① percentiles and Quartiles [

percentage :-

Ex:- 1, 2, 3, 4, 5, 6

% of numbers that are odd $= \frac{3}{6} = \frac{\text{No. of odd numbers}}{\text{Total No. of numbers}}$

$= 1/2 = 50\%$

percentiles :- A percentile is a value below which a certain percentage of data points lie.

$x = \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12\}$

① percentile Rank of $10 = \frac{\#\text{ of values below } 10}{} \times 100$

$= \frac{\cancel{12}^{4}}{\cancel{15}_{1}} \times \cancel{100}^{20} = 80$ percentile.

80 percentile = 80% of the distribution fall below the value of 10.

② what value exists at 25^{th} percentile?

$\text{value} = \frac{\text{percentile}}{100} * (n+1)$

$= \frac{\cancel{25}}{\cancel{100}} \times \cancel{16}\ 4 = 4^{th}$ element

= ~~4~~ ~~th~~ ~~cordeca~~

Suppose if value is in decimal form in that case we actually needs to take average of the last No. present at the value and next number in the sequence

$$x = \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12\}$$

when value $= 4.5 \Rightarrow \frac{4+6}{2} = 5$

Quartiles:-

$Q_1 \rightarrow 25$ percentile

$Q_2 \rightarrow$ Median $\rightarrow 50$ percentile.

$Q_3 \rightarrow 75$ percentile.

Five Number summary:-

1) Minimum
2) First Quartile (25 percentile) $\rightarrow Q_1$
3) Median (Q_2)
4) Third Quartile (75 percentile) (Q_3)
5) Maximum.

Removing the outliers:-

$$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$$

[Lower Fence $\longleftrightarrow$ Higher Fence]

Lower Fence $= Q_1 - 1.5\,(IQR)$ Inter Quartile Range $= Q_3 - Q_1$

Higher Fence $= Q_3 + 1.5\,(IQR)$

$$x = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$$

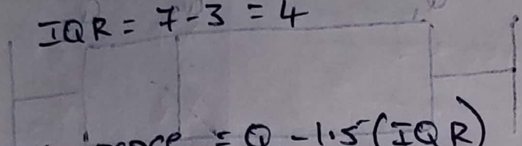
$Q_1 = 25^{t}$ percentile $= \frac{25}{100} \times (19+1)$

$= \frac{\cancel{25}^{5}}{\cancel{100}_{5}} \times \cancel{20} = 5^{th}$ element $= 3$

$Q_3 = 75$ percentile $= \frac{\cancel{75}}{\cancel{100}_{5}} \times (\cancel{20})$

$= 15^{th}$ element $= 7$

$IQR = 7 - 3 = 4$



Lower Fence $= Q_1 - 1.5(IQR)$

$= 3 - 1.5(4)$

$= 3 - 6$

$= -3$

Higher Fence $= Q_3 + 1.5(IQR)$

$= 7 + 1.5(4)$

$= 7 + 6$

$= 13$

The values which are not in this range of values $[-3, 13]$ are considered as outliers.

Box plot :- [To visulalize outliers]

