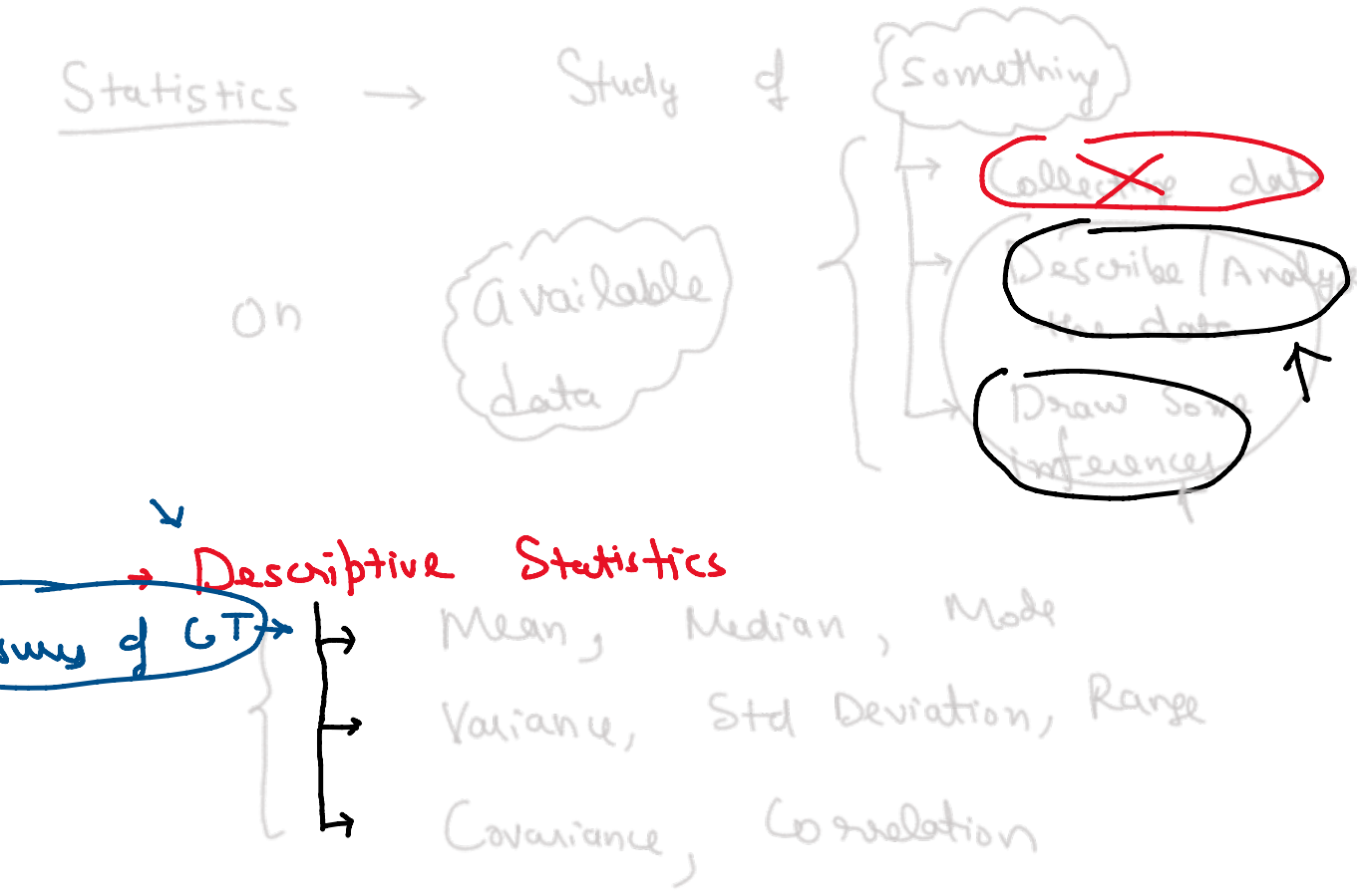
Statistics

08 July 2020 03:51 PM

Starting @ 4:07 PM

Topics]

→ Statistics



→ Discrete → countable set

→ Continuous → uncountable set

1. Measure's of Central Tendency

1. Measures of Central Tendency

1. Mean (Average)

Height (cm)	Weight (kg)	Gender
160	50	Female
180	78	Male
146	45	Female
162	51	Female
184	80	Male
180	60	Male

H, W, G

outlier → 1800

$\mu_H = 168.67$

$\mu_W = 60.66$

NOTE: Assume that all the students in the class come from same geographical region and fall under same age group

$$\mu_H = \frac{h_1 + h_2 + h_3 + \cdots + h_6}{6}$$ → Total no. of obs

Sum

$$\mu_H = \frac{\sum_{i=1}^{n} h_i}{n}$$

160, 180, 146, 162, 184, 180 → 168.67 ✓ → Central Tendency

Google form ✓

168.67 (without error)

438.67 cm (with error)

