Stats.

1. Chi Square Test
2. ANOVA TEST (F Test).

Chi Square Test

* Chi Square Test claims about population proportions.

It is a non parametric test that is performed categorical (nominal or ordinal) data.

↳ Rank

* In the 2000 U.S Census the ages of individuals in a small town were found to be the following.

<18	18-35	>35
20%	30%	50%

In 2010, ages of $n = 500$ individuals were sampled. Below are the results.

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude that the population distribution of ages has changed in the last 10 years?

Ans)

	<18	18-35	>35
Expected	20%	30%	50%

←

n=500

	<18	18-35	>35
Observed	121	288	91
Expected	100	150	250
		200	80

E

2000

① Null Hypothesis H_0 : The data meets the expected distribution

H_1 : The data do not meet the " "

② $\alpha = 0.05$ C.I = 95%

n=500 → Sample

③ Degree of freedom

$df = K-1 = 3-1 = 2$

CHI SQUARED

2 Tailed Test

④ Decision Boundary ⇒ 5.991

$\{ \chi^2 > 5.991$ Reject the Ho.

⑤ Chi Square Test statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$\chi^2 = 232.494$

$\chi^2 > 5.99$ Reject the H_0.

(e) 500 elementary school boys and girls are asked which is their favourite color : blue, green or pink. Results are shown below:

	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	230
	120	180	200	500

(2-1) (1) ⇒ (3-1) = 2

Using $\alpha = 0.05$, would you conclude that there is a relationship between gender and favourite color?

(f) Null hypothesis H_0: Gender and favourite color are related. Reject.

H_1 : Not related.

(g) $\alpha = 0.05$ C.I = 95%

(h) Degree of freedom

df: (Rows - 1) (Columns - 1)

= (2 - 1) (3 - 1)

= (1) (2) = 2

(i) Tut Stats = 5.991.

⑥ Chisquare Test Statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Observed	Blue	Green	Pink	
Boys	100	150	20	270
Girls	20	30	180	230
	120	180	200	500

$\Rightarrow (3-1) = 2$

$f_e = f_c \, f_r$ / f_n

Expected	Blue	Green	Pink	
Boys	64.8	97.2	108	270
Girls	55.2	82.8	92	230
	120	180	200	500

$$= \frac{(100-64.8)^2}{64.8} + \frac{(150-97.2)^2}{97.2} + \frac{(20-108)^2}{108} + \frac{(20-55.2)^2}{55.2} +$$

(Boy, Blue) = (120 × 270)/500 = 64.8 //

(Boy, Green) = (180 × 270)/500 = 97.2

(Boy, Pink) = (200 × 270)/500 = 108

(Girl, Blue) = (120 × 230)/500 = 55.2

(Girl, Green) = (180 × 230)/500 = 82.8

(Girl, Pink) = (200 × 230)/500

$\chi^2 = 259.78 > 5.99$ Reject the H0.

{python}

⑦ ANOVA Test (F Test) ⟶ {python}

Analysis of Variance

{ CHI SQUARE AND ANOVA } Stats { python }

Interviews Questions

C.I

① Size of all the sharks in the world.?

n = 40 | $\bar{x}$ = 5m | S = 0.5m

n=25

$\alpha = 0.05$

n > 30 → z test

Z_test .or t_1

t test

$t_{\alpha/2}$

$$\bar{x} \pm \text{Margin of Error}$$

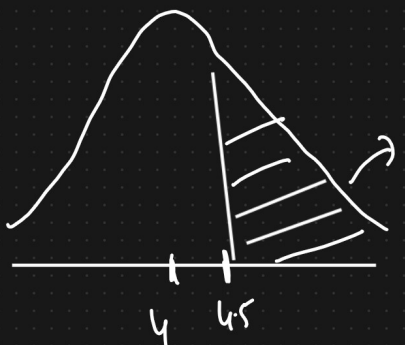
$$\bar{x} \pm Z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right).$$

Amazon

Point Estimate

Lower fence ⟷ Higher Fence

④



r = 0.5

Correlation

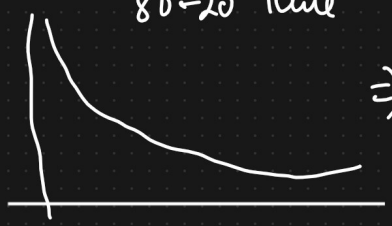
② Pearson & Spearman Rank Correlation

⇓ Linear

⇓ Non Linear Data.

