Descriptive Statistics

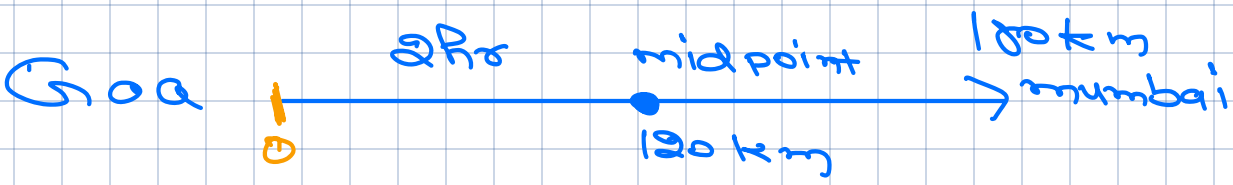
1. Mean
2. Median
3. Mode
4. Range
5. Variance
6. SD
7. weighted mean
8. Random Variable

Descriptive Statistics

- Describe
  - describing the data

Inferential Statistics

- Inference/predict



Describing Data Observed so far

- Avg speed so far → 60km/h
- 120km dist Traveled

Inference

- If we continue at same avg speed
- +1 hr to reach mumbai
- Total 3hr Goa → mumbai

Descriptive Stats

- Measure of Central Tendency
- measure of Variability

Measure of Central Tendency

① Mean

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

② Median

- $n/2$ if n is odd
- $\left(\frac{n}{2} + \frac{n}{2}+1\right)/2$ if n is even

30 30 35 40 40

- mean and median were reasonable measure of central tendency

30 30 35 40 40 300

- Outlier impact mean by a large Quantity
- Median is preferred for DS with outlier

Loss Function

1. MSE → closer to mean
2. MAE → Outliers closer to mean

50 60 30 10 100 120 15

- Sort → Find mid point (n log n)
- Calculation of median will be resource intensive

③ Mode

→ Most common Datapoint

[30 90 80 70 60]

→ Mode → 90

[30 90 80 70 60 60]

→ 90 , 60

→ Bi-modal Data

Q. i2

→ n = 4

→ [Q0 Q2 Q8 , ?]

→ mean → Q4

→ $\frac{Q_0 + Q_2 + Q_8 + x}{4}$ → Q4

→ x → Q6

[Q0 Q2 Q6 Q8] → $\frac{Q_2 + Q_6}{2} \to Q_4$

$n/2$ $n/2+1$

Weighted Mean / Avg

Subject	Credit	Grade (x_i)
A	4	4
C	3	5
D	2	3
B	3	4

Simple Avg = $\frac{(4+5+3+4)}{4}$

GPA = $\frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i}$

$\frac{4 \times 4 + 3*5 + 2*3 + 3*4}{4+3+2+3}$

w_i = weights → credits

x_i = Grades

Expected Value / Weighted Avg

Avg Pets → Pet Count → X

30% → 0 pets

40% → 1 pet

10% → 2 pets

20 → 3 pets

$$\frac{0.30 * 0 + 0.40 \times 1 + 0.10 * 2 + 0.2 *}{0.3 + 0.4 + 0.1 + 0.2}$$

1.20

x and y

$\frac{(x+y)}{2} = 40 \Rightarrow x+y = 80$

z

$\frac{(x+y+z)}{3} = 45$

$x+y+z = 45 \times 3$

$z = 45 \times 3 - 80$

Measure of Variability

↓

Spread

① Range = Maximum − minimum

30 35 35 40 300

$= 300 - 30 = 270$

Is Range impacted by Outlier?

