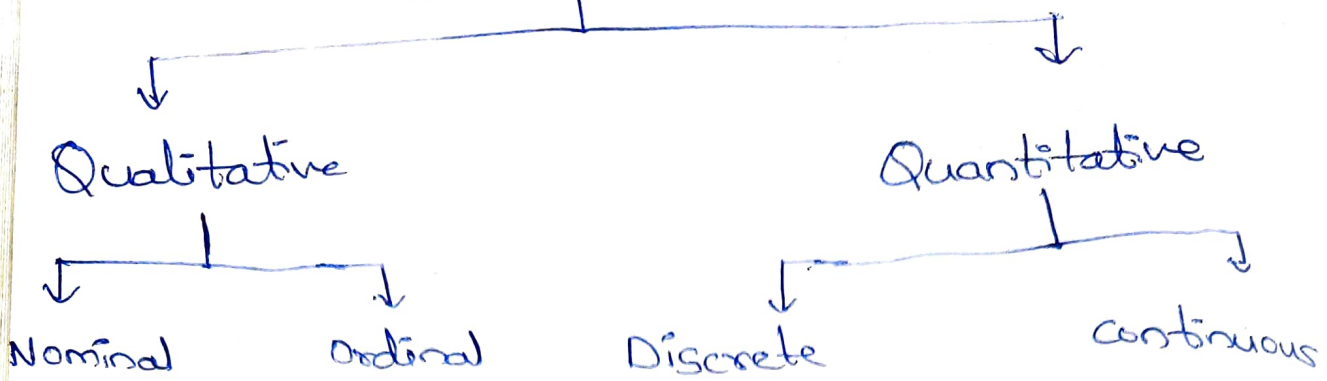
Types of Data



Qualitative	Quantitative
Category data	Numbers
Men, Women Blue, Black American, Non American	

Nominal: Men, Women
Blue, Black
American, Non American

Ordinal: Poor, Rich, Middle (Hierarchy defined)
Poor, Average, Outstanding

Discrete: 10, 20 (Not having float values)
11, 13
Ex: No. of people in Room

Continuous: Weight of people (can be int, float)

Descriptive Statistics

1) Measures of Central Tendency:
   a) Mean b) Median c) Mode

1.1) Mean: Average of the data

Ex: 10, 11, 13, 17, 19, 23

$$\Rightarrow \frac{10+11+13+17+19+23}{6} = 15.5$$

Ex2: 10, 11, 13, 17, 19, 23, 24, 500 ← outlier

$$\Rightarrow \frac{10+11+13+17+19+23+24+500}{8} = 77.125$$

Outliers: Extremely high or low value

* Problem with mean is that it is affected hugely in the case of outliers present in the data.

1.2 Median: Exact middle of a sorted data

7, 9, 13, 11, 10, 5, 4

*a) Sort the data

b) Take the middle value $\left(\frac{n+1}{2}\right)$

4, 5, 7, (9) 10, 11, 13 → odd data

↓

Median

Ex 2: 4, 5, 7, (9, 10) 11, 13, 14 → even data

$\frac{9+10}{2} = 9.5$

$\frac{\left(\frac{n}{2}\right) + \left(\frac{n}{2}+1\right)}{2}$

Advantage of Median

→ It is robust in the case of outlier present in the data.

10 11 13 (17) 19 23 24

10 11 13 (17 19) 23 24 500

$\frac{17+19}{2} = 18$

1.3 Mode :- Most frequently occuring value in a data

P_1	Yes
P_2	Yes
P_3	yes
P_4	No
P_5	yes
P_6	No

Central value

Mode = Yes

Two mode
Bimodal
Three mode
Multimodal

2 Measure of Spread/Dispersion

a) Range (b) Variance (c) Standard Deviation
d) Median absolute Deviation (e) Percentile
f) Quartiles (g) Inter Quantile Range

a) Range → Max - Min

① 19 - 10 = 9

② 550 - 5 = 545

10	5
11	50
14	55
16	105
17	300
19	550

Outlier: 10, 11, 12, 13, 14, 15, 16, 17, 110

110 - 10 = 100

* Effected by the presence of outlier in the data.

