19 November 2023 09:17 AM

* DESCRIPTIVE STATISTICS:-

1) MEASURES OF CENTRAL TENDENCY:-

> MEAN : Average

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

$[20, 21, 22, 24, 25] = \frac{112}{5}$

$= 22.4$

$$\mu = \frac{\text{Sum of all terms}}{\text{total number of terms}}$$

> MEDIAN : Mid point (Sorted data)

i) n is odd : $\frac{n+1}{2}$ th

Ex: 10, 20, 30, 40, 50

$n = 5$

median $= \frac{5+1}{2} = 3^{rd}$ term

$= 30$ Ans.

ii) n is Even: $\frac{\left(\frac{n}{2}\right)\text{th} + \left(\frac{n}{2}+1\right)\text{th}}{2}$

Ex: 10, 20, 30, 40, 50, 60

$n = 6$

median $= \frac{30+40}{2} = 35$ Ans.

> MODE : The most present term

Ex: 10 20 30 40 20 10 20 50

Ex: 10, 20, 30, 40, 20, 10, 20, 50

mode = 20

Ex: 10, 20, 30, 40, 20, 10, 20, 50, 10

mode = 10, 20

---

2) MEASURES OF DISPERSION:-

> RANGE : UB - LB (LB to UB)

> VARIANCE :

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

Ex1: 20, 22, 25, 28, 30 , $\mu = 25$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
20	-5	25
22	-3	9
25	0	0
28	3	9
30	5	25

$$\sigma^2 = \frac{68}{5} = 13.6 \text{ Ans.}$$

Ex2: 20, 30, 40, 50, 60 , $\mu = 40$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
20	-20	400
30	-10	100
40	0	0
50	10	100
60	20	400

$$\sigma^2 = \frac{1000}{5} = 200$$

> STANDARD DEVIATION :

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

Ex: 1 $\sigma^2 = 13.6$

$\sigma = 3.69$

Ex: 2 $\sigma^2 = 200$

$\sigma = 14.14$

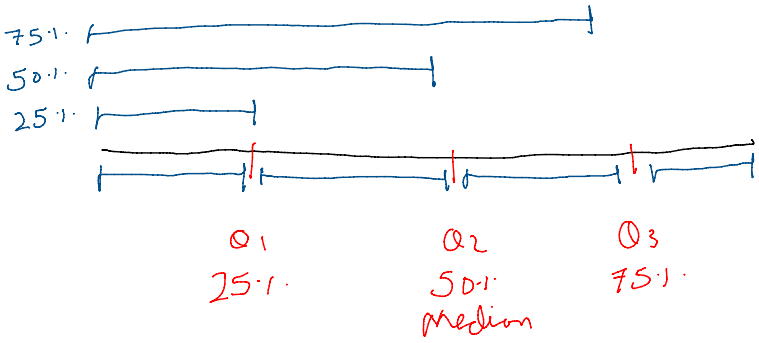
> QUARTILES :-

> Quartiles divide our data into four quarters

> Q1 : Value below which 25% of data lies

> Q2 : Value below which 50% of data lies

> Q3 : Value below which 75% of data lies



Ex: Salaries in an Org

$Q_1 = 60000$, $Q_2 = 80000$, $Q_3 = 100000$

> 25% employees have salary < 60000

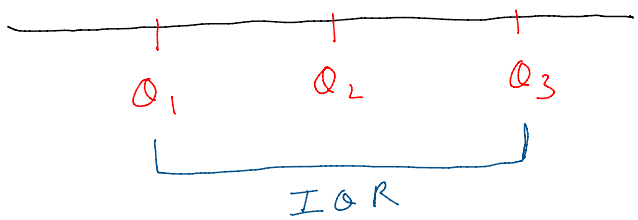
> 50% employees have salary < 80000

> 75% employees have salary < 100000

* Inter quartile range :

- [illegible] of data

* Inter Quartile range:
  > Central 50% of data
  > IQR = $Q_3 - Q_1$



---

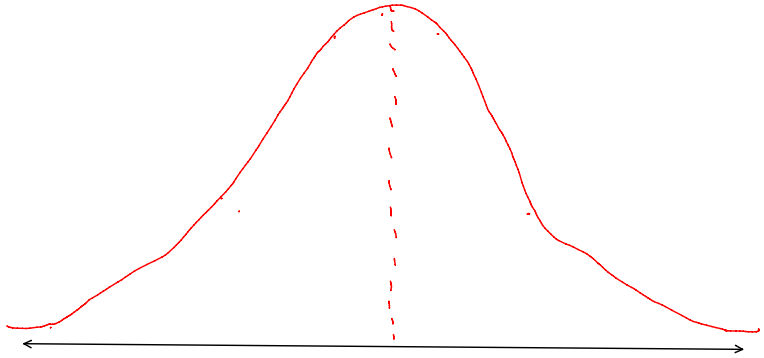
* OUTLIER:-
  > Data point which is away from the general pattern
  > Odd one out

* Effect of outlier on mean and median.
  * MEAN:-
    > 20, 21, 22, 23, 25, 27, 28, 30, 31, 33 , $\mu = 26$
    > 20, 21, 22, 23, 25, 27, 28, 30, 31, 73 , $\mu = 30$
      > Mean shifts towards the outlier.