1 error

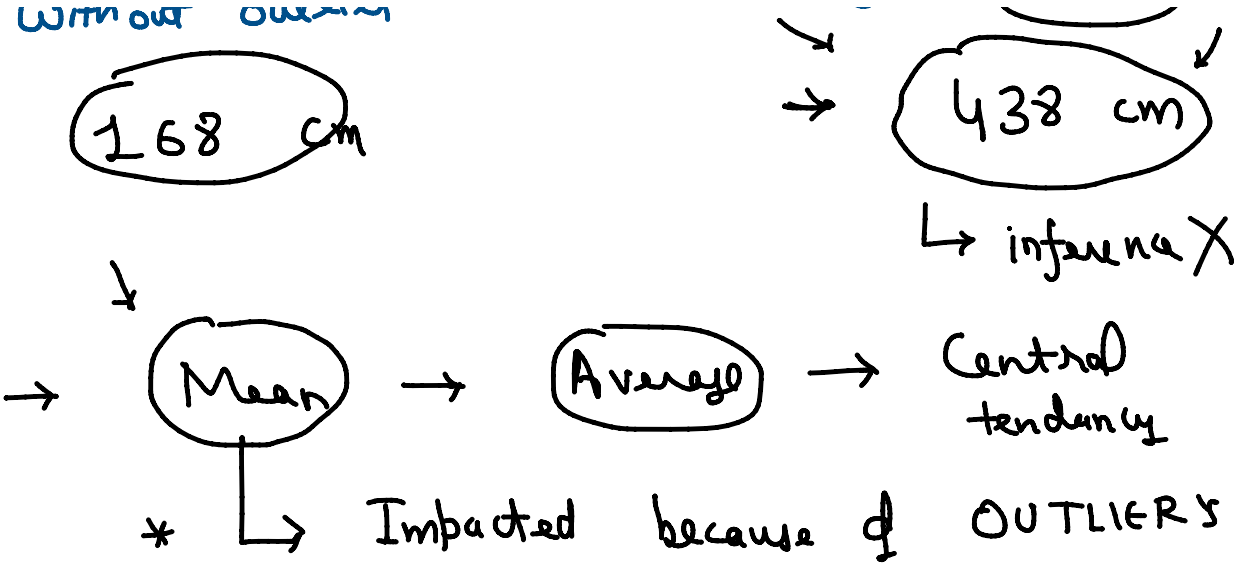
H (cm) → 5.6

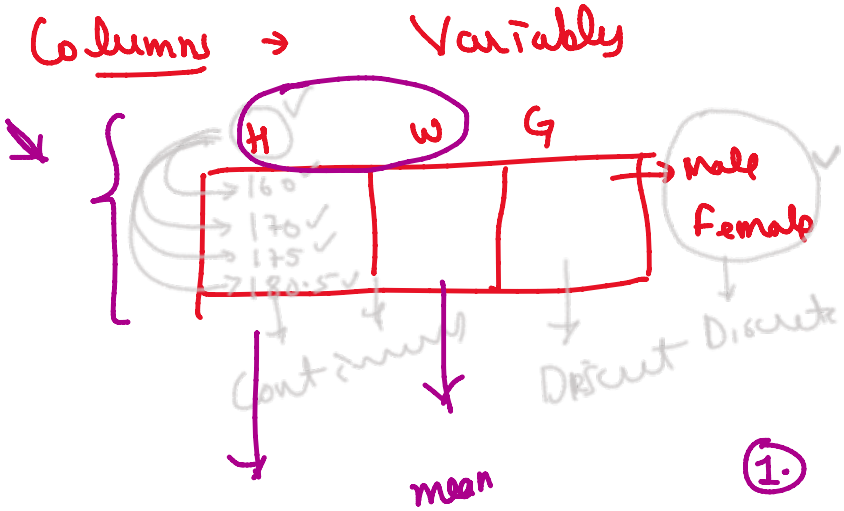
W → 18

→ outlier → Extremely small or a very large value

With out outlier

With outlier





1. 1 variable at time
2. collection

Univariate Analysis → Mean

one variable

Height (cm)	Weight (kg)	Gender
160	50	Female
180	78	Male
146	45	Female
162	51	Female
184	80	Male
180	60	Male

NOTE: Assume that all the students in the class come from same geographical region and fall under same age group

→ 3 Variables

→ H, W → Continuous Variable

→ G → Discrete/ Categorical

Mean → Univariate analysis
Measure of central tendancy

$$\mu_H = \frac{\sum_{i=1}^{n} h_i}{n}$$

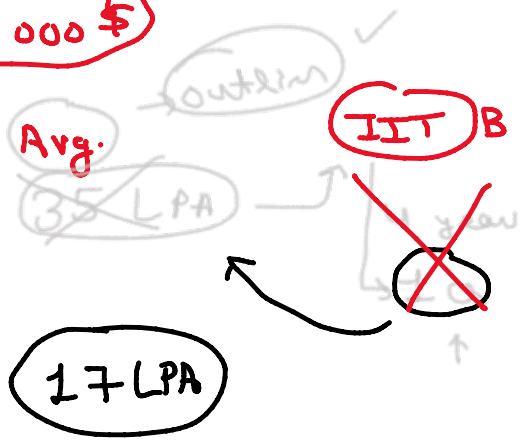
$\mu_H = \frac{\sum_{i=1}^{n} \cdots}{n}$

