Statistics

↳ The science from learning the data.

Descriptive Stats → summarizing data

Inferential Stats → methods for drawing conclusions about a population based on info from a sample.

Population	Sample
Entire Collection	Subcollection of members selected from a population.

Step I → Identify the question
↳ what are my hypothesis

Step II → Collect the data

Step III → Analyze the data

Step IV → Draw conclusions

Parameter Vs Statistics

↓ describes a population

describes a sample

Type of data you dealing

PAGE NO.
DATE

Quantitative Vs **Qualitative**

(Numerical data) (Categorical data)

(counting) (Labeling)

Discrete Data **Continous Data**

(certain values) (Any value within range)

Missing Value

↳ Delete

or ↳ input values (Replace)

Sampling Methods

i) Simple random sample

ii) Systematic sample

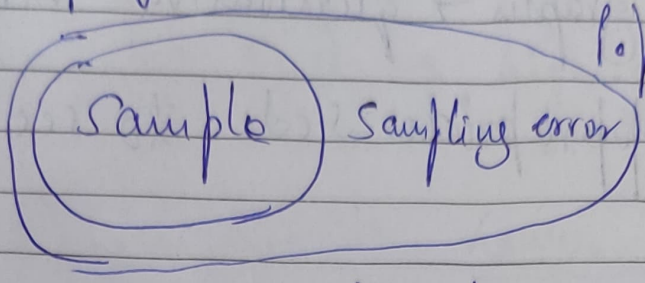
iii) Stratified sample

iv) ~~Custo mer~~ sample
Cluster

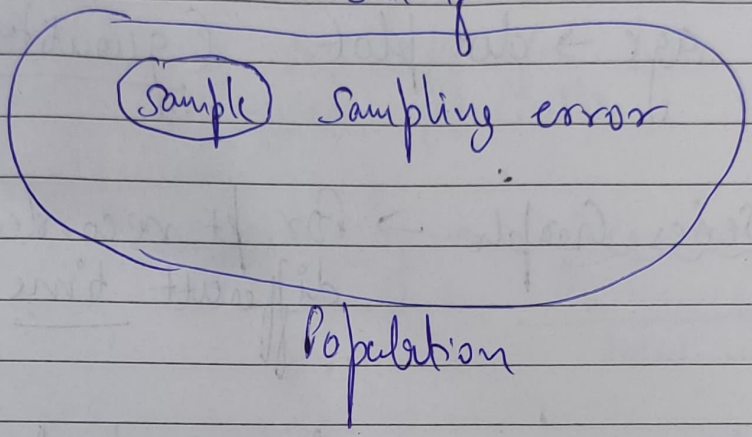
v) Convenience Sample

PAGE NO.
DATE

Sampling error



but if



Good Design of Experience

i) Randomization

ii) Replication (Repeat) → wide range

iii) Blinding (Placebo effects

Visualization

Use of bar graphs → qualitative variable

Choose appropriate scale to see volatility

Eg:- age → dot plot (quantitative var)

Time series Graph → for pt recorded at different time

Dot plots are good for comparisons

Histograms → can compare well but on same scale.

↳ frequency polygon

↳ histograms dot in middle joined

Classes

↳ Dividing data into classes (equal width)

PAGE NO.
DATE

Lower class limit → Smallest value in each class

Upper class limit → highest value in each class

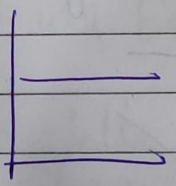
class mid point → $\frac{\text{lower} + \text{upper}}{2}$

class width → diff b/w two consecutive lower limits.

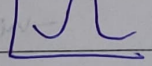
Relative frequency → $\frac{\text{frequency for specific class}}{\text{total no of data pts}}$

↳ the column sums up 100

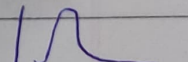
1) Uniform Distribution → Same frequency

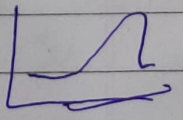
↳ 

2) Normal distribution

↳ 

3) Skewed Distributions

Right distribution → Right foot → 

Left distribution → Left foot → 

PAGE NO.
DATE

Paired Data → Related Data

(Height & Weight)

