Anova

1) Recap
2) Anova → One way Anova
3) Assumption for Anova
4) Kruskal Wallis Test
5) QQ Plot
6) Levene Test
7) Shapiro Test

⇒ Hypothesis Testing →

⇒ Checking if the results (sample data) we are seeing is significantly different than expected value.

→ Z test → • data is normal
• Chance that data lies in extreme region

→ T test → • Maths equation → changed from Pop to Samp

→ Chi-Square → Categorial variable.
→ As data is Categorial → # of Values

⇒ Exp dist^{n}/ data ⟷ Compare the diff^{n} Observed data

$$\chi^2 = \sum \frac{(Exp - Act)^2}{Exp}$$

χ^2 being large → Wrong Hypothesis
Reject Hypothesis

⇒ Exp → 1) Equal prob
2) Based on Hist dist^n

Independent Variable (Cat) →

	A	B	C	D	Total
⇒ 3 pets → Cat	10	20	30		
→ Dog	20	40	60		
→ Horse	30	60	90		

Anova → T Test → 2 Samples at a time.

Aerofit → 3 products → Income

KP781	10L	12L	30L	40L ...
KP781	8L	10L	11L	- - -
KP581	7L	12L	- - - - - -	

⇒ KP281 KP481 KP781

KP481 KP581 KP581

⇒ n dist Cat → $^nC_2$ test to perform

⇒ Higher Chances of Error → as each test will have their own error

Why not to do pairwise T-Test

A	62	65 → −3
B	65	65 → 0
C	68	65 → 3
	Real	pred

error is 0

I only get 1/3 Conv

⇒ ① T Test → 2 sample

χ^2 Test → 2 Categorial variab [Cat vs Cat]

Anova → >2 Sample

Num vs Cat

↳ 2 val

TTest

⇒ Gender (2 value Cat) vs Income (Num) → T-Test

⇒ Gender (Cat) vs Product (Cat) → X^2

⇒ Product (3 value Cat) vs Income (Num → Anova

---

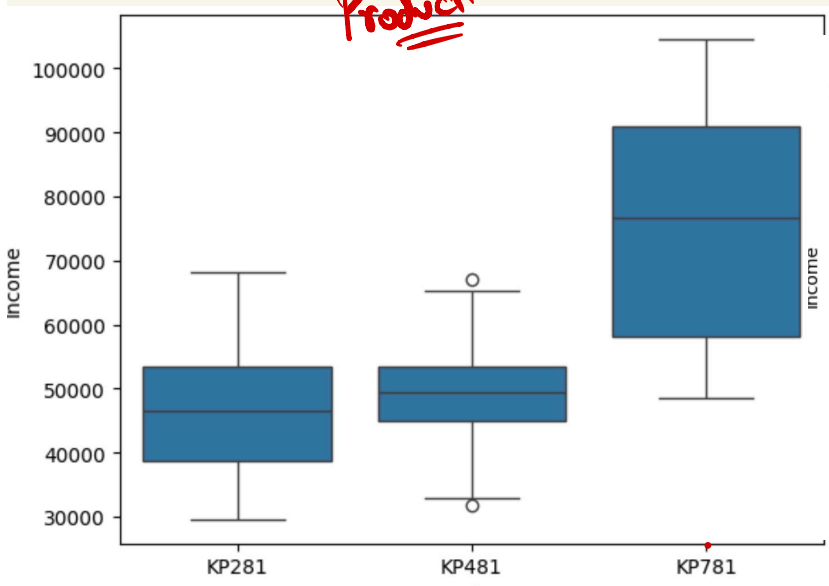
1) H0 → They are independent / no impact

2) Dist^n + Tail + p-value → to calculate Stat

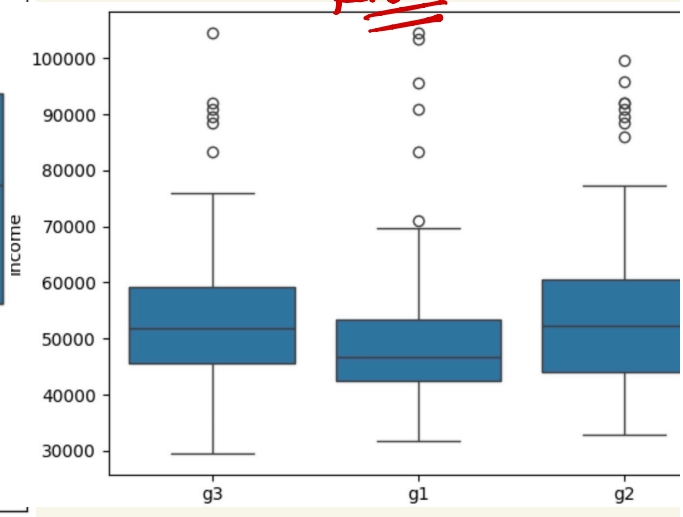
⇒ Cat^4 Product vs Income

→ Case2 divide dat in random 3 groups and Compare in Cond.

