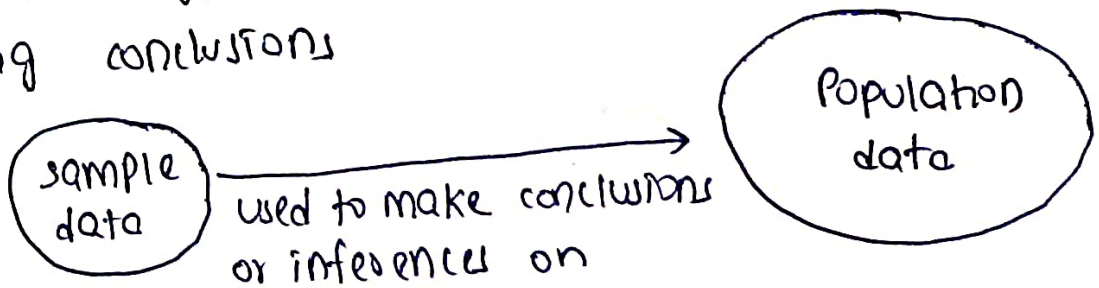
(c) Data distribution
Histogram, box plot, pie chart, PDF, PMF

(d) Summary statistics
Five number summary

Some examples of descriptive statistics can be finding out mean test scores of students in a class, plotting histogram of students in a school etc.

(ii) Inferential statistics :-

It involves methods for making predictions or inferences about a population based on a sample of data. This allows for hypothesis testing, estimations and drawing conclusions



Inferential statistics mainly consists of:-

(a) Hypothesis testing

(b) P value

(c) confidence interval

(d) Statistical analysis tests like Z-test, t test, ANOVA test and chi-square test.

Some examples for inferential statistics can be finding p-values in test scores comparison, finding 95% confidence interval for average height, predicting house prices and comparing test scores of different schools.

(A) Population and sample data:-

→ A population is the entire set of individuals or objects in interest in a particular study. It includes all members of a defined group that we are studying or collecting information on.

→ A sample data is a subset of population data that is used to represent the entire group. sampling involves selecting a group of individuals/observations from the population to draw conclusions about whole population.

→ Characteristics of population data:-

(i) Population data is the complete set.

(ii) Parameter is the numerical values that summarizes the entire population data like
- Population mean (μ)
- Population variance (σ^2)

(iii) some examples can be finding average height of all students in a class, average income of all the employees in an organization etc.

→ Characteristics of sample data:-

(i) sample data is a subset of population i.e represents only some portion of the population.

(ii) we use statistic which is a numerical value that summarizes the sample data like
- Sample mean
- sample variance

(iii) we use random sampling to avoid bias to select a subset of data from population data.

(iv) some examples of sample data can be predicting election results where we interview a certain groups of people to provide conclusions about entire population.

(A) Types of sampling techniques :-

Sampling techniques are majorly divided into two categories, which are

(i) Probability sampling

(ii) Non-probability sampling

(A) Probability sampling

(i) simple random sampling :-

Every member of the population has an equal chance of being selected.

Example of simple random sampling can be selecting people randomly, selecting a students randomly from the class based on student ID.

(ii) systematic sampling :-

selecting every n^{th} member of the population after a random starting point.

Example can be to be the n^{th} customer to get discount in a shop, selecting every n^{th} member for a feedback survey etc.

(iii) stratified sampling :-

Dividing population into strata (groups) based on specific characteristics and then randomly sampling from each strata.

Example can be dividing employees in a company based on department and random sampling within each department.

Another example can be dividing people into three strata (age ≤ 12, age 13-18 and age > 18) and then sampling from each strata.

(iv) cluster sampling :-

Dividing population into clusters, randomly selecting clusters and then sampling all the members from selected clusters.

Example can be randomly selecting several schools from a district and surveying all teachers within those schools.

(v) multi-stage sampling :-

It involves combining several sampling methods like selecting clusters and then random sampling within those clusters.

Randomly selecting cities and then at each selected cities, randomly selecting households for surveys.

(*) Non-probability sampling :-

In this sampling we select individuals that are the easiest to reach.

(i) convinience sampling :-

selecting individuals whom are easiest to reach.

For example, surveying people at a mall.

(ii) Judgmental / purposive sampling :-

Selecting individuals based on researcher's judgment.
Here, sampling is about usefulness / representation.
Example for this is choosing machine learning experts to survey about a machine learning paper.

