Descriptive Statistics

Measure of Dispersion

1. Range → max − min → extreme values
2. IQR (Inter Quartile Range) → get rid of extreme values → $Q_3 - Q_1$ (25% … 75%)

Percentage

0 1 2 3 4 5 6 7 8 9

56, 64, 75, 80, 83, 90, 92, 95, 98, 100

$n = 10$ (even num of values)

Percentile

Test scored of learners

50% → Median → $\frac{83+90}{2} = \frac{173}{2} = 86.5$

25% → 75

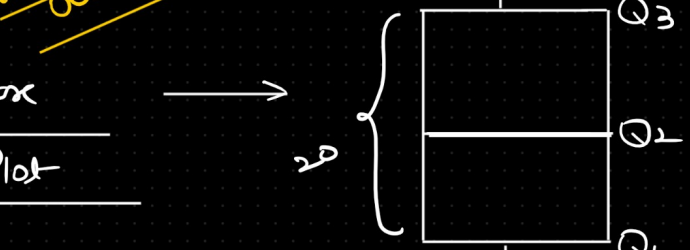
75% → 95

$$IQR = 95 - 75 = 20$$

Standard Value

Upper whisker → $Q_3 + 1.5 * IQR$

Box Plot (detect the outliers)



Lower whisker → $Q_1 - 1.5 * IQR$

Outliers

Interpretation: The IQR of 20 means that the middle 50% of test scores are within 20 points of each other, giving you a better understanding of the students' typical performance without being influenced by the extreme scores.

Variance & Standard Deviation

↳ spread of the data wrt mean value

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

N → size of population data

μ → Population mean

Task 1

Why??

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

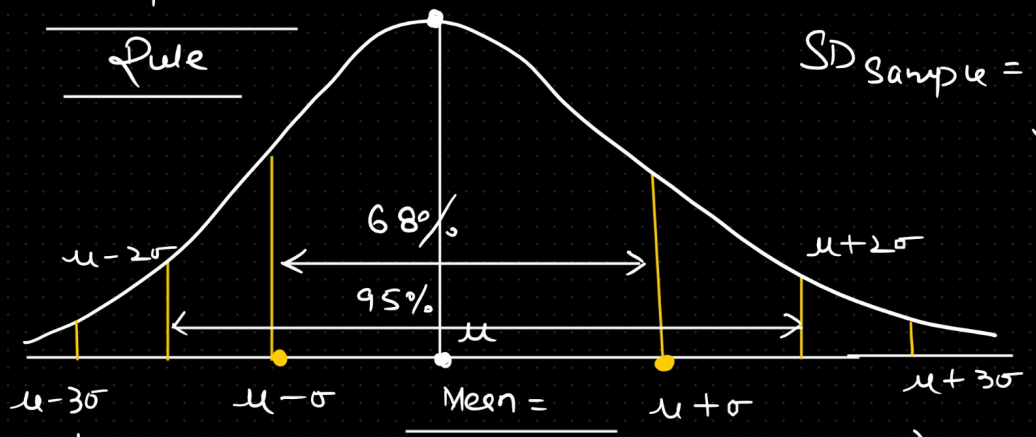
n → small subset (sample data)

σ^2 → variance

$\sqrt{\text{variance}}$ → $SD_{Pop} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

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

Empirical Rule



Mean =

Median =

Mode

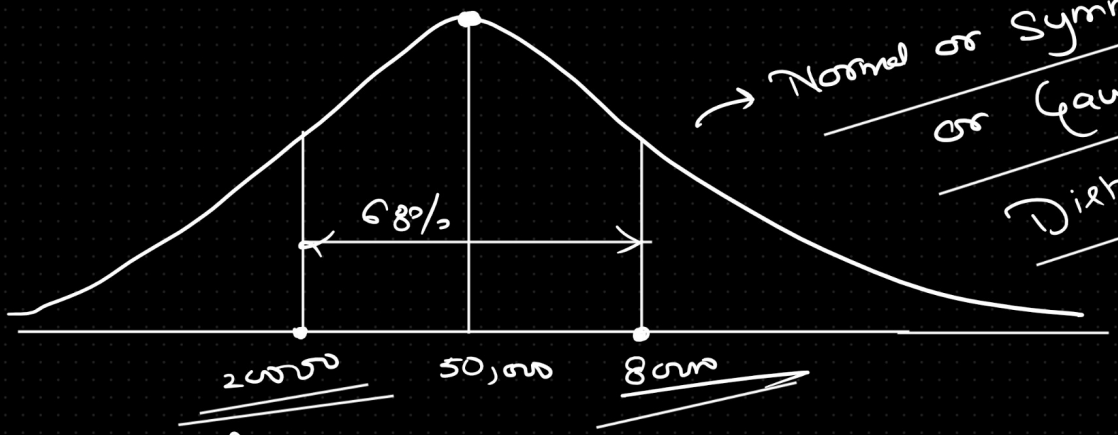
99.7%

Variance vs SD ??

Variance quantifies the spread of data points around the mean but can be hard to interpret due to squared units.
Standard Deviation offers a more intuitive measure of spread in the same units as the original data, making it easier to understand and communicate the variability in your dataset.

Metric is very widely used

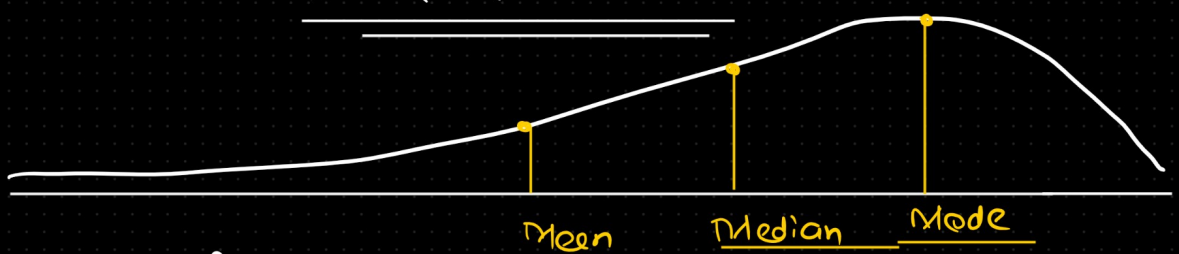
Budget or financial Planning



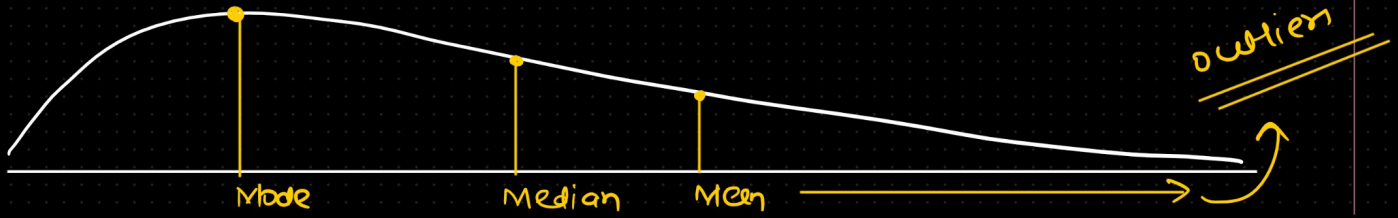
→ Normal or Symmetric or Gaussian Distribution

Negative
Left Skewed

Mean < Median < Mode



Positive
Right Skewed → Mode < Median < Mean



Income Distribution → Right skewed distribution

Gamma Distribution → Right skewed data

- $\alpha > 1$ → Right skewed
- α → shape
- β → scale → Larger