Product



Random



A Qty which changes when Act data differ from exp.

⇒ Anova → Analysis of Variance.

$$\text{F-ratio} = \frac{\text{Variance b/w groups}}{\text{Variance within groups}}$$

Varian b/w group → is higher when there is a lot of dependance on group.

Variance within group → should be higher for random group.

⇒ 14 A → 18-20

23 B → 20-25

29 C → 27-30

A → 12 — 30

B → 12 — 30

C → 12 — 30

	Case 1	Case 2
Variance within group	→ Low	High
Variance b/w groups	→ Same	Same

⇒ Var within the group is Lower for dependent case.

Var b/w the group is higher for dependent case.

$$f\ ratio = \frac{\text{Var b/w the group}}{\text{Var with the group}}$$

f ratio is higher when data is dependent on group.

f ratio is above a threshold → reject H_0

↓

High f ratio → Low p-value → reject H_0

Assumptions of Anova →

1. Data is gaussian
2. Data is independent across each record
3. Equal variance in diff^{n} groups

If assumptions of Anova fails → Kruskal Wallis Test

⇒ Kruskal →

H_0 → Population median of all groups is same.

=

H_a → at least 1 median is different.

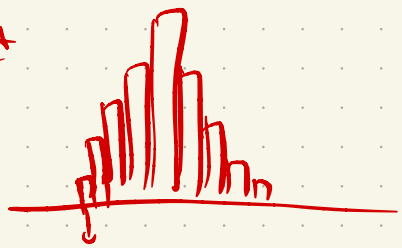
One-way ANOVA (Analysis of Variance) assumes the following:

Homogeneity of variances:
This assumption states that the variances within each group being compared are roughly equal.
Normality of data:
This assumption refers to the distribution of residuals or errors being approximately normal.
It does not necessarily require the raw data to be normally distributed.
Independence of observations:
This assumption means that the observations within each group are not related or dependent on each other.
While equal sample sizes are desirable for one-way ANOVA, they are not always a strict assumption.

① Normality Test → If data is Gaussian or not.

A) QQ plot → Quantile Quantile plot.

⇒ plot a histogram of your data
+
plot a histogram of ideal Normal Curve.

plot 

Compare each bar → normal bar

each Quantile → normal Quantile

0-5%
5-10%
10-20%
.
.

divide data in 100 parts →

⇒ (μ, σ) →

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$ ←

⇒ Data → get μ, σ → Plot a Normal curve for this (μ, σ)

Shapiro Test →

H_0 → Data is Gaussian

H_a → Data is not Gaussian.

⇒ To check normality → QQplot, Shapiro Test

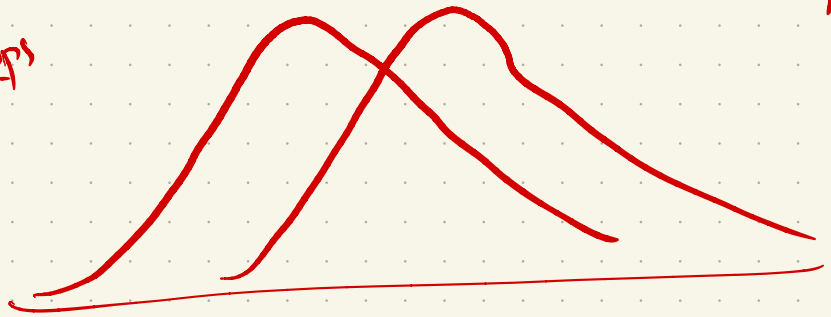
⇒ Equal Variance →

Levene Test → If there is equal variance in diff groups.

H_0 → Have same variance

H_a → Have diff variance.