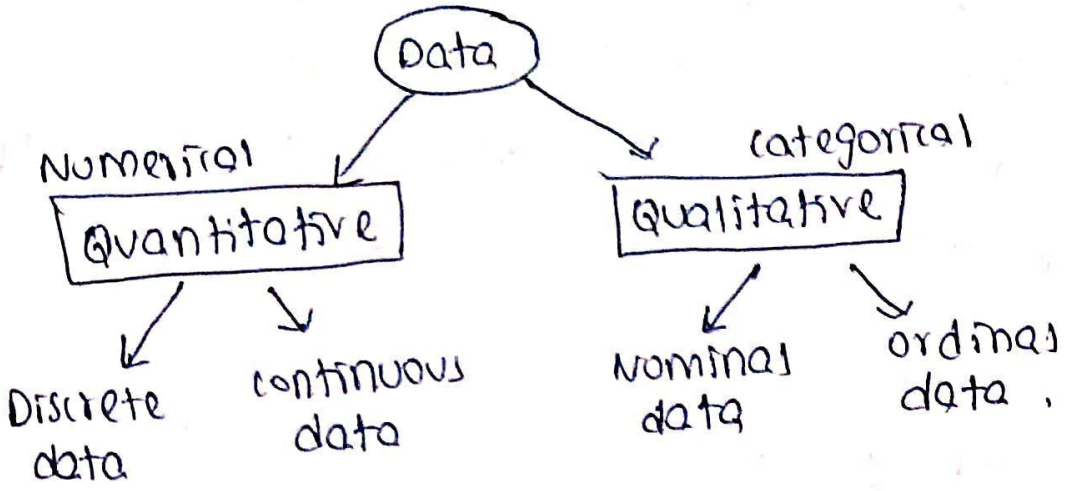
(iii) Snowball sampling :-

It is a sampling method where existing study participants recruit future participants from among their acquaintances.
For example, surveying people having rare disease and asking them if they know anyone having the same disease.

(iv) Quota sampling :-

It is sampling method based on individual's quota which can be gender / caste / religion.
Example for this is surveying people of certain gender about their perspective on social issues.

ⓐ Types of data :-



⇒ **Discrete data :-**

It is a type of quantitative data which contains positive whole numbers.

Example of discrete data :-
- Number of bank accounts (1/2/3/4 ---)
- Number of children in a family (0/1/2/3 -)

→ **continuous data :-**

It is a type of quantitative data that can contain any numerical value including decimals and negative values.

Example of continuous data :-
- weights of students (80.52, 82, 65)
- Temperature (-27.4°C, 28°C, 50°C...)

→ **Nominal data :-**

It is a type of qualitative data that can contain a value from a set of values.

Examples of nominal data :-
- Role in cricket (batsman / bowler / keeper / all-rounder)
- Gender (male / female / others)

⇒ **ordinal data :-**

It is a type of qualitative data that can hold a value from a set of values. The only difference from the nominal data is that the ordinal data can be ranked.

Examples of ordinal data :-

- Customer feedback (Good / Bad / worst)
  It can be ranked as Good > Bad > worst.

- Levels in a game (hard / moderate / easy)
  Levels can be ranked as hard > moderate > easy.

(x) Scales of measurement of data:-

The scales of measurement of data describes the nature of information within the value assigned to variables. The primary scales of measurement of data are:-

(i) Nominal scale
(ii) Ordinal scale
(iii) Interval
(iv) Ratio

⇒ Nominal scale:-

Nominal scale classifies data into distinct categories that donot have an intrinsic order.
Since we are dealing with categories, this scale is for qualitative / categorical data.

Characteristics of nominal scale:-

→ Data is categorised based on labels/names/qualities.
→ The categories in nominal scale are mutual exclusive. i.e one record can contain only one category.
→ No logical order exists among the categories i.e categories cannot be ranked.

Examples of nominal scale:-

→ Gender (male or. Female)
A record can only have male/female but not both at the same time.

→ Colors (yellow, red, ...)
A record can hold only one color. so this data can be measured in nominal scale.

→ Types of cuisines (Indian / Italian / chienes ..)
These categories cannot be ordered and a record can contain only one category. So, this can be measured through nominal scale.

⇒ Ordinal scale:-
This scale classifies the data into categories that can be ranked or ordered.

Characteristics of ordinal scale:-
→ Data is categorised (qualitative) and ranked in a specific order / intrinsic order.
→ The interval between the ranks are not necessarily equal.

Examples of ordinal scale:-
→ Education levels (High school / Bachelors / masters / Ph.D)
categories can be ranked as,
Ph.D > masters > Bachelors > High school

→ customer feedback (statisfied / very satisfied / not satisfied)
categories can be ranked as,
very satisfied > satisfied > not-satisfied

⇒ Interval scale:-
The interval scale not only categorises data and orders them, but also specify the exact difference between the intervals. It lacks a true zero point.

Characteristics of interval scale:-

→ Data is ordered with consistent intervals between the values.

→ Interval scale allows for meaningful comparison of differences.

→ It has no true zero point.

Examples of interval scale:-

→ Temperature in Fahrenheit (10°F, 20°F .... 60°F)
0°F means no temperature which makes no sense.
Differences are consistent i.e
20°F - 10°F = 10°F } same differences
30°F - 20°F = 10°F

→ IQ scores of students in the class {80, 85, 90, 95--}
0 IQ simply means no IQ which is useless.
Differences are consistent i-e
85-80 = 5 } same differences between IQ
90-85 = 5

→ Ratio scale:-
In this scale, the data is ordered, differences are measurable and contains a 0 starting point. Ratio can be measured in this case.

Example of ratio scale:-

→ Student marks {0, 30, 45, 60, 75 -}
The differences are measurable in this.

student marks can be ordered as 75 > 60 > 45 > 30 > 0. The ratio also makes sense here sine we can say that the person who scored 60 has worked 2 times harder than the person who scored 30 since

$$60/30 = 2/1 = 2:1$$

→ <u>some examples for scales of measurements of data:-</u>

(i) Length of different rivers in the world
It can be measure by ratio scale because, the lengths can be ordered, differences between the lengths are measurable. Ratio also makes sense since a river of length 60km can be interpreted as twice longer as another river of length 30km.

(ii) Favorite food based on gender
It has nominal scale of measurement since the food cannot be ordered and is categorical.

(iii) IQ measurement:-
This can be measured through interval scale since there is no true zero point.