Impacted in the presence of outliers

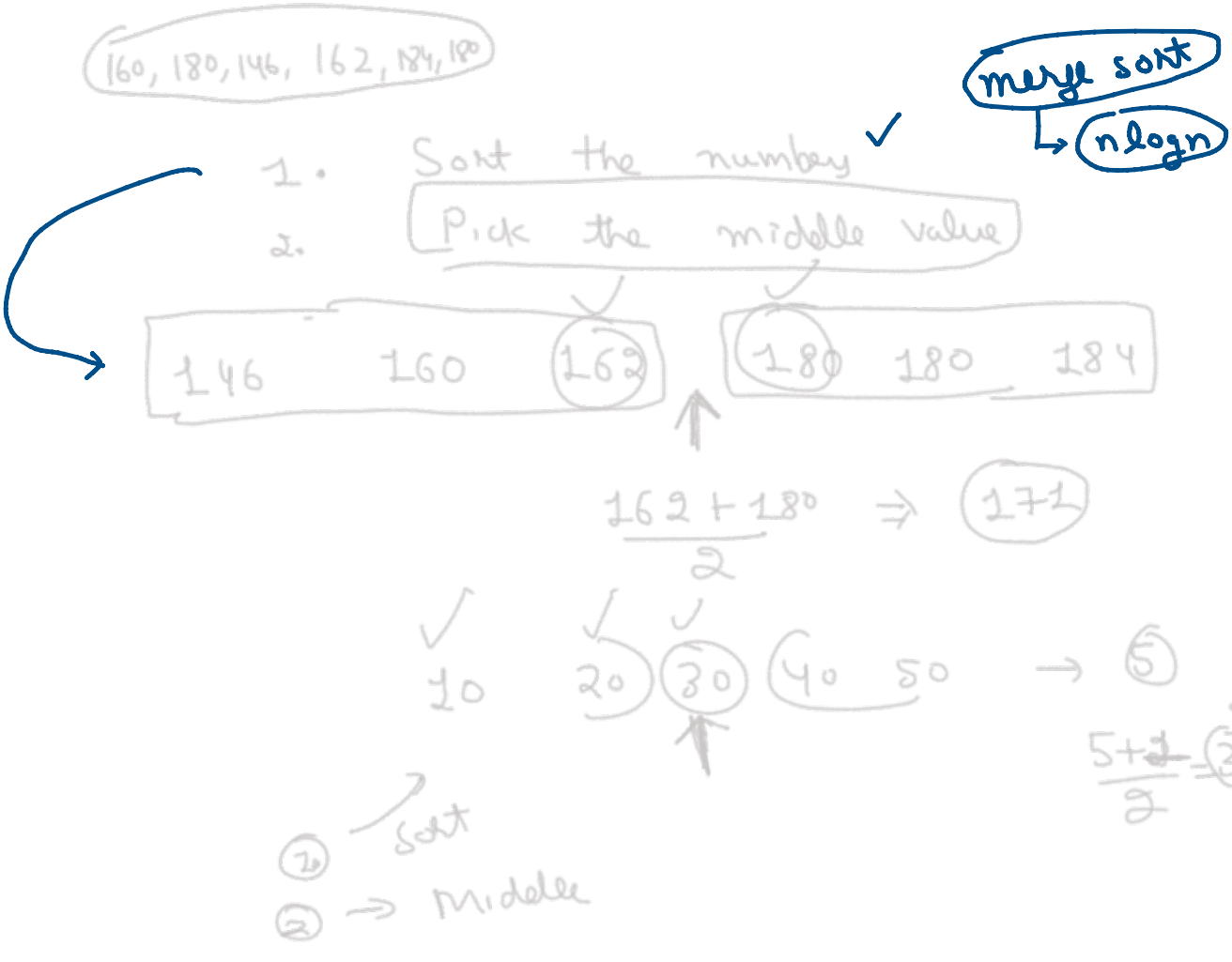
↳ Solution → Drop the outliers ✓

Americans → 50,000 $

Admission →



1. Mean
2. Median ↴



1. Sort the values in increasing order

2. If no. of observation is

Even → median = $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$

Odd → median = $\left(\frac{n+1}{2}\right)^{th}$ value

→ Measure of central Tend ↓

1. Mean → 3·5 → O(n)

2. Median → 1. Sorting → O(nlogn)
2. Middle → O(1)
O(nlogn)

Measure Central Tendency ↓

160, 180, 146, 162, 184, 1800

[146 160 162 180 184 1800]

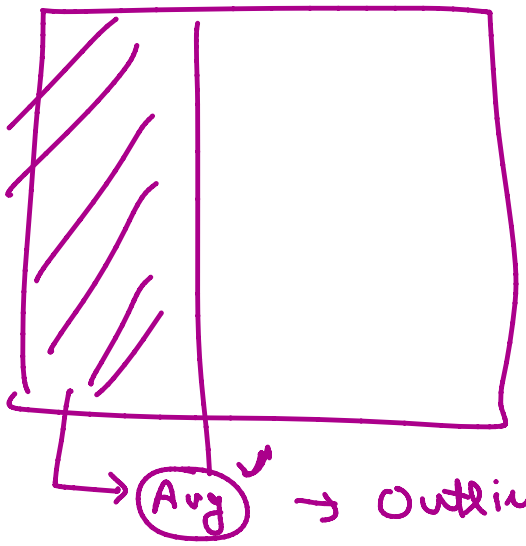
Mean → 168 cm (if no outlier)

438 cm (outliers)

Median → 171 cm (no outliers)

171 cm (outliers)