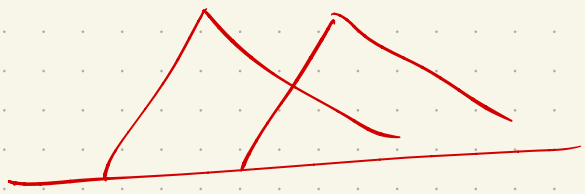
2 groups

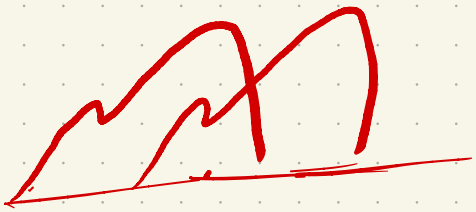


Normal vs observed

group1 vs group2

Quantile





⇒ Anova → More than 2 Cat and

You want to compare impact of Cat on a numerical Column.

→ F-Stats ⇒ $\frac{\text{Var b/w group}}{\text{Var within group}}$

F ↑ p-value ↓ → data is dependent on group

→ Assumption → Normal
→ equal variance
→ independent

→ If Assumption fail → KW Test

→ Normality → QQplot
→ Shapiro → Shapiro Wilks

→ Equal var → levene Test

Doubts

1. → 2 Test → 7600 problem

⇒ H_0 = Steps = 8000

H_a = Steps ≥ 8000

Right tailed test

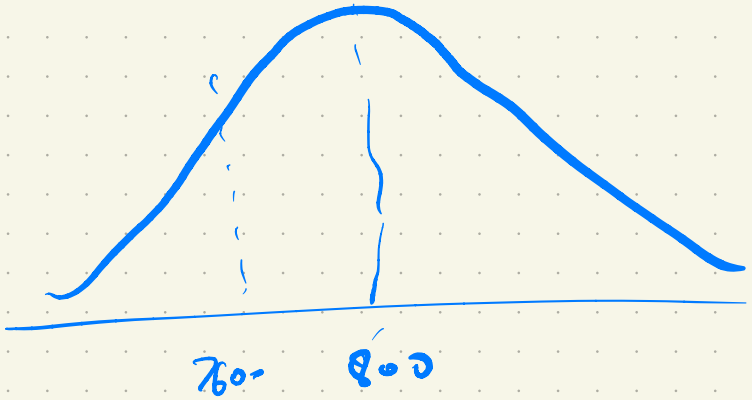
⇒ Defining our dist^n → Normal dist^n → (μ, σ)

→ Population

→ SM μ

σ → CLT

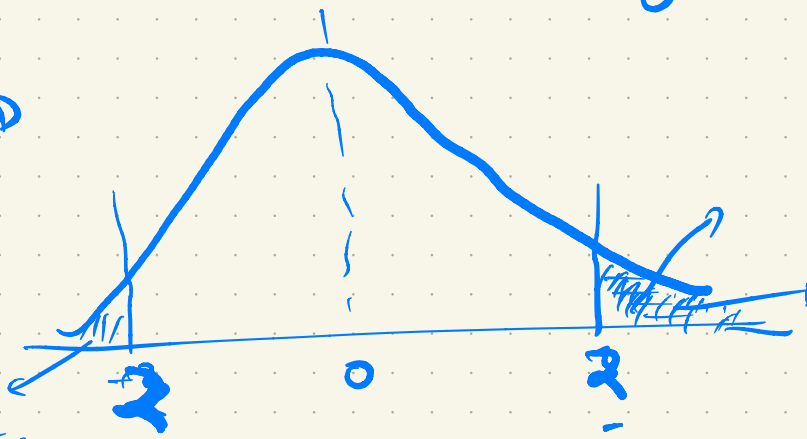
⇒ H_0 be true → $N\left(8000, \frac{\sigma}{\sqrt{n}}\right)$



p_value
↳ data lying on right side of observed

⇒ Define dist with assuming H_0 is true.