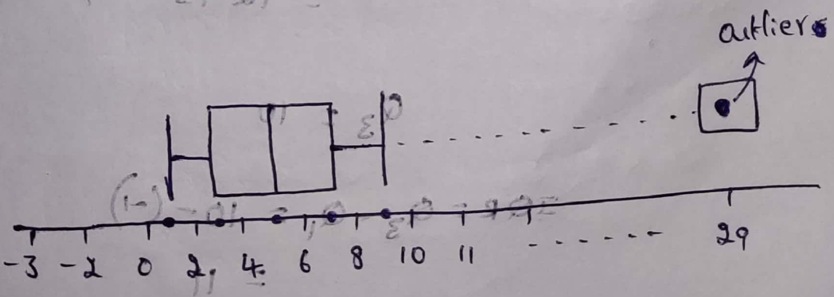
① min value $= 1$

② $Q_1 = 3$

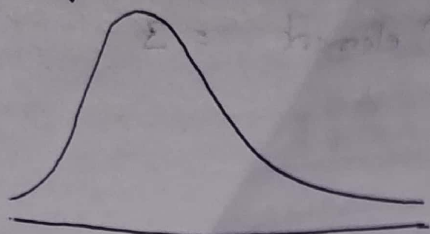
③ median $Q_2 = 5$

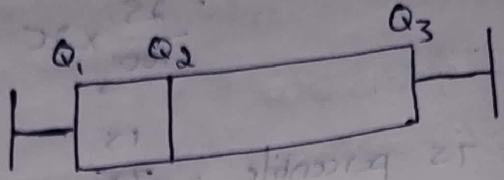
④ $Q_3 = 7$

⑤ Maximum $= 9$



Right skewed

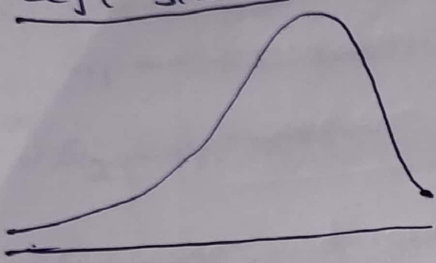


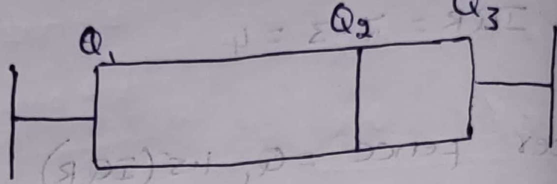


$Q_3 - Q_2 > Q_2 - Q_1$

mean > median > mode

Left skewed





$Q_2 - Q_1 > Q_3 - Q_2$

mode > median > mean

Find the outliers and draw box plot ?

$y = \{-13, -12, -6, -5, 3, 4, 5, 6, 7, 7, 8, 10, 10, 11, 24, 55\}$

$Q_1 = 25^{th}$ percentile $= \frac{25}{100} \times (16+1)$

$= \frac{25}{100} \times 17 = 4.25$

$Q_1 = -\frac{2}{2}$

$Q_3 = 75^{th}$ percentile $= \frac{75}{100} \times (16+1)$

$= 12.75 \Rightarrow \left(\frac{10+10}{2} = 10\right)$

$Q_3 = 10$

$IQR = Q_3 - Q_1 = 10 - (-1)$

$= 11$

Lower fense $= -1 - 1.5(10+1)$

$= -1 - 1.5(11)$

$= -17.5$

Higher fense $= 10 + 1.5(10+1)$

$= 10 + 16.5$

$= 26.5$

→ Here 55 is an outlier

$z = \{1, 2, 4, 6, 7, 12, 18, 34, 77, 66, 108, 99, 14\}$

<u>step</u> sort

$z = \{1, 2, 4, 6, 7, 12, 14, 18, 34, 66, 77, 99, 108\}$

$Q_1 = \frac{\cancel{25}^{1}}{\cancel{100}_{4}} \times 14 = 3.5$ $\left(\frac{4+6}{2} = 5\right) = 5$

$Q_3 = \frac{\cancel{75}^{3}}{\cancel{100}_{4}} \times 14 = 10.5$ $\left(\frac{66+77}{2}\right) = 71.5$

$IQR = 71.5 - 5 = 66.5$

$LF = 5 - 1.5(66.5)$

$= -94.75$

$HF = 71.5 + 1.5(66.5)$

$= 171.25$

→ There are no outliers in this dataset.

Covariance and correlation :-

→ one of the very important topic for data preprocessing, data analysis and feature selection.

Let us consider two random variables :-

size(x)	price (y)
1200 sqm	100 k $
1500 sqm	200 k $
1800 sqm	300 k $
2000 sqm	400 k $

Relationship between size and price

X ↑	Y ↑
X ↓	Y ↑
X ↑	Y ↓
X ↓	Y ↓

→ Covariance and correlation are two statistical concepts that are used to measure the relationship between two variables. Although they are different concepts and have different interpretations.

→ covariance :- measures how two variables vary together specifically covariance measures how much two variables vary from their respective means at the same time

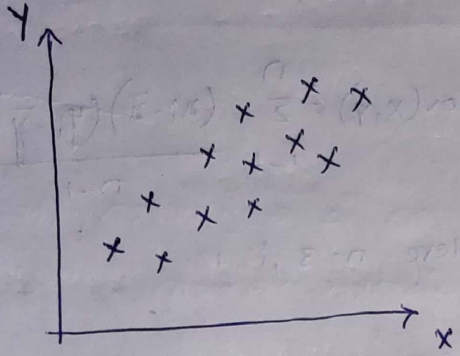
→ A positive covariance means that the two variables tend to increase or decrease together.

→ while a negative covariance means that one variable tends to increase while the other decreases.

covariance tells us about direction of relationship.