Time ↑



→ Median Package

C T

1. Mean
2. Median
3. Mode → Maximum freq. value.

1 2 3 1 2 3 1 2

→ 1, 2 → Bimodal

Shopkeeper → Garments → Shopping

Male Female

Wholesaler

25 S 25 L
75 XXL 75 XL
25 M

50 S 50 M
50 L 50 XL
50 XXL

5:18 to 5:23 BREAK

Statistics

↳ Measure of Central Tend

Univariate analysis

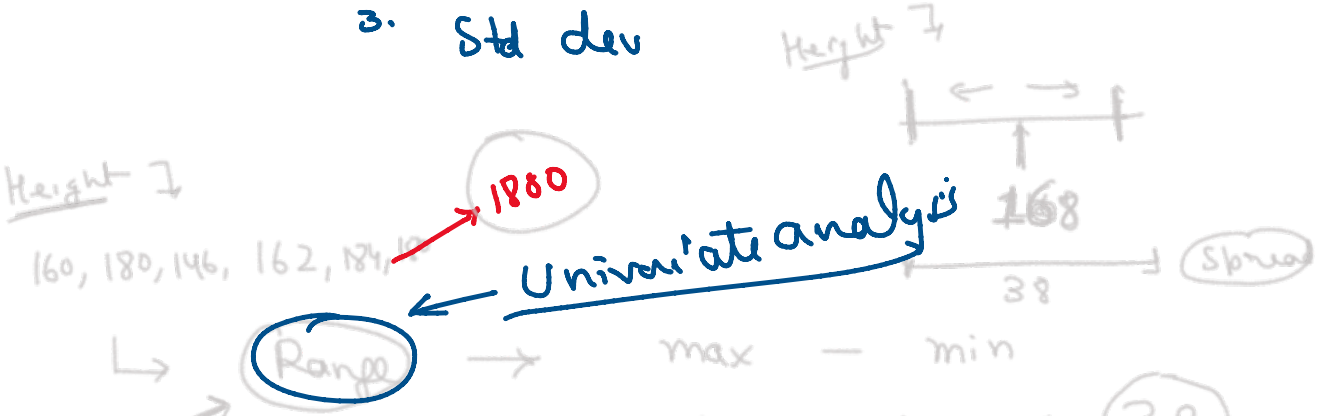
1. Mean → Continuous columns
2. Median
3. Mode → discrete data/ Column

Mean → Pick if no outliers exists. Fast to compute

Median → Pick if outliers exists
More time complexity

→ Measure of Spread

1. Range
2. Variance
3. Std dev



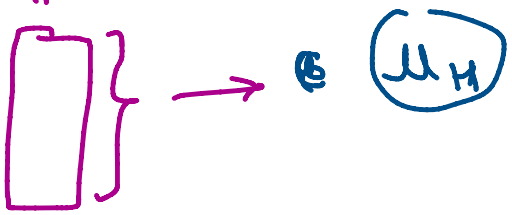
Univariate

2. Variance

→ Variance measures how far a dataset is spreaded out around the central

is spreaded out around the central tendency

H

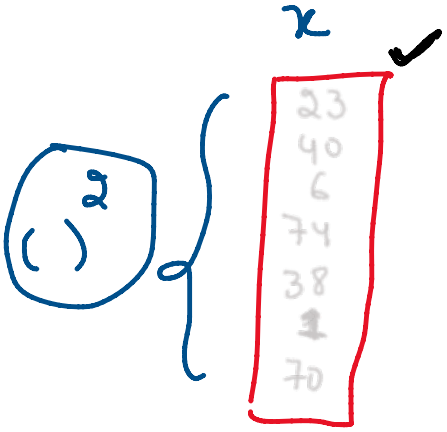


$\mu_H = 160$

$\sigma^2 = 0$

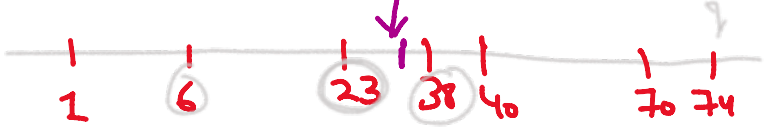
→ { No variability or Zero variance if all the values are same }

x



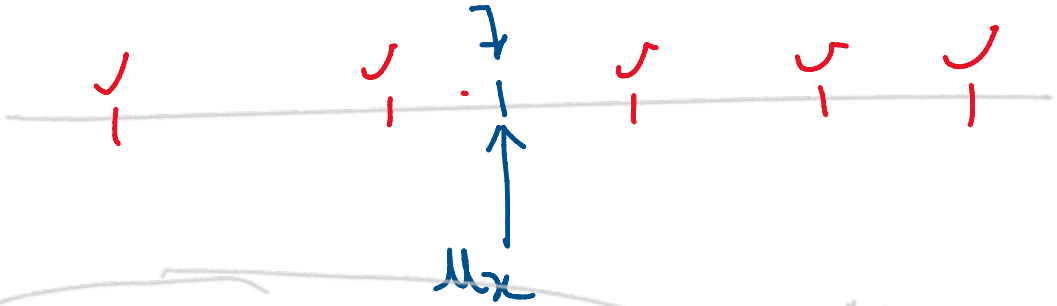
$\mu_x = 36$

36 → μ_x ←



$(36-1)^2 + (36-6)^2 + (36-23)^2 + (36-38)^2 + (36-40)^2 + (36-70)^2 + (36-74)^2$