2.2 Variance → Spread of data around mean

$$Var = \frac{\Sigma \text{ datapoint} - \text{mean}}{}$$

10, 11, 13, 17, 19, 23

① Mean → 15.5

② datapoint - mean ⇒

10 - 15.5 = -5.5 +
11 - 15.5 = -4.5 +
13 - 15.5 = -2.5 +
17 - 15.5 = 2.5 +
19 - 15.5 = 4.5 +
23 - 15.5 = 7.5 = 2

$$= \Sigma (\text{datapoint} - \text{mean})^2$$

$$\Rightarrow (-5.5)^2 + (-4.5)^2 + (-2.5)^2 + (2.5)^2 + (4.5)^2 + (7.5)^2$$

=

Because of squaring each value, there will be explosion of value

$$Var = \frac{\Sigma (datapoint - Mean)^2}{No.\ of\ points}$$

(1.3) Standard Deviation →

$$SD = \sqrt{Variance}$$

Q) Are variance & SD affected by outliers?

A) Yes, because of presence of mean in calculation of var & SD.

(1.4) Median Absolution deviation →

$$\frac{\Sigma (datapoint - median)}{N}$$

* It is not effected by outlier

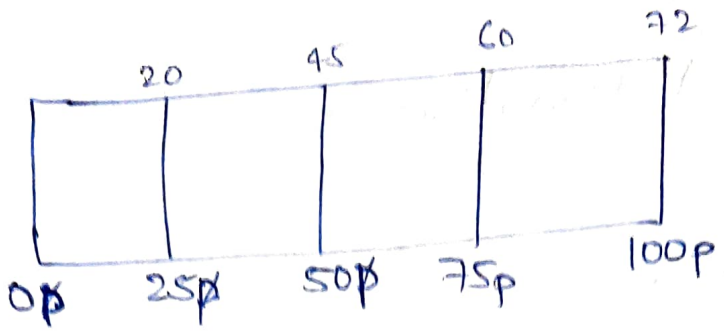
(1.5) Percentile

Stud = $\frac{61}{100}$ = 61% → 99 percentile