(iii) Oil Distribution in a place follows which distribution.

80-20 Rule

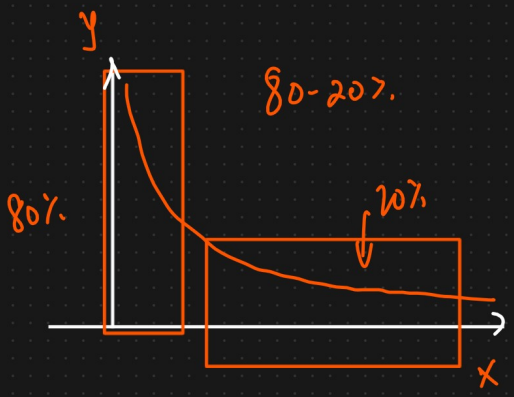


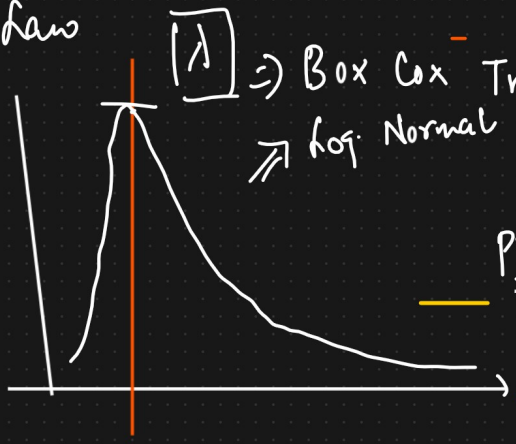
⇒ Pareto distribution

⇓

Power Law

Qatar





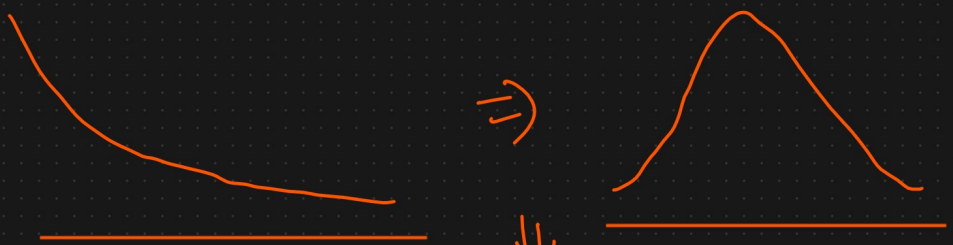
[λ] ⇒ Box Cox Transform

⇗ log Normal

Python

Eg:

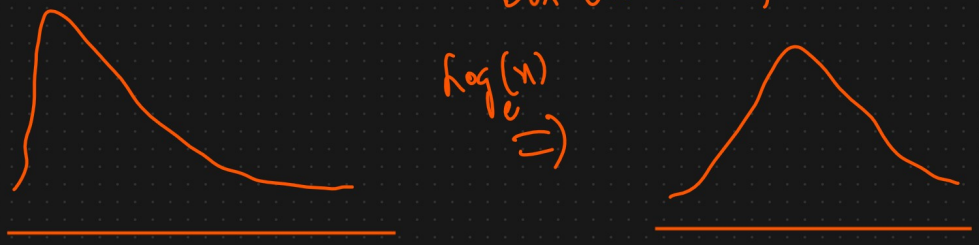
(iv)



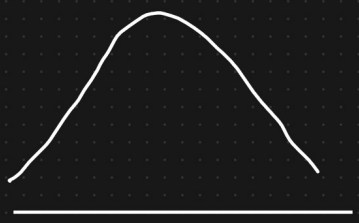
⇓

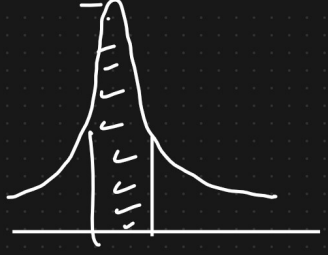
Box Cox Transformation.

(v)



(vi) What is the disadvantage of replacing nan values with mean?





Null values

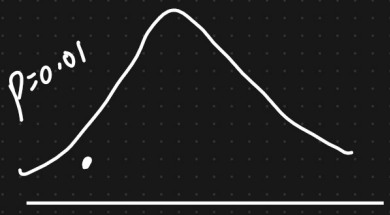
(#) Outliers : 5 Number Summary, Box plot.