⇒ Yes

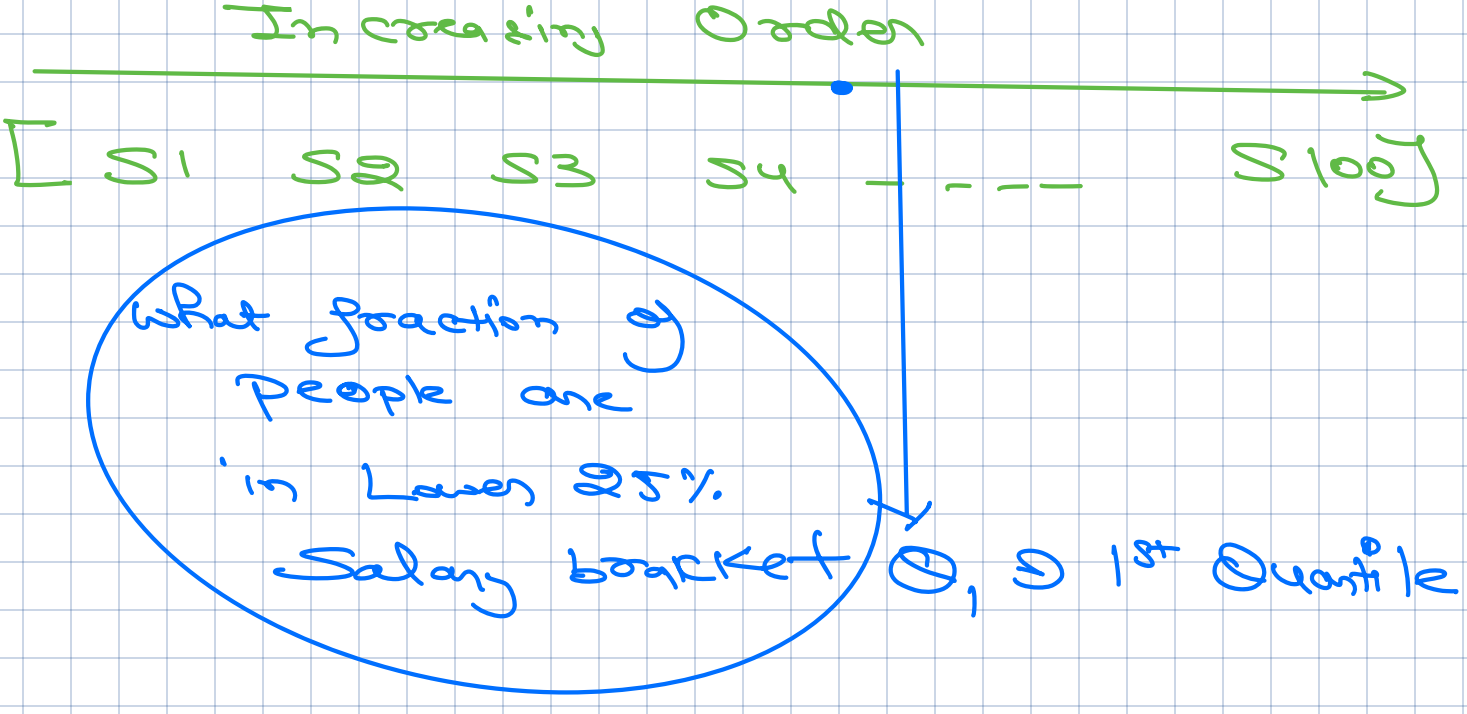
Any measure of spread that is Not Impacted by Outliers

