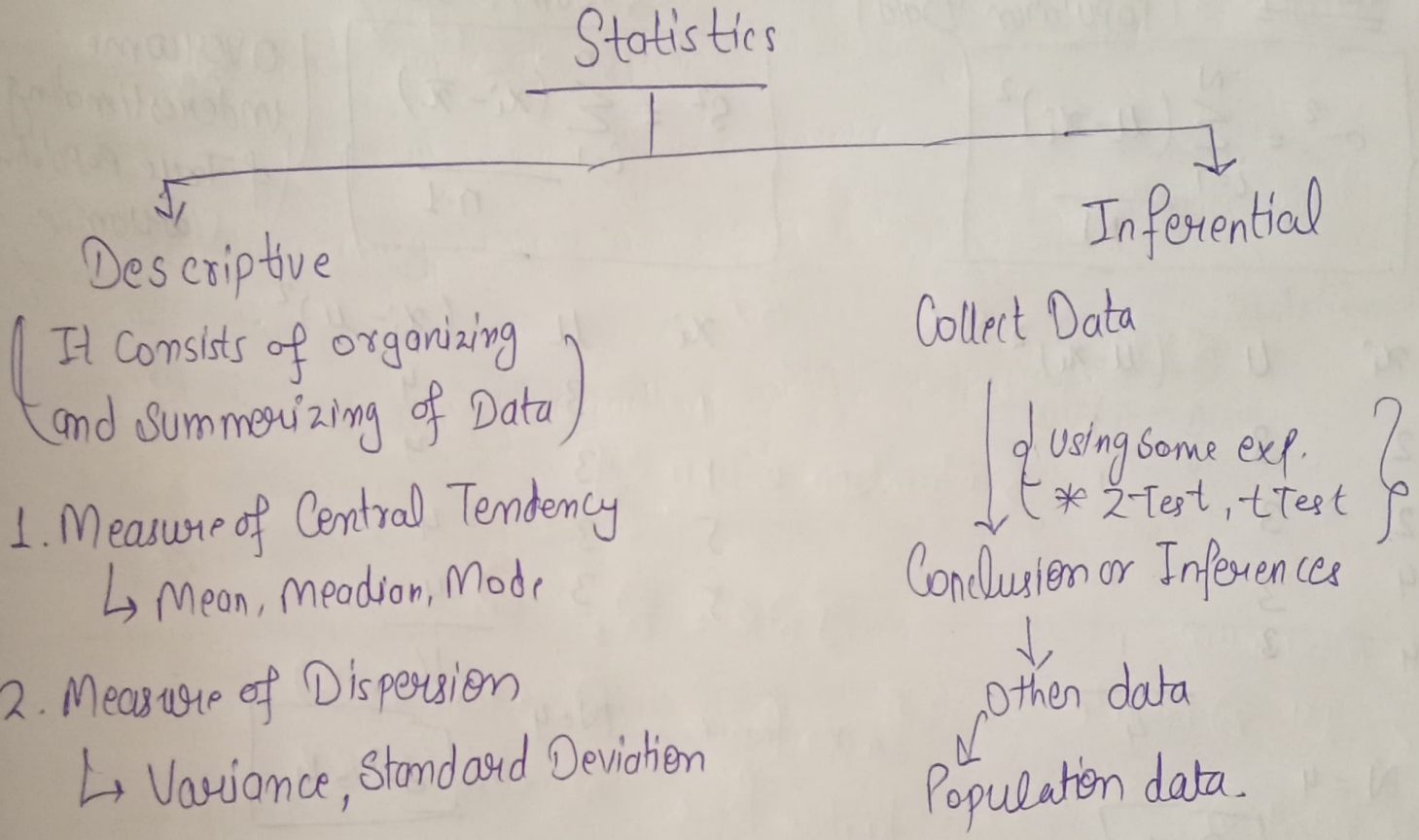
Statistics

=> Statistics is a field that deals with collection, organization, analysis, interpretation and presentation of data

⇓

Decision making



Population = N

Sample = n.

