9th - November - 2020

Unit → Summarizing Quantitative Data

Statistics

Descriptive Inferential

⇒ Measuring Central Tendency

- Mean
- median
- mode

- Choosing the 'best' measure of central

when there is outlier → median

else → mean

- Mean acts as a balancing point.

⇒ Inter Quartile Range (IQR)

- Steps →
  i) Find the median
  ii) Split the data along the median
  iii) Find medians of the two parts (Q_1, Q_3)
  iv) $IQR = Q3 - Q1$

3, 5, 7, 1, 1, 8, 4, 6 ⇒ 1, 1, 3, 4, 5, 6, 7, 8

$Q_1 = 2$ ↑ 6.5 ⇒ Q_3 ↑

1 | 1 3 | 4 | 5 | 6 7 | 8

- Why ~~prefer~~ IQR when we already have range?
  Range is highly sensitive to outliers.
  ∴ IQR is a better representative.

⇒ Variance and standard deviation of a population.

- We had earlier covered about central tendency (mean, median & mode)
- But let's say we have two distributions,

  -10, 0, 10, 20, 30 8 9 10 11 12
- In both of them, mean/median is same, what's varying is the dispersion.
- How to calculate it ?
  - Range
  - Variance (σ^2) = $\sum_{i=1}^{n} \frac{1}{n} (x_i - \bar{x})^2$
  - Std. deviation (σ)

Note :

Std. dev. as same units as it were in the original distribution.

- Standard deviation cannot be ~~zero~~ negative.

when all points are closer to mean, std dev is close to 0.

Note:

$$\text{S.D. for sample} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = s_{n-1}$$

⇒ Faster way of calculating σ^2.

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

⇒ Population and sample standard deviation

The formula we used for standard deviation depends on whether the data is being considered a population or the data is a sample representing a large population.

For population:

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

For sample:

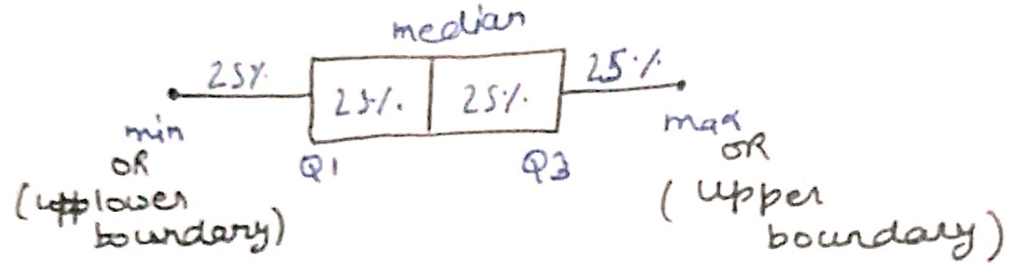
$$s_{n-1} = \sqrt{\frac{\Sigma(x_i - \bar{x})}{n-1}}$$

Note:

s_{n-1} is a better estimate of σ than s_n.

⇒ Box plots

displays a 5 number summary



⇒ Judging outliers

The numbers lying ~~outside~~ above upper boundary & or below lower boundary are outliers.

upper boundary = Q3 + 1.5 * x IQR

lower boundary = Q1 - 1.5 x IQR

⇒ OTHER MEASURES OF SPREAD :

Range = max - min

mid Range = $\frac{max + min}{2}$

Mean Absolute deviation (MAD) = $\frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$