→ Measure how far dataset is spreaded out around mean.



$$\sigma^2 = \sum_{i=1}^{n} (\mu_x - x_i)^2$$

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

$\mu_x = 36$, $\sigma^2 = 702$

$\sqrt{\sigma^2} = \sqrt{702}$

Std. dev → $\sigma = 26.49$

→ Measure of central Tendency
- Mean (impacted by outliers)
- Median
- Mode

→ Measure of spread
- Range
- Variance σ^2 } (impacted by outliers)
- Std. deviation σ }

} Univariate Analysis statistical tarike

↳ $$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n} (\mu_x - x_i)^2$$

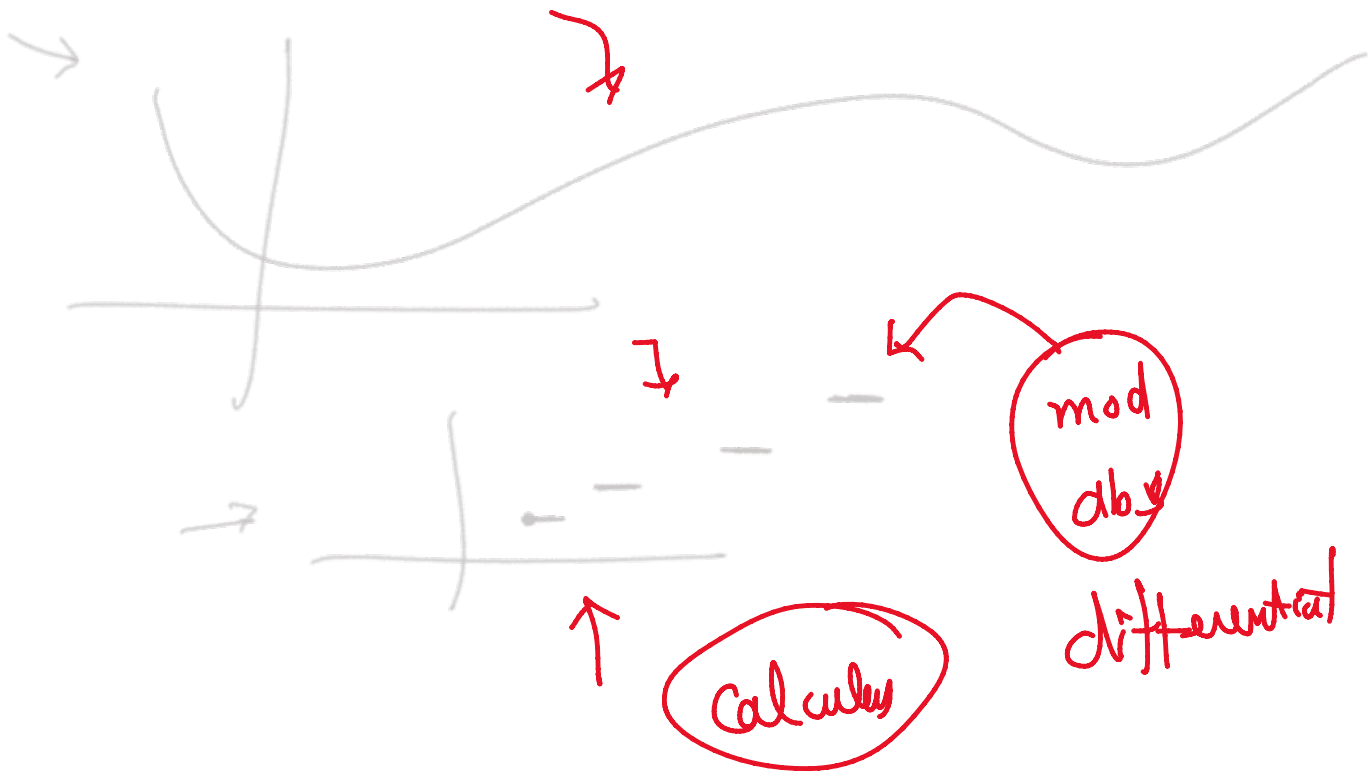
+ve OR -ve

Variance → Average squared Distance

Variance → Average squared Distance of each point from mean

Std dev

Continuous & Discontinuous func



Starting @ 4:05 PM

TOPICS

→ Statistics (Part-2)

1. Measure of Central Tendency
   - Mean (outliers impact)
   - Median ✓ (Time complexity)
   - Mode ✓

2. Measure of Spread
   - Range
   - Variance
   - Std. dev
   - IQR ✓ (Time complexity)

(outliers impact)

Univariate Analysis

Qus: Suppose

45/100

Vgood
Good
Bad ✓
VBad ✓

99th percentile

{CAT GATE}

OR

45/100 → 45%

99th percentile

99% of the total population scored less than 45 marks

100
50

100

990

10

Percentile tells us the rank

90th percentile

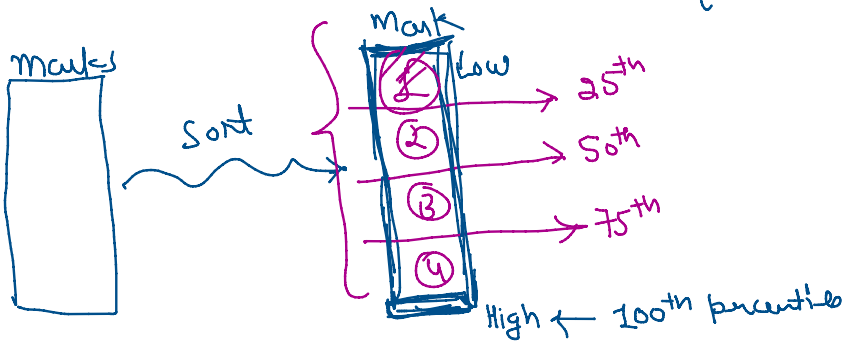
90% of the total population scored less than

