ANOVA

Analysis of Variance

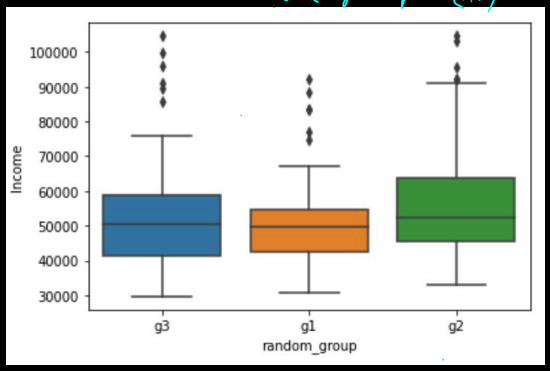
	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	18	Male	14	Single	3	4	29562	112
1	KP281	19	Male	15	Single	2	3	31836	75
2	KP281	19	Female	14	Partnered	4	3	30699	66
3	KP281	19	Male	12	Single	3	3	32973	85
4	KP281	20	Male	13	Partnered	4	2	35247	47

Income (Num) Vs Gender (Categorical - 2 categories) → T-test

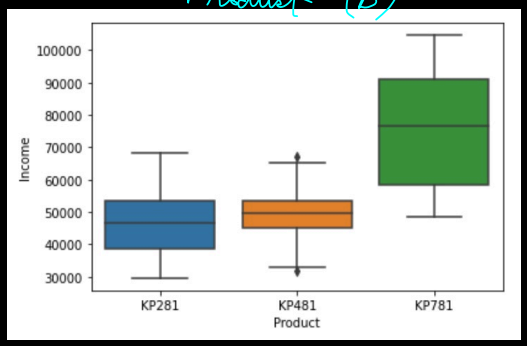
Gender (Cat) Vs Product (Cat) → Chi-Squared Test

Income (Num) Vs Product (Categorical 2 or more categories) → ANOVA

Random group (A)



Product (B)



H_0: all means are equal → high p-value

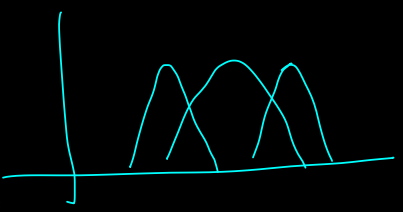
H_a: Atleast one mean is different from other → low p-value

ANOVA : deep dive Height

Setup 1

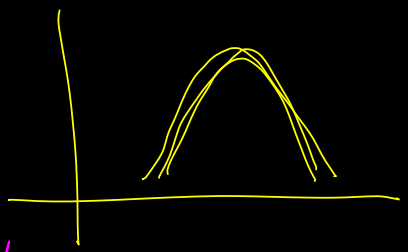
American Basketball players
Indonesian college students
Indian cricket team

① variance within group
② variance b/w group



Setup 2 group people alphabetically

A to G
H to N
O to Z



$$f\text{-ratio} = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

f-ration will be high in setup 1 $\longrightarrow$ H_a $\longrightarrow$ right tailed
will be low in setup 2 $\longrightarrow$ H_0

Assumptions of ANOVA:

1. Data should be Gaussian → Q-Q plot, Shapiro
2. Independence
3. Equal variance in each group → Levene

When assumptions of ANOVA don't hold we use Kruskal-Wallis

QQ Plot: Test if data is Gaussian

$x_1, x_2, x_3, \ldots, x_{10000}$

Empirical rule? 68 / 95 / 99
1σ / 2σ / 3σ

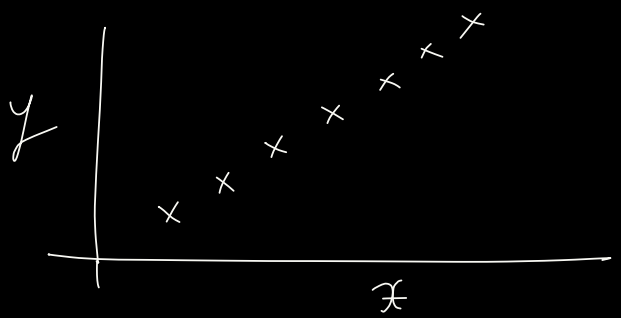
$y_1, y_2, y_3, \ldots, y_{1000}$ → actually Gaussian

1st percentile of x ≃ 1st percentile of y

2nd percentile of x ≃ 2nd percentile of y

⋮

100th ≃ 100th percentile of y



Shapiro: Test whether data is Gaussian or not

H_0: Data is Gaussian

Post Read (Optional)

iPhone sales in 3 stores

	A	B	C	
	25	30	18	
	25	30	30	
	27	25	29	
	30	24	29	
	23	26	24	
	20	28	26	
	25	26.5	26	25.83
	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}$

H_0: All means are equal H_a: Means are different

Step 1 Compute individual group means $\bar{Y}_1 = 25$ $\bar{Y}_2 = 26.5$ $\bar{Y}_3 = 26.5$

Step 2 Compute mean of these 3 values $\bar{Y} = \frac{25 + 26.5 + 26}{3} = 25.83$

Step 3 Between groups

$$SSB = 6(25 - 25.83)^2 + 6(26.5 - 25.83)^2 + 6(26 - 25.83)^2 = 6.9$$

DF = 3 - 1 = 2

$$MSB = \frac{SSB}{DF} = \frac{6.9}{2} = 3.49$$

Step 4 Within groups

$$SSW = (25 - 25)^2 + (25 - 25)^2 + (27 - 25)^2 + \cdots + (20 - 25)^2$$
$$+$$
$$(30 - 26.5)^2 + (30 - 26.5)^2 + (25 - 26.5)^2 + \cdots + (28 - 26.5)^2$$
$$+$$
$$(18 - 26)^2 + (30 - 26)^2 + (29 - 26)^2 + \cdots + (26 - 26)^2$$
$$= 223$$

DF = 18 - 3 = 15

$$MSW = \frac{SSW}{DF} = \frac{223}{15} = 14.9$$

$$F = \frac{3.49}{14.9} = 0.23$$

$$F = \frac{MSB}{MSW}$$