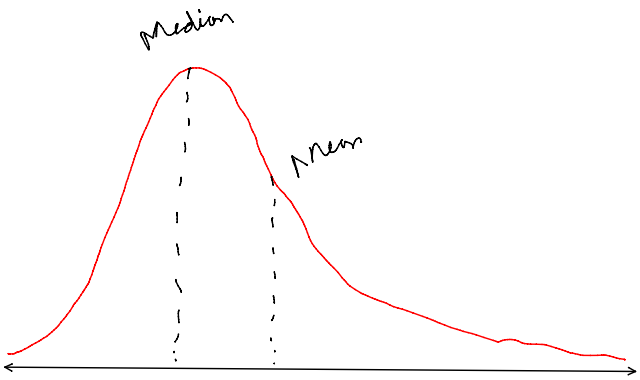
  * MEDIAN:-
    > 20, 21, 22, 23, 25, 27, 28, 30, 31, 33 , mod = 26
    > 20, 21, 22, 23, 25, 27, 28, 30, 31, 73, mod = 26

---

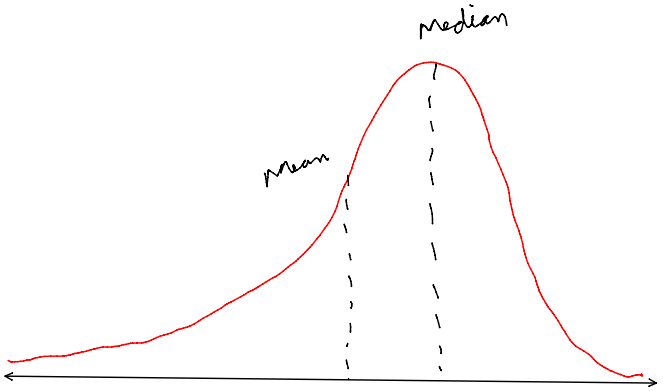
* SKEWNESS:-



Symmetric data
Mean ≈ Median



Right Skewed
Positive Skewed
Mean > Median



Left Skewed
Negative Skewed
Mean < Median

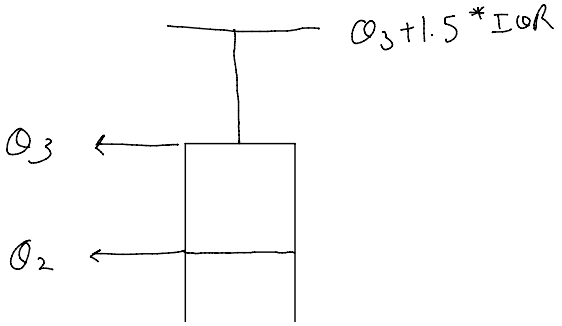
---

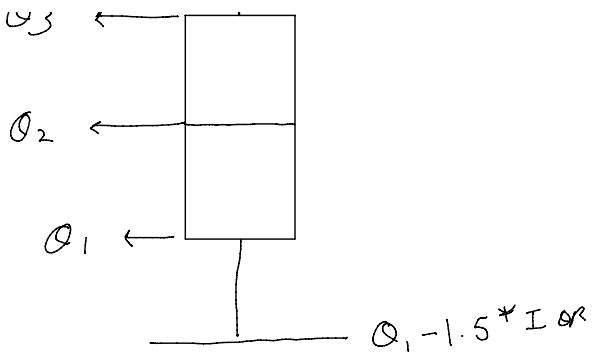
* OUTLIER DETECTION:-

1) BOX PLOT
2) Z SCORE

. — outlier

1) BOX PLOT :-





2) Z SCORE:-

$$Z = \frac{x - \mu}{\sigma}$$

Ex $\mu = 30$

$\sigma = 5$

$X = 45$

$$Z = \frac{45 - 30}{5} = 3$$

* Outliers have high z scores. (-/+)

Ex: 20, 22, 25, 26, 28, 59

$\mu = 30$

$\sigma = 13$

x_i	$x_i - \mu$	$(x_i - \mu)^2$	z $(x_i - \mu)/\sigma$
20	-10	100	-0.76
22	-8	64	-0.61
25	-5	25	-0.38
26	-4	16	-0.31
28	-2	4	-0.15
59	29	841	2.23

* COVARIANCE :-

> measure of linear association b/w two features

$$Cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

* CORRELATION :-

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

> Range = -1 to +1
> $Cor(x,y) \approx +1$, Very high +ve Cor
> $Cor(x,y) \approx -1$, Very high -ve Cor

---

* VISUALISATIONS