Introduction part :

- check the quality of the data
  → descriptive Statistics
- To make a statement or a conclusion
  → Inferential Statistics

* what is statistics ?

It is the science of collecting, organising and analysing data (IS better decision making)

* what is data?

Data means facts and pieces of information that can be measured.

eg: the IQ of a class student.

{98, 97, 68, 57, 110}

avg IQ min, max

* Descriptive Statistics

It consists of organising and summarising data.

Inferential Statistics

Techniques where we used the data that we have measured to form conclusion

&

To make a statement / conclusion on a descriptive statistics we use inferential Statistics

eg question

Are the avg marks of java class student is same as python class student in bearal.

→ Inferential Statistics

what is the avg marks SQL student

→ Descriptive Statistics

population (N) and Sample (n)

The entire group of the data we call it as population.

eg All people in India

A subset of a population we call it as a sample

eg: I call people from different region of India

Keypoints :

- populations are larger than Samples
- Samples should be representative of the popul
- Sample allow for easier faster and less costly collection

Types of Sampling techniques

1. Simple random Sampling

every member of a population has an eq

Chance of being selected for our Sample

eg: Avg mileage of a bike.
Avg ratio of married people in bngl

2. Stratify Sampling
where the population is split into non overlapping group
eg: people is male or female

3. Systematic Sampling
from the population every nth sample we are going to select.
eg: while doing survey in the mall on the topic of moderisation, collecting information of every fifth person who is coming out from mall

Convinience Sampling
The Sample is collected based on our convinience from the particular domain experts.

le: Sampling technique selection always depends problem statement.

Variable : It is a property that can take any value

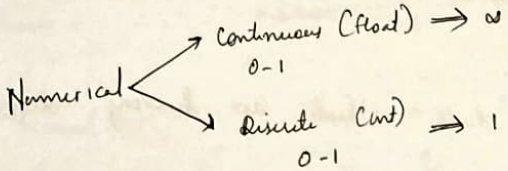
Two kinds of Variable

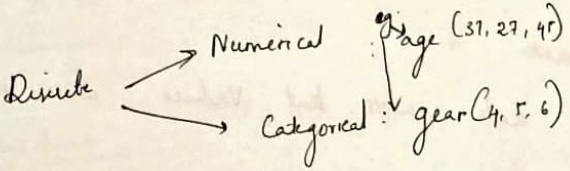
1) Quantitative (Numerical) Variables.
2) Qualitative (Categorical) Variable

① Quantitative Variable

A value can be measured and we can perform mathematical operation like ($\overset{+}{A}, \overset{-}{S}, \overset{\times}{M}, \overset{\div}{D}$)

eg: mpg, weight, height.
gear → discrete Categorical data.





2) Qualitative / Categ...

Non-measurable data and based on some characte...
we can devide categorical Variable.

eg: gender → male, female, other

working type → IT, non-IT

Blood group → A+, B+, O+, AB+

Variable measurement Scales

Types of measured Variables

1) Nominal data.

The categorical data which are having different classes.

2) Ordinal data.

... of the data matters but Values does not.

80	46	32	98	57
↓	↓	↓	↓	↓
2nd	4th	5th	1st	3rd

68	92	88	49	70
4th	↓ 1st	↓ 2nd	↓ 5th	↓ 3rd

3) Interval data

Order matters and Value also matters but natural zero is not present.

(that Value never reaches zero)

4) Ratio data

The ratio data can be measured order equidist... and have meaningful zero

Discriptive Statistics

1. measure of central tendency

mean median mode

population mean (μ)

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

Sample mean ($\bar{x}$)

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

Median

- Sort the Value either asce or desc
- Choose the mid Value
- if you get mid 2 Value take any of those

[1, 2, 2, 3, 4, 5] [1, 2, 2, 3, 4, 5, 100]

Mean: 2.8 Mean : 16.7

median: 2.5 median : 3

Mean will be affected by outlier where as median won't affect by outlier

for null Value Imputation :- (mean and median is used)

Mode

most repitative Value

Measure of Dispersion

eg:

	Person 1	person 2	
Mon	7.30 am	8 am	
tues	7.45 am	11 am	
wed	8 am	9 am	Variance high
	7.15 am	7 am	predictions low
	7 am	10 am.	
	? 7-8	? 9-10	

Variance

Standard deviation

nge

population Variance (σ^2)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample Variance (s^2)

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$n-1$ = degree of freedom

Calculate the Variance

{1, 2, 2, 3, 4, 5}

~~8.02~~

$$\sigma^2 = \frac{(1-2.8)^2 + (2-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 + (4-2.8)^2 + (}{6}$$

Va

$$\frac{3.36 + 0.69 + 0.69 + 0.02 + 1.36 + 4.69}{6} = \frac{10.83}{6} =$$

$\sqrt{\sigma^2} = \sqrt{1.8} = 1.34$

$\sigma = 1.34$ km

$\mu = 2.8$

Standard deviation = $\sqrt{\text{Variance}}$

population = SD

$\sigma = \sqrt{\sigma^2}$

$\sigma = \sqrt{1.8}$

$\sigma = 1.34$

Sample

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Range = max − min
= 5 − 1
= 4

Sample = SD

$S = \sqrt{s^2}$

$S = \sqrt{2.16}$

$S = 1.46$.

Percentile and Quartile.

Percentile is a value below which a certain observation will come under.

{1, 2, 3, 4, 5, 5, 6, 7, 7, 8}

How much % of data will come below value 6.

Percentile rank of x = (No of value below x / N) × 100.

$= \frac{7}{11} \times 100$

$= 63\%$ observation data is < 6.

• Quantile helps to find the value which is present at the given percentile rank.

{ 1, 1, 2, 3, 4, 5, 5, 6, 7, 7, 8 }.

which value is present at 25%?

$$\text{Value} = \frac{\text{percentile}}{100} \times (n+1)$$

$$\frac{25}{100} \times 12$$

$$= 3 \longrightarrow \text{Index}$$

$$= \text{Value} = 2$$

If it is 90%.

$$\frac{90}{100} \times 12$$

10.8 = Index

10 → Index

7 = Value.