29/10/25

Statistics :

Introduction :

→ To check the quality of the data.

for this check we have to use 2 stastics.

1. Descriptive Statistics
2. Inferential Statistics.

1. Descriptive Statistics :

To make a statement or a conclusion.

2. Inferential Statistics :

It uses sample data to make inference or generalizations about a population.

What is Statistics ?

→ It is Science of collecting, organizing and Analyzing data. (for better decision making).

What is data ?

→ Data means facts or pieces of information that also can be measured.

→ Information contain data but data didn't contain information.

Example :

The IQ of a class Students.

{ 98, 97, 68, 57, 110 } → Avg, Min, Max.

Descriptive Statistics :

It Consist of organizing and Summarizing Data.

Inferential Statistics :

Technique where we used the data that we have measured to form conclusion.

(OR)

To make a statement / Conclusion on a Descriptive Statistics we use inferential statistics.

Ex :

1. Are the Avg marks of the Java class Students is same as python class Students in the Besant.?

A) The above statement is inferential Statistics.

2. What is the Avg Marks SQL Students ?

A) Descriptive Statistics.

Population (N) and Sample (n)

Population(N) :

The entire group of the data we call it as a Population.

Ex: All people in India.

→ A Subset of a population we call it as a Simple.

Ex:

One lakh people from different region of India.

Key points:

→ populations are larger than Samples.

→ Sample should be representative of the population.

→ Samples allow for easier, faster and less costly Data collection.

Types of Sampling techniques:

1. Simple random sampling.

Every member of a population as an equal chance of being selected for our Sample.

Example:

→ The Avg milage of a bike

→ What is Avg ratio of married people. in banglore.

2. Stratified Sampling:

→ Strata is nothing but a Group.

→ Where the population is split into non overlapping Groups.

Example :

1. When the man is alive or dead.
2. The person is good ~~but~~ or bad. ~~also~~ and.

3. Systematic Sampling :

From the population every n^{th} sample we will collect.

Example :

While doing survey in the mall on the topic of modernaization collecting information on every 5^{th} person who is coming out from the mall.

4. Convienence Sampling :

The sample is collected based on our convience from the particular domain experts.

30/10/25

Note : Sampling technique selection always depends on problem statement.

Variable : A variable is the property that can take an any value

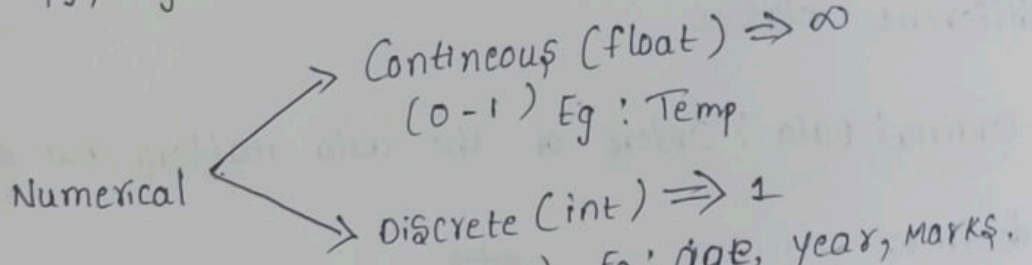
Two kinds of variables :

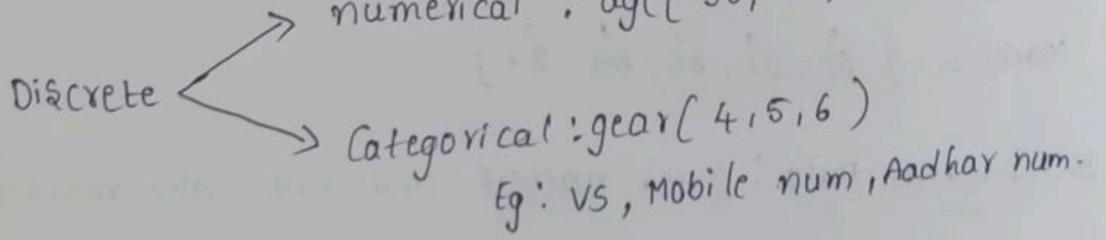
1. Quantitative (Numerical) variables.
2. Qualitative (Categorical) variables.

1. Quantitative variable :

A value can be measured and we can perform mathematical operation like [Add, Sub, Mul, Div).

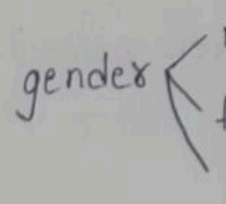
Ex: mpg, height, Weight,

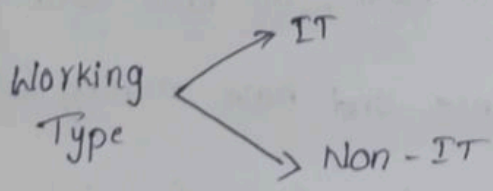


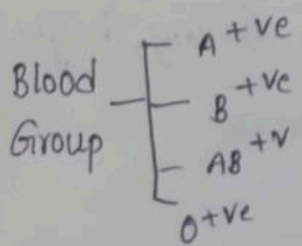


2. Qualitative variable :

Non-Measurable data and Based on some characteri-stics we can derive categorical variables.







Variable Measurement scales:

4 types of Measured variables.

1. Nominal Data: The categorical Data which are having different classes.

2. Ordinal Data: Orders of the data matters but value doesn't.

Marks1 = { 95 (3), 98 (2), 99 (1), 92 (4), 90 (5) } ranking

Marks 2 = { 80 (5), 82 (4), 85 (3), 89 (1), 87 (2) }

3. Interval Data: Order matters and value also matters but natural zero is not present.

Ex: Eye Sight, Temperature, Waves

4. Ratio Data: The ratio data can be measured, order, equi-distance and have mining ful zero (true zero point).

→ The value start from '0'.

Ex: Speed, Ratings, No. of Students, Salary, height, Weight age.

Descriptive Statistics :

1. Measure of Central tendency. (Mean, Median, Mode).

Mean :

1. population Mean (μ)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

2. Sample Mean ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n}$$

Median :

* Sort the value either Ascending or Descending order.
* Choose the mid value.
* If you get Mid 2 value take avg of those 2 values.

→ Mean will be affected by outliers where as Median won't affect by outliers.

→ We eg for null value Imputation. using Mean & Median.

Mode : Most repeatedly repetative values.

fx : [1, 2, 3, 1, 2, 2, 3, 4, 5]

More repeated value is 2.