⇒



5%

1-norm.cdf(2)

norm.cdf(-2)

5%

p_value = 2 × (1 - norm)

height_women

p value $\neq$ norm.cdf(#.74) − norm.cdf(#1.74

p. val = $\left(1 - \text{norm.cdf}(1.74)\right) + \left(\text{norm.cdf}(-1.74)\right)$

Q 60% → 1 car
28% → 2 car
12% → 2+ car

---

73 owns → 1 car ←
38 owns → 2 car ←
18 own → 2+ car ←
— 129

60 28 12

73/129 38/129 18/129

60/100 129 28/100 129 12/100 129

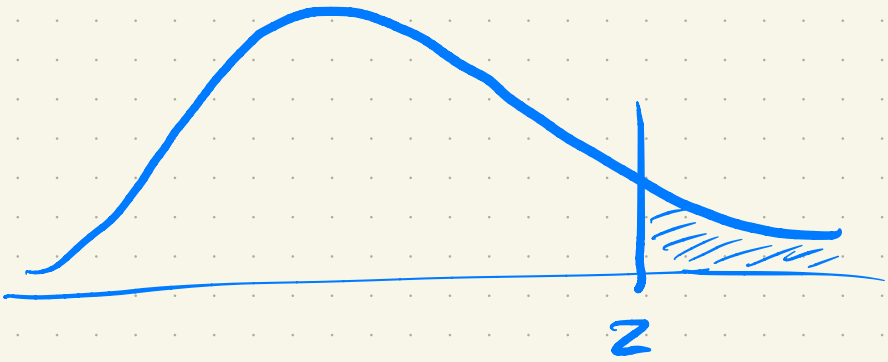
73 38 18

⇒ Critical value → p-value Comparing it against α.

Values associated with p-value = α.

→ SM datat → Z score → p-value

p-value = α → Critical Z score
↓
Critical SM.



If Right tail $\alpha = 0.05$ → 1 - norm.cdf(z) = 0.05

z = norm.ppt(0.95)
↓
Critical z.

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{\bar{x} - \mu_{pop}}{\sigma_{pop}/\sqrt{n}}$$

$\bar{x}$ → Critical SM

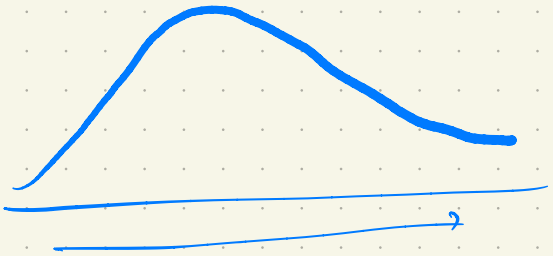
⇒ power of Test

$1 - \beta$

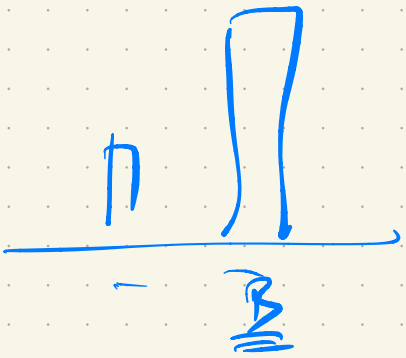
β ↑ power ↓

β → false negative error.

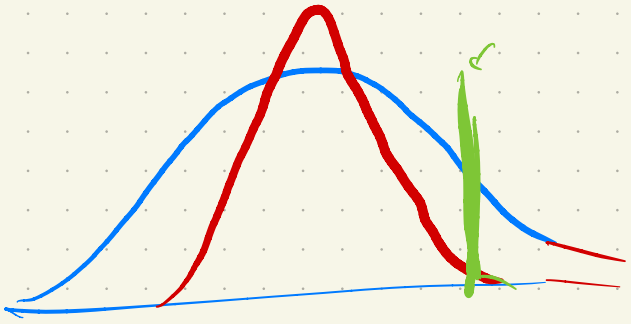
⇒ Std dev



A



β



STD ↓ β ↓ power ↑

If everyone is innocent

$\alpha = 0.02$ ⇒ 98 people →

2 criminal →

$\beta = 0$

$\alpha =$

98	0
2	0

If everyone is criminal

98 innocent
2 criminal

0	98
0	2

→ $\beta = 0.98$

$\beta = 1 - \alpha$

⇒

92	3
2	3

CLT → Std dev of SM dist^n = $\frac{\sigma_{pop}}{\sqrt{n}}$

mean of SM dist^n = μ_{pop}

→ A → 30 → $\mu_A = 100$

→ B → 30 → $\mu_B = 120$

180
170
300

⇒ A → [1, 30, 40, . . .] ← 300 → $\mu_A = 95$

B → [5, 27, 28, . . .] ← 300 → $\mu_B = 48$

https://colab.research.google.com/drive/11syQehH05QZwzsEkuXOx2af8WuXzPHkC?usp=sharing

Collar link : https://colab.research.google.com/drive/11syQehH05QZwzsEkuXOx2af8WuXzPHkC?usp=sharing