⇒ Correlation

↳ does not imply causation

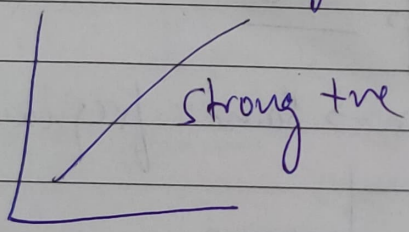
∨

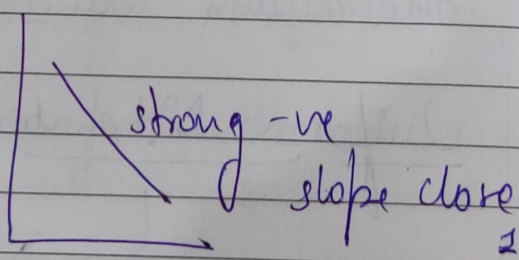
Lurking variable

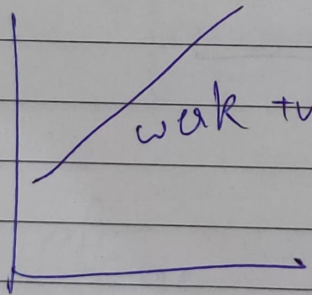
↳ indirectly involved

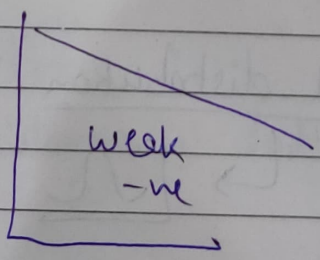
Linear correlation → straight line relationship

regression line









PAGE NO.
DATE

Center of your data

Resistance → a measure of center is resistant if the presence of extreme values doesn't cause it to change very much.

Mean → Avg of data → $\frac{\text{Sum of all values}}{\text{no of total values}}$

Sample → $\bar{x} = \frac{\sum x}{n}$ → sample

Population Perimeter → $\mu = \frac{\sum x}{N}$ → population

So, mean is not resistant

↳ coz any outlier is there, it will be added and hence it will be changed.

Median → Value that falls in exact middle of data set → when in ascending order

So median is resistant to extreme values

Mode :- Greatest frequency of data

The mode is resistant to extreme values bcoz outliers would never be in mode

Symmetric data → Mean ~ Median

Measure of Spread

Range → Diff b/w largest & smallest
↳ Not resistant

Standard Deviation :- measure of how much individual values deviate away from mean.

Sample ⇒ $s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$ → like range 1 to n

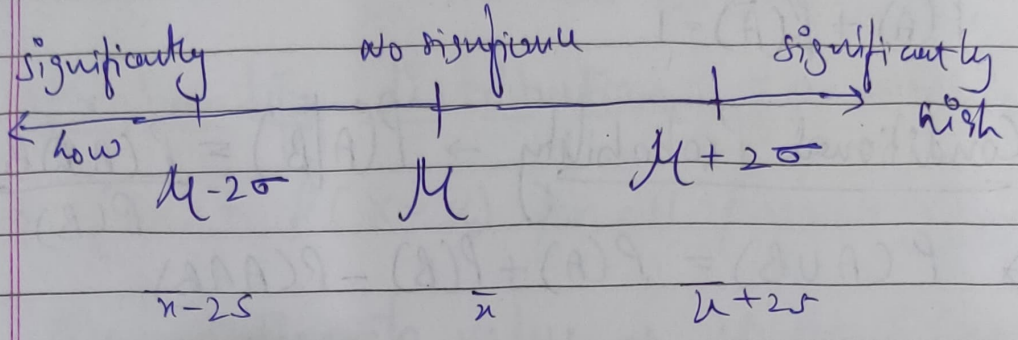
Population → $\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$

Degree of freedom

Prop of SD

i) Never be negative
ii) 0 only if all data same
iii) large values → more variance & wider spread
iv) Not resistant → will have extreme values

Variance = $(SD)^2$



Measure of location

Percentile → divide in 100 groups → 1% in each

Percentile of x = $\frac{\text{number of values less than } x}{\text{total no of values}} \times 100$

Quartiles → 25% of gap

Q_1 → 25%.
Q_2 → 50% → Median
Q_3 → 75%.

where 0% is minimum & 100% in max

IQR = $Q_3 - Q_1$

Outlier → Above $Q_3 + (1.5 \times IQR)$
Below $Q_1 - (1.5 \times IQR)$

PAGE NO.
DATE

Probability

$P(A) + P(\bar{A}) = 1$

Conditional Probability → $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Mutually exclusive event → cannot happen at same time

$P(A \cap B) = 0$

Sampling with replacement Vs **without**

↓ independent ↓ not independent

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|(A \cap B))$

Contingency table → (Venn diagrams)

↳ Group histograms table

PAGE NO.
DATE

Discrete random Variable

↳ that take countable values

Probability distribution of x → table $x = 0, 1, \ldots$

PMF → $f(x) = P(X = x)$ for all possible value of x

Mean discrete value → μ or $E(x)$

$\Downarrow$

$\sum x \cdot f(x)$ → pmf of x

Variance of Discrete random Variable

↳ $Var[X] = \sigma^2 = E[X^2] - (E[X])^2$

$= E[X^2] - \mu^2$

$\downarrow$

$\sum x^2 \cdot f(x)$

S.D of Discrete random Var

$SD[X] = \sigma = \sqrt{E[X^2] - (E[X])^2}$

Bernoulli Experiment → Either Success or fail

↳ no → sum of all success

Binomial distribution

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$x = 0, 1, \ldots n$

Mean $= np$

Var $= np(1-p)$