Frequently used Percentiles

Divides the data into 4

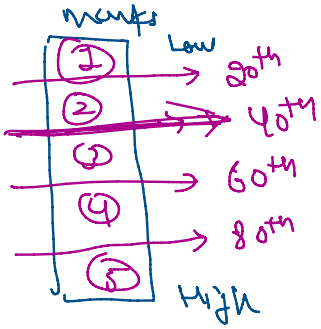
frequently used percentiles

→ **Quartile** → Divides the data into 4 equal parts



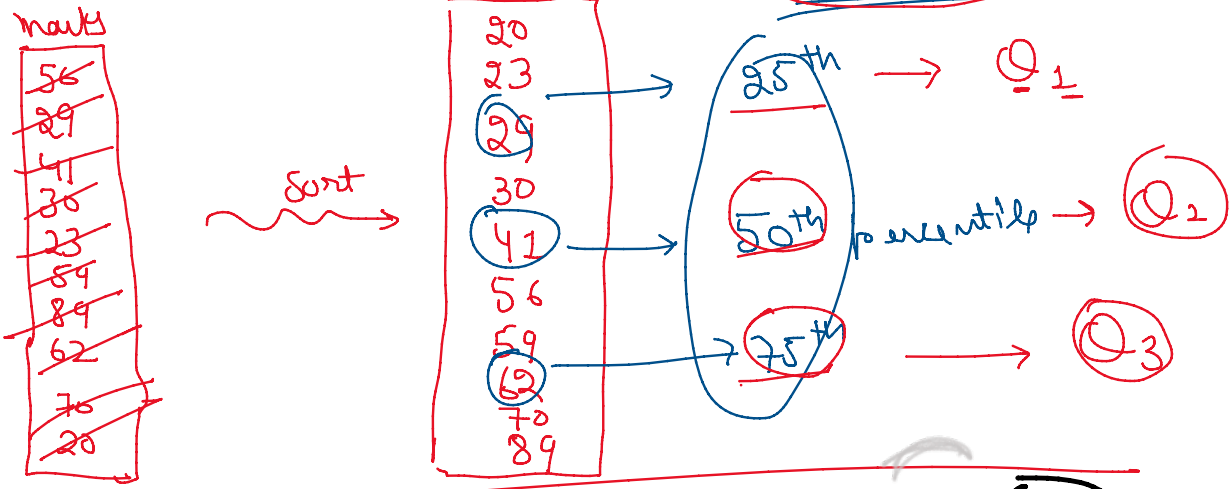
Quartile
↳ 25th, 50th, 75th

→ **Quintiles** → Divides the data into 5 equal parts



Quintiles
20th, 40th, 60th, 80th

Quartiles

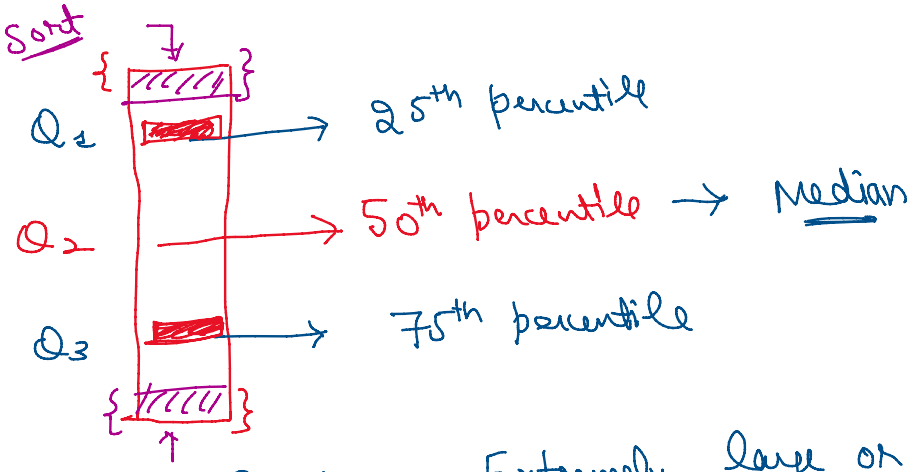


→ IQR ~ (Interquartile Range)

IQR = 75th value − 25th value

IQR = $Q_3 - Q_1$

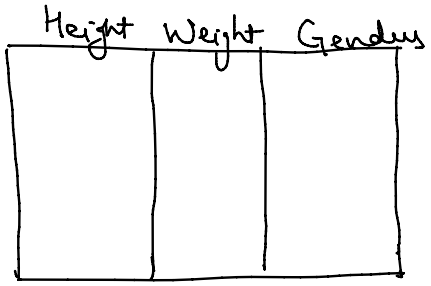
$$IQR = Q_3 - Q_1$$



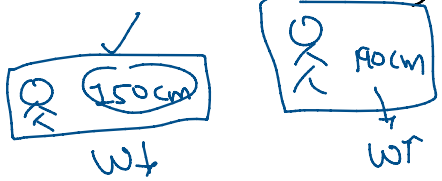
Outlier → Extremely large or small value