1% students who score > 61%

1.6) Quartiles

25p = <20

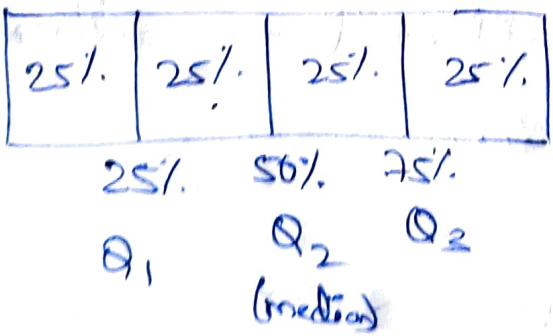


Step1 → Sort the data

25% person score less than 20
75% " " more than 75

100% is scored by 72

1.7) Inter Quartile Range



IQR = Q3 - Q1

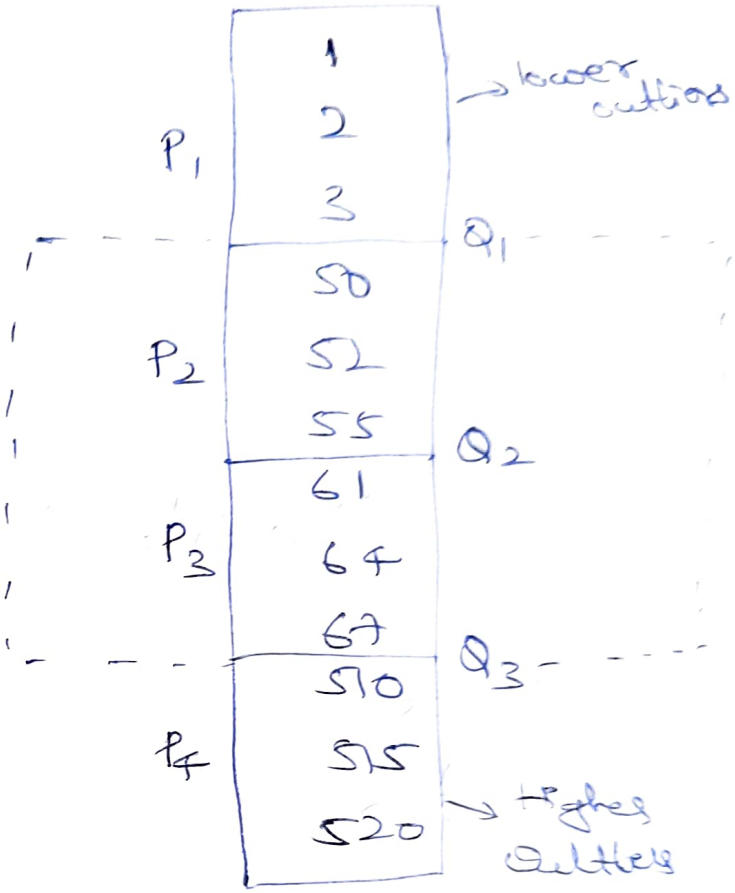
* IQR helps us remove outliers from the

data.

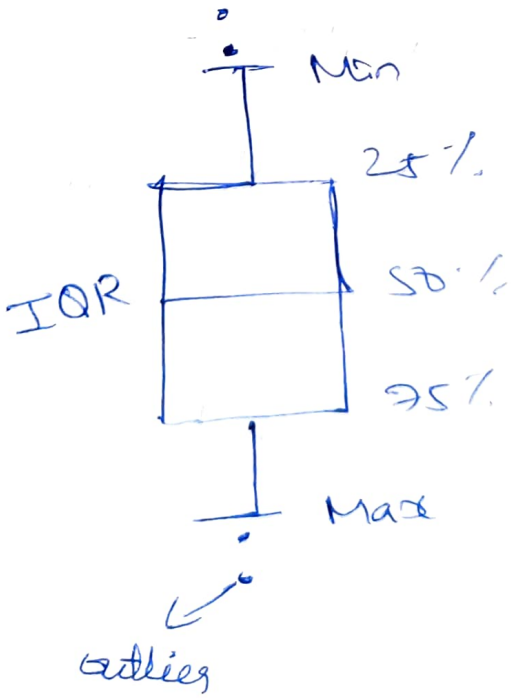
1, 2, 3, 50, 52, 55, 61, 64, 67, 510, 515, 520

① Sort the data.

$IQR = Q_3 - Q_1$



Boxplot ⇒ Five point summary of the data



Lower Outliers $\leq Q_1 - 1.5 * IQR$

Upper Outliers $\geq Q_3 + 1.5 * IQR$

Boxplot helps us detect outliers & also helps remove outliers.

Sometimes it may remove genuine data

Bivariate Analysis

1. Covariance
2. Correlation

Covariance: Measures the direction of relationship between two variables

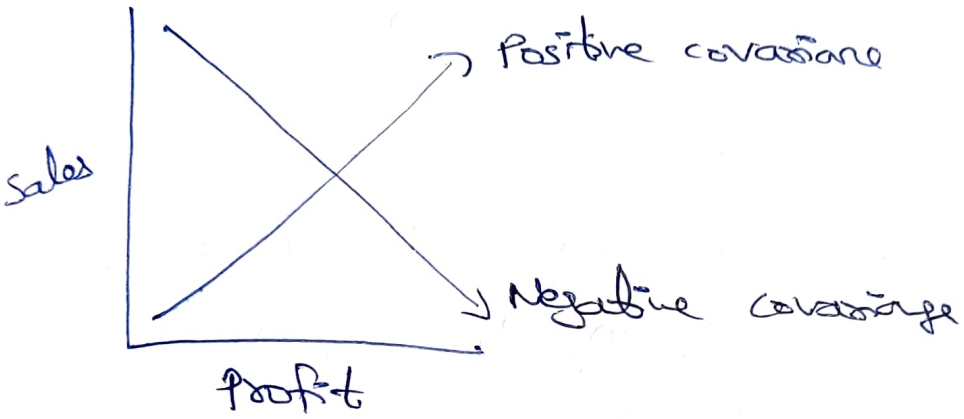
* Positive Covariance: Means that both variables tend to be high or low at the same time.