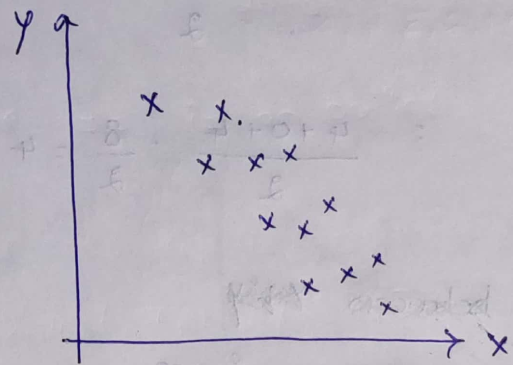
if x ↑, Y ↑ = +ve direction

if x ↑, Y ↓ = −ve direction



$X \uparrow \quad Y \uparrow$

$X \downarrow \quad Y \downarrow$



$X \uparrow \; Y \downarrow$

$X \downarrow \; Y \uparrow$

Covariance :-

$$\text{cov}(x,y) = \sum \frac{(x_i - \bar{x})(y - \bar{y})}{n-1}$$

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

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

$$\boxed{\text{Var}(x) \Leftarrow \text{cov}(x,x)}$$

$\text{Cov}(x,y)$ $\boxed{\begin{matrix} X\uparrow Y\uparrow \\ X\downarrow Y\downarrow \end{matrix}}$ $\Rightarrow$ +ve covariance

$\boxed{\begin{matrix} X\uparrow Y\downarrow \\ X\downarrow Y\uparrow \end{matrix}}$ $\Rightarrow$ −ve covariance

Ex:-

X	Y
2	3
4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

$$cov(x,y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Here $n=3$, $i=1$

$$= \frac{[(2-4)(3-5)+(4-4)(5-5)+(6-4)*(7-5)]}{2}$$

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4 \quad \text{+ve covariance}$$

~~conclusion: Relationship between x&y~~

→ x and y are having a positive covariance.

Advantage

(i) Tells us about relationship between x & y.

disadvatage

(1) Covariance does not have a specific limit value. ($-\infty$ to ∞)

pearson correlation coefficient :-

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

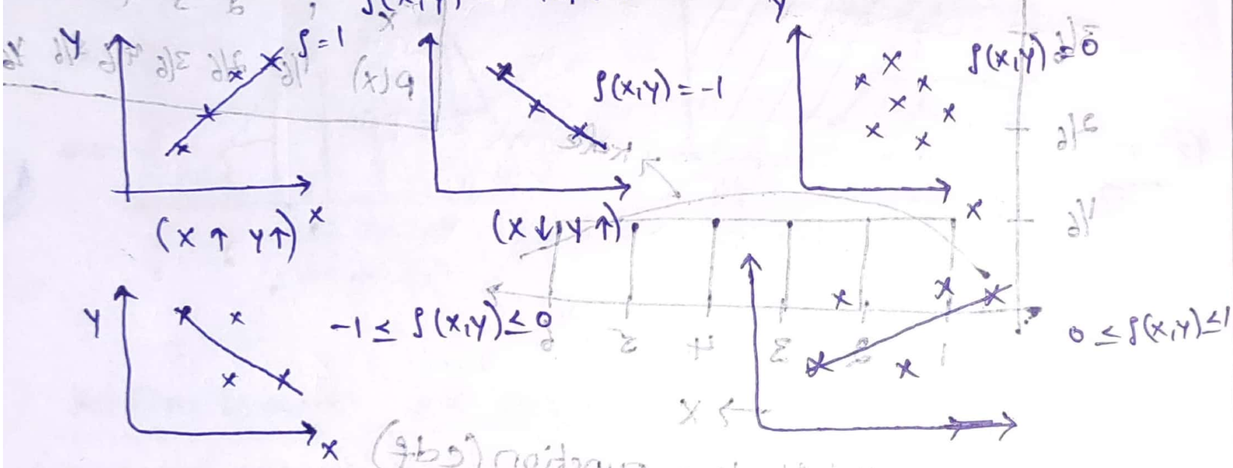
→ correlation measures the strength and direction of the linear relationship between two variables, without being affected by units of measurements.

→ The more the value towards +1 the more +ve correlated it is.

→ The more the value towards −1 the more −ve correlated it is.

→ when we try to find the pearson correlation coefficient the range of values lies between

$\rho(x,y) = -1 < \rho < 1$



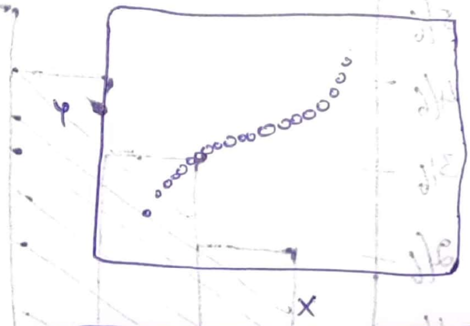
Spearman's correlation coefficient :-

→ spearman's rank correlation measures the strength and direction of association between two ranked variables.

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \cdot \sigma_{R(y)}}$$

spearman correlation − 1

pearson correlation − 0.88



↳ spearman's correlation coefficient ranges between −1 and 1.

→ Here correlation 1 represents a perfect monotonic relationship between the two variables being correlated.

→ which means that as one variable increases the other variable also increases in a perfectly predictable manner.

→ This is observed when the two variables are perfectly ranked.

X	Y	R(X)	R(Y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	5
2	2	4	4