Univariate Analysis → One variable at a time

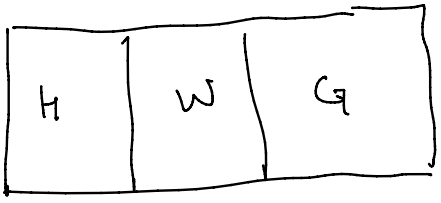
Bivariate Analysis → Two " " " "



1. ——
2. ——
3. Measures of Relationship (Bivariate Analysis)
   - Co variance
   - Correlation

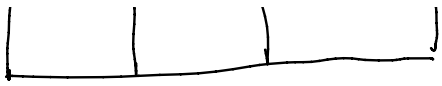


Co variance



H↑ ⇒ wt ↑ or wt

H↓ ⇒ wt ↑ or wt



H↑ ⇒ W↑ or W↓

H↓ ⇒ W↓ ✓

H↑ ⇒ W↑ ✓

H ∝ W ✓

{ Cov(H, W) = +ve → H ∝ W
OR −ve → H ∝ 1/W }

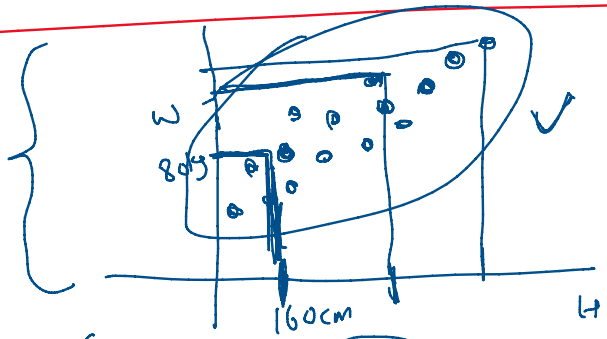
Co variance

Variance → $\frac{1}{n}\sum\{(\mu_H - h_i)(\mu_H - h_i)\}$

$\text{Cov}(H,W) \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(\mu_H - h_i) * (\mu_W - w_i)$

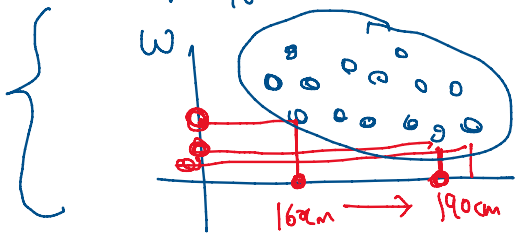
Cov(H,W) = +ve value ⇒ H↑ ⇒ W↑

Cov(H,W) = −ve value ⇒ H↑ ⇒ W↓



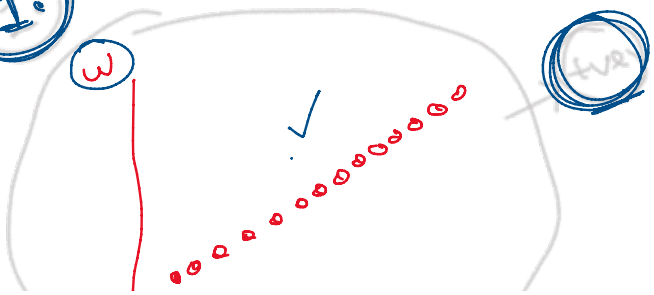
H↑ ⇒ W↑

Cov(H,W) = +ve value

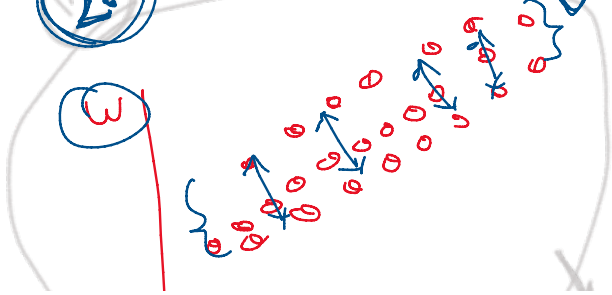


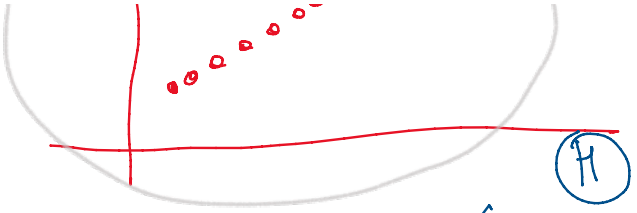
H↑ ⇒ W↓

1.