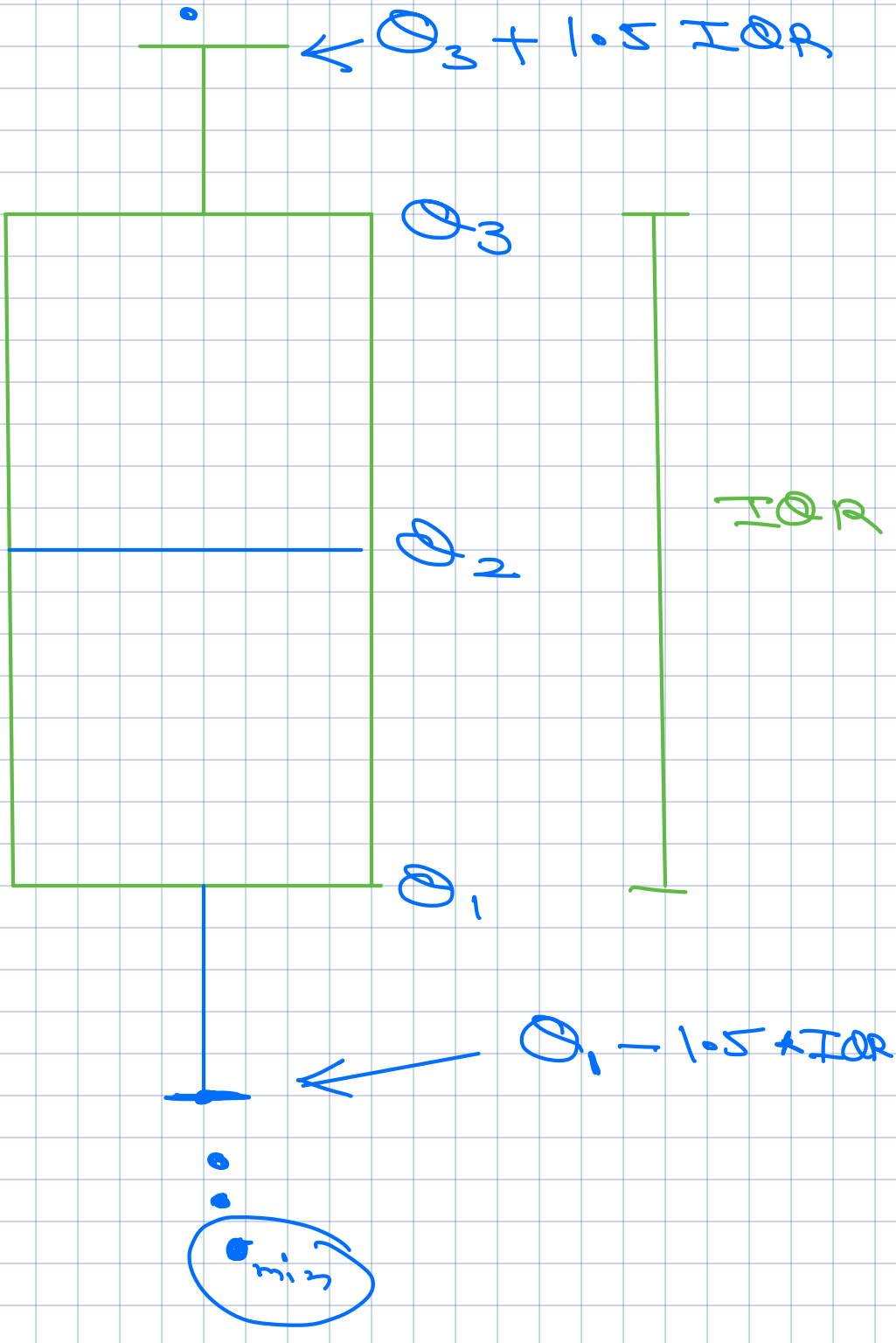
IQR ⇒ Inter Quartile Range

$$IQR \Rightarrow Q_3 - Q_1$$

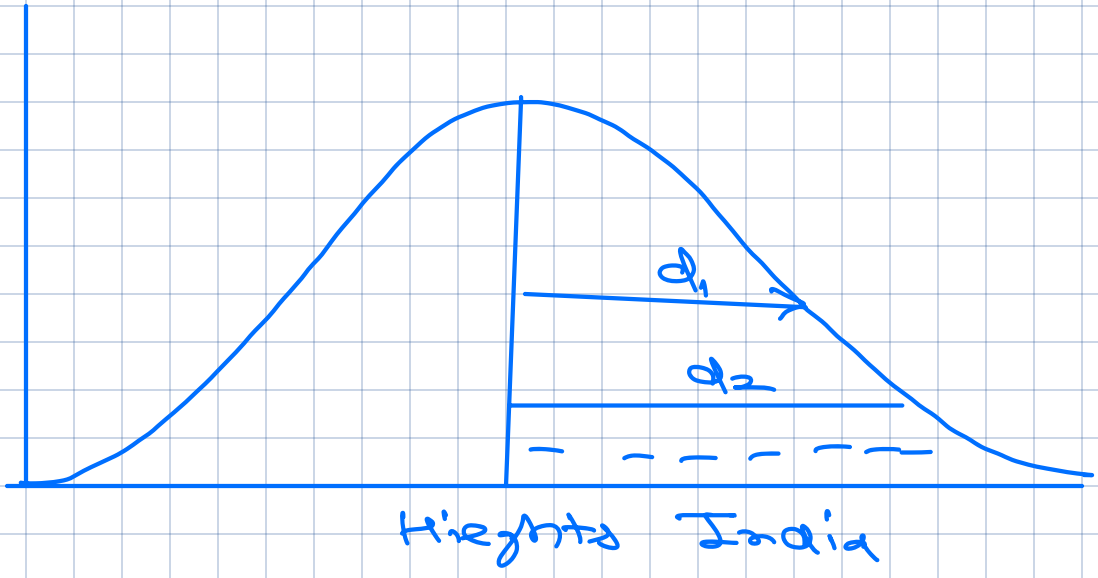
Q_3 → 75th %

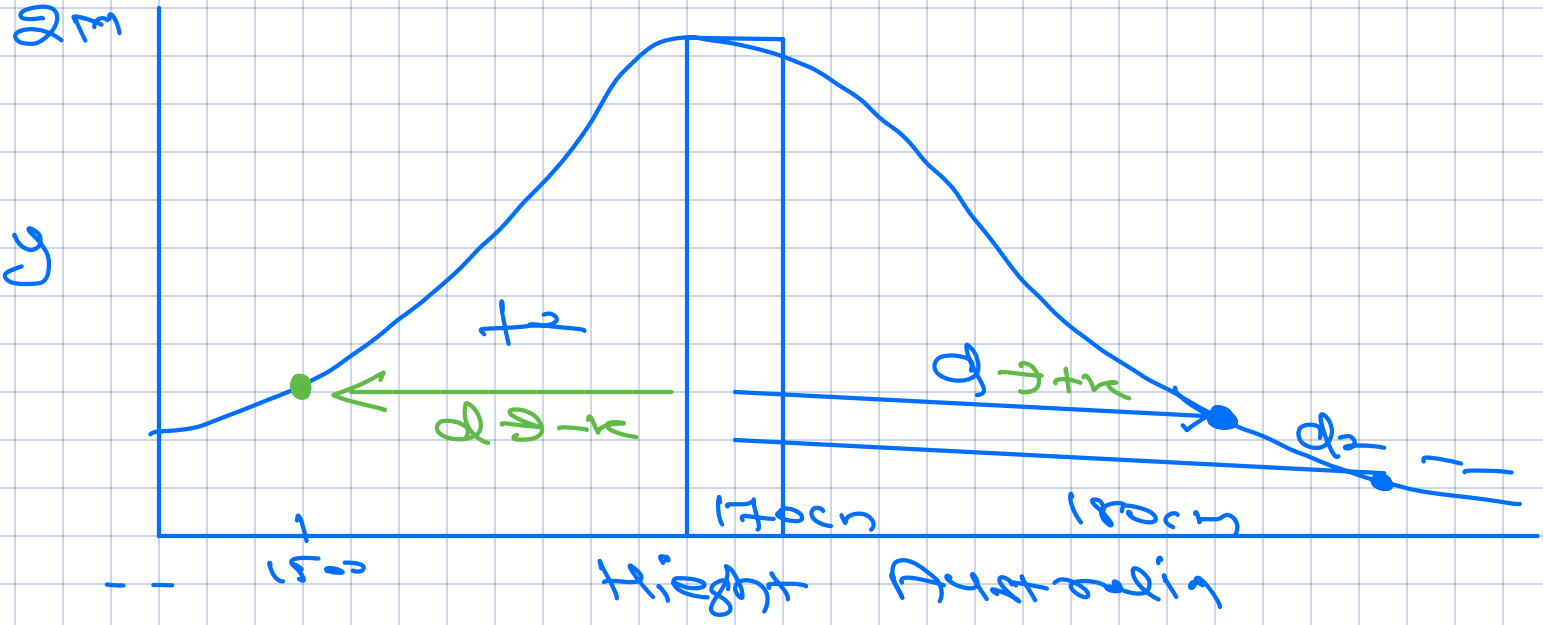
Q_1 → 25th %




$Q_3 + 1.5\ IQR$
Q_3
Q_2
Q_1
IQR
$Q_1 - 1.5 * IQR$
min

Variance





① Avg distance of each point from mean

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

(Variance)