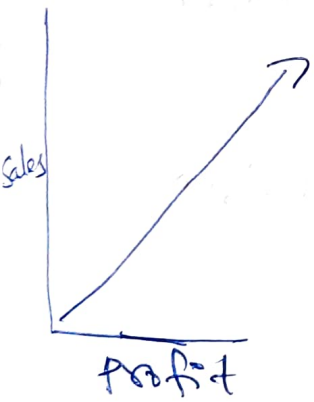
* Negative covariance: Means that when one variable is high, the other tends to be low.

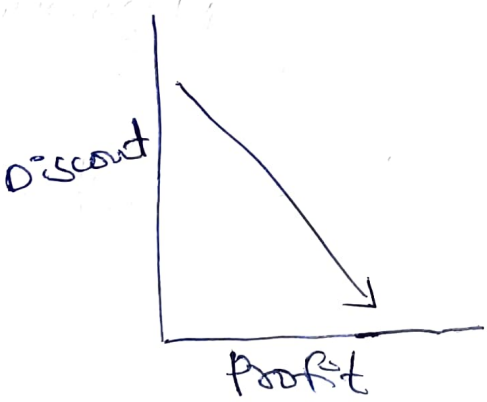
Covariance Formula :-

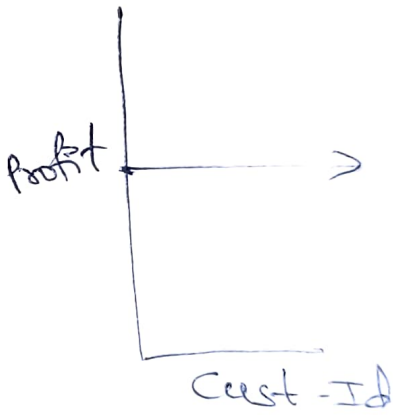
For Population :

$$Cov(X,Y) = \frac{\Sigma(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$









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

$\bar{x}$ = mean of x data

$\bar{y}$ = mean of y data

Sales (X)	Profit (Y)
150	10
170	25
190	35
210	60
240	80

$X_i = 1, 2, 3, 4 \ldots$ $\quad x_1 = 150, x_2 = 170, x_3 = 190 \ldots$

$y_1 = 10, y_2 = 25, y_3 = 35 \ldots$

	$(x_i - \bar{x})$	$(y_i - \bar{y})$	
①	−	+	} → Negative Relationship
②	+	−	
③	+	+	} → Positive Relation
④	−	−	

+ve Positive −ve Negative

If Sales ↑ Profit ↑ → positive relationship

Sales ↓ profit ↓ → " "

Sales ↑ profit ↓ → Negative "

Sales ↓ profit ↑ → Negative "

$\Rightarrow \bar{X} = 192, \quad \bar{Y} = 42.$

$\Rightarrow$
$(150-192)(10-42) = -42 * -32 = +$
$170-192 = -22 * 25-42 = -17 +$
$190-192 = -2 * 35-42 = -7 +$
$210-192 = 18 * 60-42 = 18 +$
$240-192 = 48 * 80-42 = 38 =$

$(-42 * -32) + (-22 * -17) + (-2 * -7) + (18 * 18) + (48 * 38)$

$= 3880$

$\Rightarrow \frac{3880}{5} = 776$ (positive Relationship)

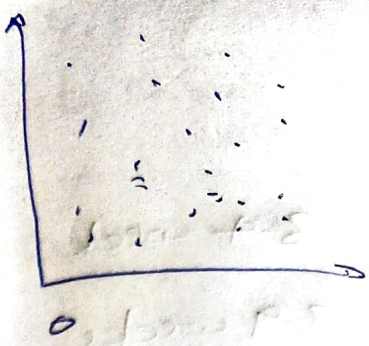
Correlation (Pearson correlation coefficient)-

Measure the direction and strength and Direction of relationship bw two variables

- Positive Corelation
- Negative Corelation
- No corelation.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \Rightarrow$$

S_{xy} is covariance
S_x & S_y are the standard deviations



Correlation coefficient r is number between -1 to $+1$ and tells us how well a regression line fits the data

$$Corr = \frac{Cov(X,Y)}{SD(X) * SD(Y)}$$

-1 to $+1$

$(-1$ to $+1)$

If a number is close to $+1 \rightarrow$ +ve relationship
close to $-1 \rightarrow$ -ve relationship
close to $0 \rightarrow$ weak relationship