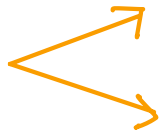


# Numerical - Categorical Relationship

⑤ Tests  T-test  
Z-test

⑤ Constraint on Num - Categorical  
⑤ num-categories = 2

**Questions** What test would you use for following Variable pairs

① Product vs Gender  
chi - square

② Income vs Gender  
T-test / z-test

③ Product vs Income?  
( > 2 Categories) (Numerical)

⑤  $n \geq 2$  T-Tests ( $n \rightarrow$  num-categories)

⑤ One Way ANOVA

# One-Way ANOVA Test

- \* One  $\Rightarrow$  A single categorical Variable (Ex: Product)
- \* Way  $\Rightarrow$  Dependent Variable (Ex. Income)
- \* ANOVA  $\Rightarrow$  Analysis of Variance
- \* Uses F-Distribution to generate F-statistic / F-ratio
- \* This test is suitable when dealing with

Numerical Variable

vs

Categorical Variable  
(num-categories  $> 2$ )

- \* Ass<sup>n</sup>: Is there a significant Difference between Income of Different product buyers

$H_0$  : Means across all groups are Same

$H_a$  : Atleast mean of one Group is Different

In Anova

① Analys Variance of Data using F Distribution

② Two types  $\begin{cases} \text{Variance Between the Groups} \\ \text{Variance Within the Groups} \end{cases}$

F-ratio

$$\frac{\text{Variance Between the Groups}}{\text{Variance Within the Groups}}$$

F-ratio

F statistic

$$F\text{-ratio} \propto \frac{1}{p}$$

\* Interpretation

① F-score is  $\approx 1$  (if Both within and Between Variances are close to each other)

② Population is probably common for all Group/Samples

② F-score  $\gg 1$

↳ Samples might be from Population with Different means

## Assumptions of ANOVA

- ① Normality of Data (Residual)
- ② Homogeneous Variances (Residual)  
(Variances across groups should be approximately Equal)
- ③ Independent Observations

---

## Questions

- ① Do we have Alternative when Data is Not Gaussian?
- ② How do we check if Data is Gaussian?

# Kruskal-Wallis Test

$H_0$  : All groups have same population median

$H_a$  : There is a significant difference in medians of at least two group

## Advantages

① No assumption of Normality

② Robust to Outliers

## Limitation

① Less powerful compared to ANOVA when assumption are met.

② How do we check if Data is Gaussian?

Ans ① Plot Histogram

② Plot QQ-plot

③ Shapiro Wilkins Test

$H_0 \Rightarrow$  Data is Gaussian

$H_a \Rightarrow$  Data is Not Gaussian

$\alpha \Rightarrow 0.05$

test-stat, pvalue  $\Rightarrow$  Shapiro(data)

# Levene-Test

Used for checking if variances in 2 samples are equal or Not

$H_0$  : Variances are Equal i.e.  $\sigma_1^2 = \sigma_2^2$

$H_a$  : Variance are not Equal

- ↳ Returns a p-value and a Test-statistic
- ↳ P-Value can be compare with  $\alpha$  to reject or fail to Reject:  $H_0 (\sigma_1^2 = \sigma_2^2)$