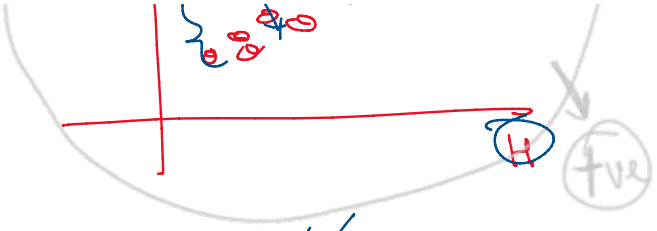
2. Cov(H,W) = −ve value





H ✓

✓ H↑ ⇒ W↑ ✓



+ve ✓

✓ H↑ ⇒ W↑

→ CORRELATION → Pearson Correlation ✓

→ ~~Spearman Rank correlation~~

Pearson Coeff → $\rho_{H,W}$

$$\{-1 \leq \rho_{H,W} \leq +1\}$$

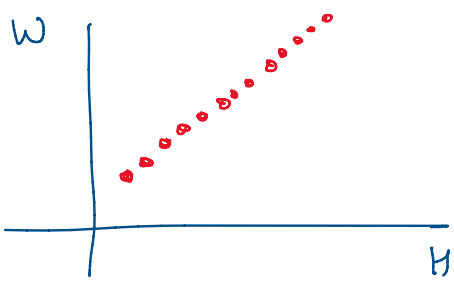
$$\rho_{H,W} = \frac{COV(H,W)}{\sigma_H * \sigma_W}$$

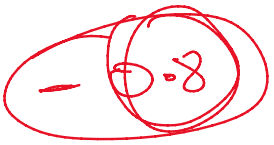
Pearson → ✓ H↑ ⇒ W↑ OR W↓ ← Covariance

→ ✓ Dependency ✓

$\rho_{H,W} = \oplus 1$ ✓ ✓

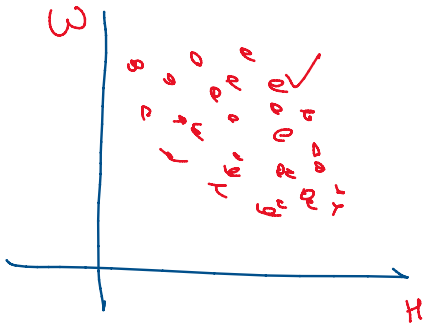
$\oplus$ → $H \propto W$ ρ_H

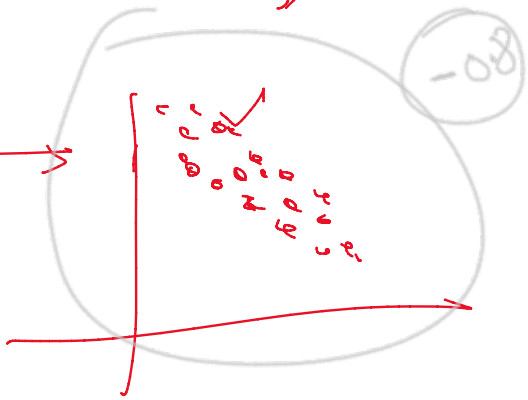


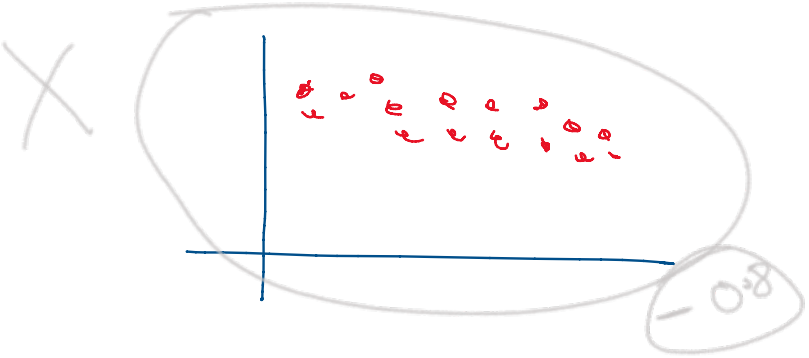


$\rho_{H,w}$ = 0.5

$H\uparrow \Rightarrow w\uparrow$







05:03 to 05:08 → 5 min

1. Measures of Central Tendency (outlier impact)
   - Mean
   - Median
   - Mode

2. Measures of Spread → (outlier impact)
   - Range
   - Var & Std
   - IQR

} Univariate Statistical Analysis

3. Measures of Relationships (Bivariate Analysis)
   - Covariance
   - Correlation (Pearson) (ρ)

↳ Covariance

↳ Correlation (Pearson) (ρ)

↳ Dependent

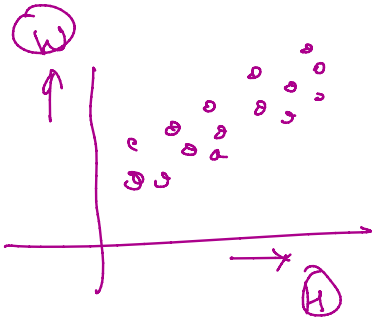
$x\uparrow \Rightarrow y(?) \Rightarrow \uparrow\downarrow$ ✓ ✓

$\rho_{H,W}$ — height — weight — np.corrcoef(height, weight)

height 1.0 0.88

weight 0.88 1.0

$+0.88$



$$-1 \leq \rho_{H,W} \leq +1$$

$\rho_{H,W} = 0 \Rightarrow$ No relationship b/w H, W

$\rho_{H,W} = \oplus 1 \Rightarrow H\uparrow \Rightarrow W\uparrow$ && H, W are totally dependent

$\rho_{H,W} = \ominus 1 \Rightarrow H\uparrow \Rightarrow W\downarrow$ && " "

$$\rho_{H,W} = \frac{Cov(H,W)}{\sigma_H * \sigma_W}$$

$\rightarrow 1 \rightarrow$ Numpy $\rightarrow$ Opensource free