μ = mean of Datapoint

Avg squared Distance from mean

$$\sigma^2_{Aus} > \sigma^2_{ind}$$

$$\sigma^2_{Aus} = 170\ cm^2$$

Standard Deviation

$$Std = \sqrt{Variance}$$

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

Avg Distance from mean

$$Std_{Aus} > Std_{ind}$$

$x_1 = 120$ $\quad (x_i - \mu)^2$

mean $= 150$

$x_3 = 170$

Random Variable

$X_i \longrightarrow$ Random Variable

A variable which has uncertainity assigned to it

$O_i \rightarrow$ Outcome (y Coin toss)

O_i < H (0.5), T (0.5)

① Discrete RV ⇒ Outcome of Coin Toss

② Continious ⇒ Stock price of Reliance,

Temperature in Next 6 mons,

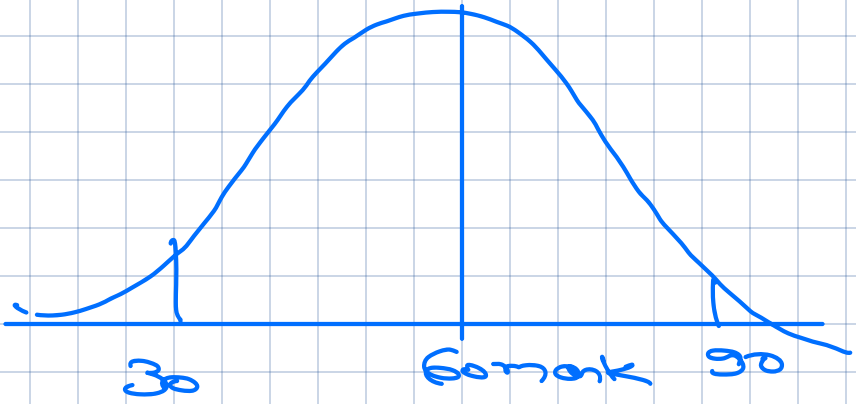
Rainfall Value,

Hieght → 170 cm, 171 cm ∞ → Continous

ML Model → Predict Mark Student will get

⇒ Assignment Solve

⇒ Classes Attended

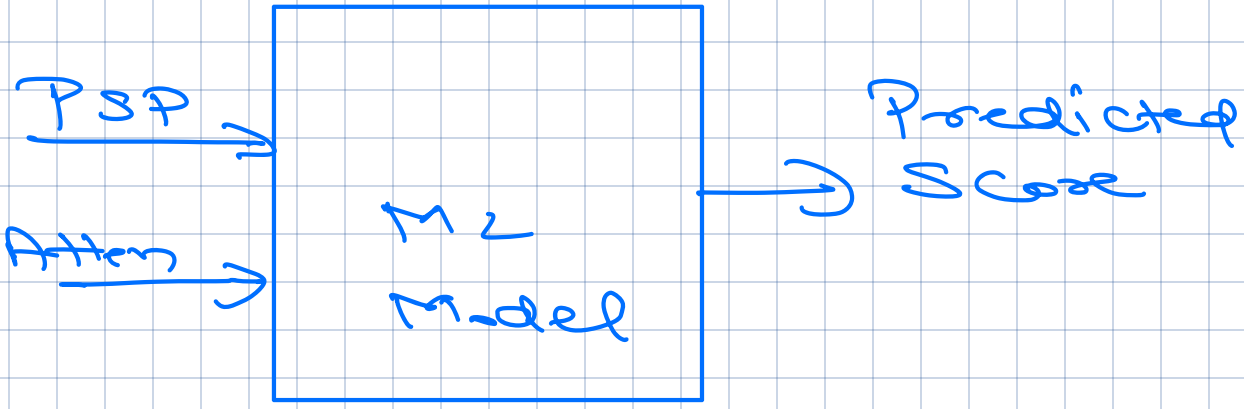


∴ PSP → 95 Att → 80 → Predicted → 75

80 Actual from Past Data

→ Model made mistake

→ 5 units off



∴ PSP → 75 Att → 70 → Predicted 65

Actual 60

→ Model made mistake

→ -5 units off

Goal of an Model

- minimize the mistakes
- minimizing Errors

MSE $= \frac{\sum_{i=1}^{n} (\text{Actual} - \text{Prediction})^2}{n}$

RMSE $= \sqrt{MSE}$

Perplexity / Cross entropy

MSE

and