Measure of Central Tendancy

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

Measure of Dispersion

Age1 = {2, 2, 4, 4} Age2 = {1, 1, 5, 5}

$\mu = 3$ $\mu = 3$. {Spred is more}

1. Variance

(Population Data)

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

x_i	μ	$(\mu - x_i)^2$
2	3	1
2	3	1
4	3	1
4	3	1
		4

$N = 4$

$\sigma^2 = 1$

(Sample Data)

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

Overcome underestimating of True population variance

x_i	μ	$(x_i - \mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4
		16

$N = 4$

$\sigma^2 = 4$

Bessel's Correction

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

if divided by n then

$\bar{x} <<< \mu$ underestimate the true

$s^2 <<< \sigma^2$ population variance

Standard Deviation

-> ~~Variance~~ denotes the distance of a data point from Mean.

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

Variable (entity which can have any value)

- Quantitative
  - Discrete (Whole No)
    - No. of students
    - No. of Bank accounts
  - Continuous (Decimal)
    - Age
    - Weight.
- Qualitative / categorical
  - Gender :- Male / Female
  - Color
    - Red
    - Blue
    - Green.

Random Variable

$X \rightarrow$ function $\rightarrow$ value

Process / Experiments $(y = 5x + 2)$

$X = \{ \begin{matrix} 0 & H \\ 1 & T \end{matrix} \}$ Tossing of Coin

Histograms

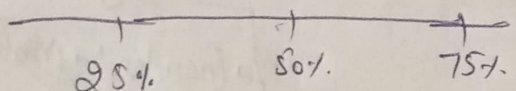
⇒ Percentiles :→ is a value below which certain percentage of observation lies.

$$\text{Percentile of } x = \frac{\text{no. of values below } x}{n} \times 100$$

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

⇒ Quartiles

25% → 1st Quartile Q_1
50% → 2nd " Q_2
75% → 3rd " Q_3



5 Number Summary

1. Minimum
2. 1st Quartile
3. Median
4. 3rd Quartile
5. Maximum

$$IQR = Q_3 - Q_1$$

$$\text{lower fence} = Q_1 - 1.5(IQR)$$

$$\text{higher fence} = Q_3 + 1.5(IQR)$$

Covariance & Correlation (C&C)

C&C are two statistical measure used to determine the relationship b/w two variables. Both are used to understand how changes in one variable are associated with changes in another variable.

⇒ **Covariance:→** is a measure how much two random variables change together. If the variable tends to inc. and dec. together, the covariance is +ve. If one tends to inc. when the other dec. the covariance is -ve.

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Cov(x,x) = Var(x) = \sigma^2$$

→ Correlation → Pearson Correlation Coefficient
→ Spearman Rank Correlation

① **Pearson Correlation Coefficient** → [-1 to 1]

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \, \sigma_y}$$

(a) The more the value towards +1 the more +ve correlated x & y
(b) " " " " -1 " " -ve " "

② **Spearman Rank Correlation**

$\rho = -1$, $\rho = -1$ to 0, $\rho = 0$, $0 < \rho < 1$, $\rho = +1$.

Non-linear data

Probability

⇒ Mutual Exclusive Event :→ two events cann't occur same time.

→ P(H or T) = P(H) + P(T) {Additive Rule for MEE}

⇒ Non-mutual exclusive event :↓

* Taking out card from deck

P(k or ♡) = P(k) + P(♡) - P(k and ♡)

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} \frac{4}{13}$

⇒ Multiplication Rule

⇒ Independent Event :→ 2 events don't affect each other.

e.g. P(H and T) = P(H) * P(T)

$= \frac{1}{2} * \frac{1}{2} = 1/4$

⇒ Dependent Event :↓

* Take a King from the deck and then the Queen card from the deck.

P(K and Q) = P(k) * P(Q/k) → Conditional Probability