Advance Statistics

① Sample vs Population

	Sample	Population
	n	N
numbers →	Statistics (a subset of population)	Parameters (collection of all items of interest)

Why prefer sample
- → easy to manage
- → less time consuming
- → less costly

A sample has to be random !

↓

(Each member is equally likely to be chosen)

② Covariance

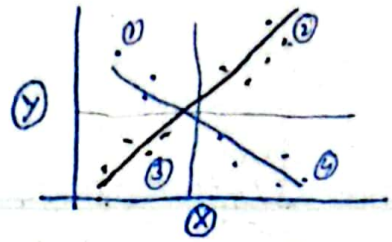
$$Cov(x, y) = E\left[(x - E(x)\, x\, (Y - E(Y)\right]$$

$$Cov_{x,y} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

(It measures the direction of relationship b/w two variables) (+ive, -ive, no

① & ④ → -ive Cov values

② & ③ → +ive Cov values



→ Covariance value doesn't tell us about steepness of the line or how close the data points are to the line.
(It just tells us about nature of relationship)

→ Cov value can be changed even if relationship doesn't (i.e by changing scale value).

→ Used for PCA (Principle Component Analysis)

→ Just like correlation but its values can be varied more than (-1 to 1)

---

③ Pearson's Correlation (γ)

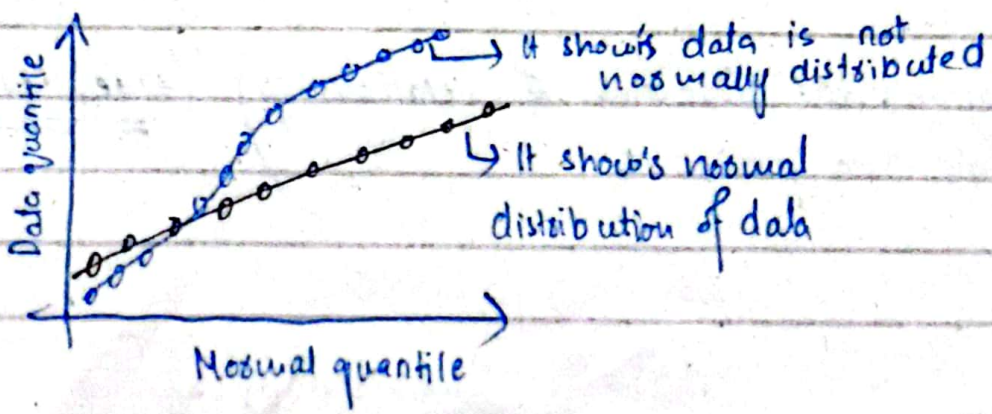
- "Explain relationship b/w two numerical variables"
- Value → (~~1~~ −1)
- Doesn't depend on scale of data
- Educated guesses ∝ Amount of data points
- Educated guesses ∝ 1/p_value

$$\gamma = \frac{Cov(x,y)}{(S_x \cdot S_y)}$$

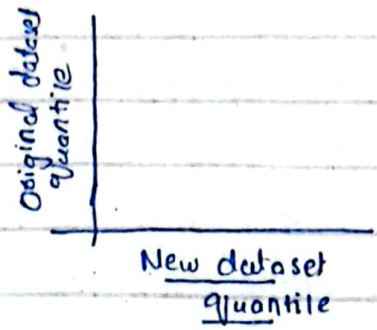
→ this denominator value ensures that the γ value becomes independent of scale.

---

④ Quantile - Quantile Plots (Q-Q plots)



We can also compose two data sets:



---

⑤ Confidence Interval (P-value)

Value → 0.05 or 5% (universally accepted)

understanding through bootstrapping

→ Take 15 books and weigh them
→ Calculate mean
→ Now take random sample of these 15 books which is again is in number (double or tripple pick is ok!)

It is done to estimate actual mean of book weight of all available books in shop.

→ Bootstrapping # > 10,000

Now here p-value of 0.05 or confidence interval of 95% means

"All the space values in which 95% of the mean's are present"

⇒ Important in hypothesis testing ✓

(vi) Null hypothesis / Hypothesis testing

(vii) T-test / Chi-square test / Anova / → When to use ?

One categorical data set → One sample proportion test
if it gives value less than p-value of 0.05 then
→ reject null hypothesis and accept alternative Hypothesis

Two categorical data set → Chi-square test

One numerical data set → T-test

Two numerical data set → Correlation test (r)

One numerical & One or two categorical data set → ANOVA
(further subsets present in data)

If,
$P < \alpha$ H_0 is rejected

If,
$P \geq \alpha$ H_0 is accepted