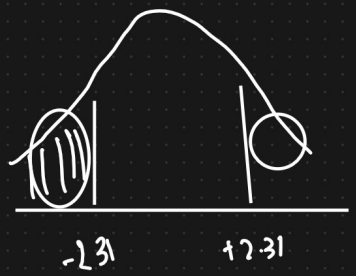
(#) Z-score

(#) Why Sample variance by n-1 ?

Bessels Correction Degree of freedom

(#) 0.1 ⇒ p-value ✓.





p-value is the probability of the null hypothesis to be True

(#) t test vs z-test

Bernoulli Distribution { Binary }. outcome.

2 outcomes

Tossing coin ⇒ H or T

⇓

Binomial distribution

P(H) = 0.5 P(T) = 0.5

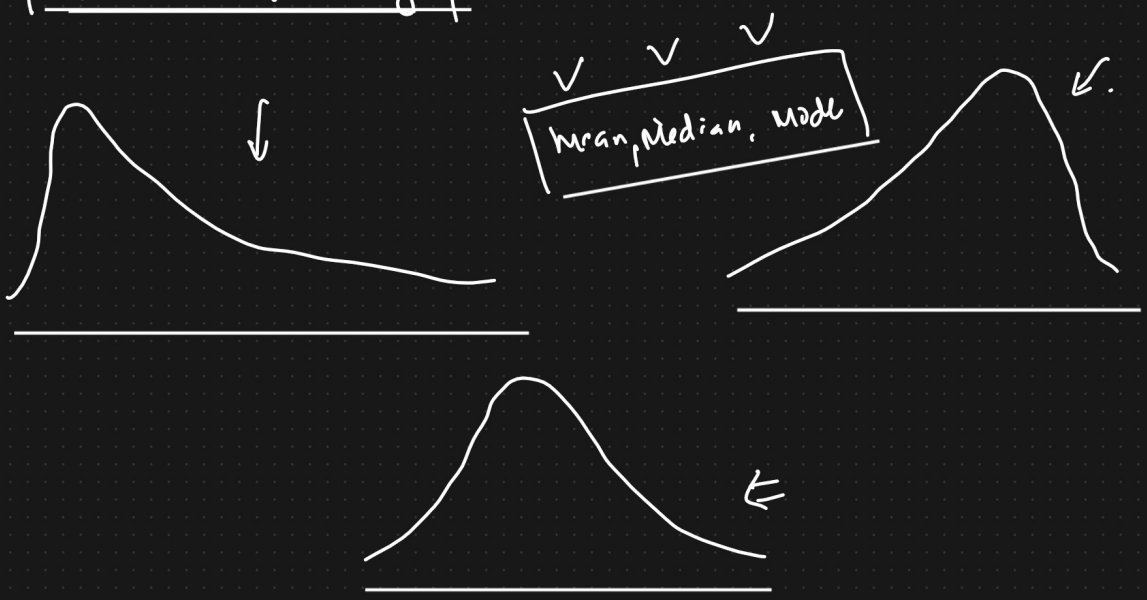
P q = 1-p

Many Bernoulli experiments

Tossing a coin 10 times

{ P P P P P P / q q q q q q }

Python Programming



Z-test for Proportion, Two Samples

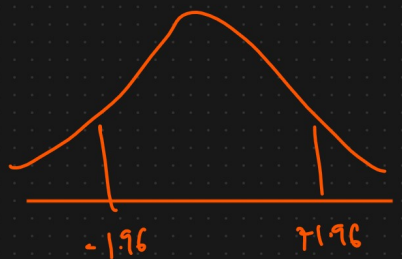
① Researchers wants to test the effectiveness of a new anxiety medicines. In clinical testing, 64 out of 200 people taking the medicine report symptoms of anxiety. Of the people receiving a placebo, 92 out of 200 reports symptoms of anxiety. Is the medication working any different than the placebo? $\alpha = 0.05$ Test the claim?

1) Null Hypothesis $H_0 = \boxed{p_1 = p_2}$

$H_1 = \boxed{p_1 \neq p_2}$

② $\alpha = 0.05$ $CI = 95\%$

③



④ Calculate Test Statistic

$n_1 = 200$

$n_2 = 200$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$\hat{p}_1 = \frac{64}{200} = 0.32$

$\hat{p}_2 = \frac{92}{200} = 0.46$

$= -2.869.$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{64 + 92}{200 + 200} = 0.39$$

$-2.869 < -1.96$ Reject the Null Hypothesis