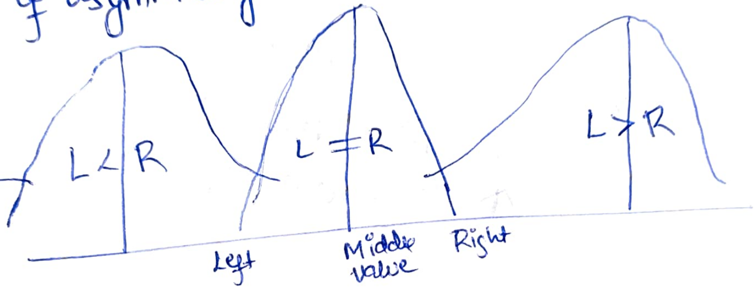
Skewness:- Skewness is the deviation from Symmentry observed in data.

* measure of asymmetry.

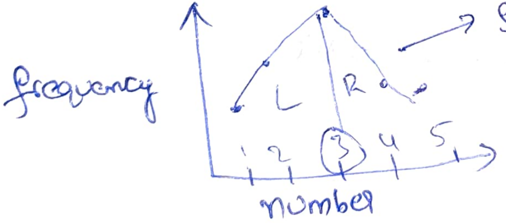
Negative Skewness
(left-modal)



positive Skewness
(right-modal)

(No Skewness)

Ex: data: {1, 2, 2, 3, 3, 3, 4, 4, 5}



Symmetrical

median = 3 (which divides data into 2 equal parts)

$\boxed{\text{mean} = \text{median} = \text{mode}}$

all the Central tendency values are Equal.

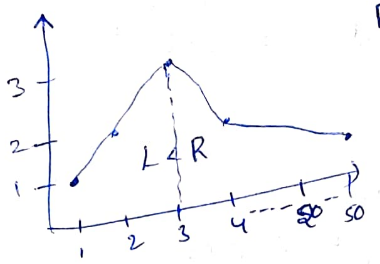
$\mu = \frac{27}{9} = 3.$

median = 3

mode = 3

Ex2: data = $\{1, 2, 2, 3, 3, 3, 4, 4, 50\}$

positive skewed



$\mu = \frac{72}{9} = 8$, median = 3, mode = 3.

* mean is prone to outliers.

Ex3: data = $\{-8, 2, 2, 3, 3, 3, 4, 4, 5\}$

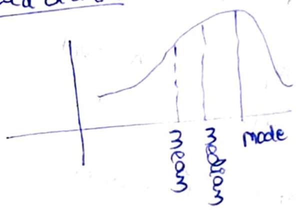
left skewed data.

$\mu = \frac{18}{9} = 2$, median = 3, mode = 3

Positive (or) Right skewed data:

mean > median > mode

Negative (or) left skewed data:



mean < median < mode

Outliers:- * Data point which is away from general pattern.

* odd one out.

mean
20 21 22 23 24 25 27 $\boxed{\mu = 23}$
20 21 22 23 24 25 70 $\boxed{\mu = 25}$ → Outlier

median
20 21 22 23 24 25 27
20 21 22 23 24 25 70

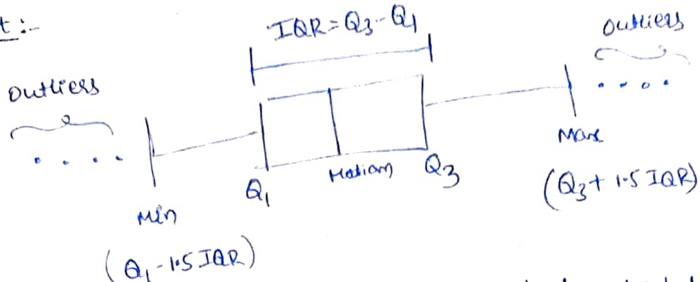
median = 23 for both data.

* It is suggestable to take median, when data have outliers.

Outlier detection:- is the process of detecting data points which are not as per the general pattern in data.

Two ways: Boxplot & Z-score

Boxplot :-



* The Outliers are the points that are present beyond & below the upper & lower whiskers.

Z-score:- Z score tells how many standard deviations (sd) away a data point is from the mean standard score.

$$Z\text{-score} = \frac{X - \mu}{\sigma}$$

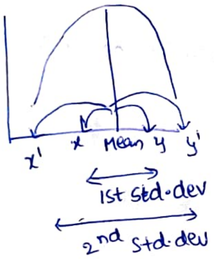
(obs) ← X; → (mean) μ; → (std.dev) σ

Ex:

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

$x = 45$

$\sigma = 5$ $\quad$ The point x is 3 std.devs away from mean



$x = \mu - SD$ $\quad x' = \mu - 2SD$

$y = \mu + SD$ $\quad y' = \mu + 2SD$

3 std.dev → 99% of data
2 std.dev → 95% of data
1 std.dev → 65 - 60% of data will be covered

Ex2: | 20 | 22 | 25 | 26 | 28 | 59 | → outlier

$\mu = 30$
$\sigma = 13$

x	$x_i - \mu$	$(x_i - \mu)^2$	Z score
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

-0.76 … -0.15 → values falls in 1st S.D

2.23 → falls in 3rd S.D more compared to other values.

Covariance: Measure of lineal association b/w two features.
proportionality — numeric

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

$(x_i-\bar{x})$: diff of i^{th} obs w.r.t Mean $(\bar{x})$

Ex: hours: {1, 2, 3, 4, 5}
Marks: {20, 40, 60, 80, 100}

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

$\bar{x} = 3$ $\bar{y} = 60$

x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
1	20	-2	-40	80
2	40	-1	-20	20
3	60	0	0	0
4	80	1	20	20
5	100	2	40	80
				200

$$Cov(x,y) = \frac{200}{5}$$

$$\boxed{Cov(x,y) = 40}$$

+ve cov → the two features have +ve relationship.
-ve cov → the " " will " -ve relationship.

Correlation: Standardized measure of strength and direction of the lineal relationship b/w two variables.

Range ⇒ {-1 to +1}
$Cor(x,y) = +1$, very high +ve corr
$Cor(x,y) = -1$, very high -ve corr

$$\boxed{Correlation = \frac{Cov(x,y)}{\sigma_x * \sigma_y}}$$

-ve relationship   +ve relationship

-1   0   +1

Pick above example:-

$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
4	1600
1	400
0	0
1	400
4	1600
10	4000

$\sigma_x^2 = \frac{10}{5} \Rightarrow \sqrt{2}$

$\sigma_y^2 = \frac{4000}{5} \Rightarrow \sqrt{800}$

$$Correlation = \frac{40}{\sqrt{800}\sqrt{2}} = \frac{40}{20\times\sqrt{2}\times\sqrt{2}}$$

$= \cancel{0.025}\ \frac{40}{40}$

$= 1$.

∴ features are highly correlated.