e.g. Dataset

Input Features

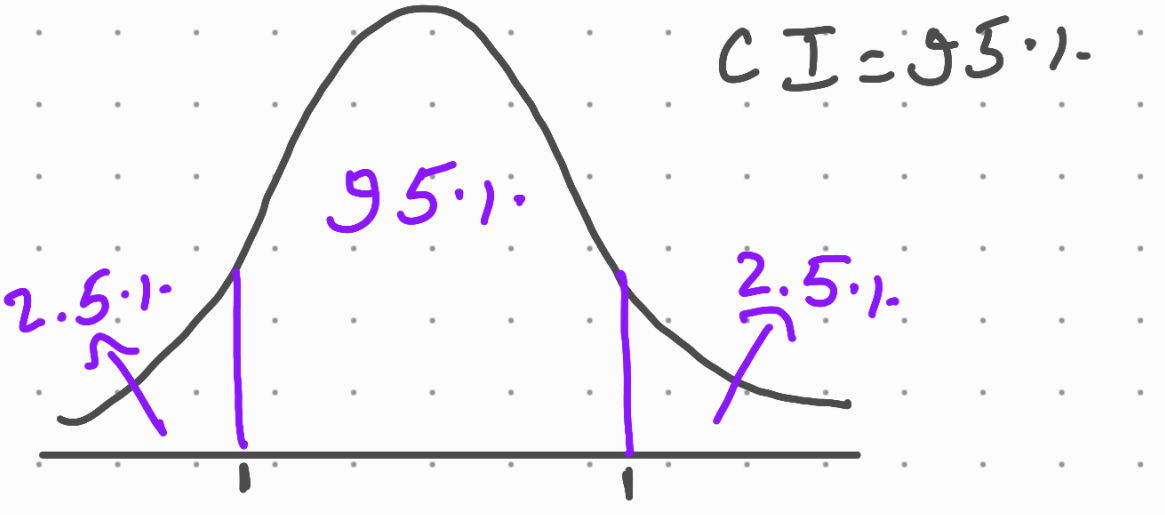
Output Feature

Size of House	No. of Rooms	Location	Price
x_1	x_2	x_3	y

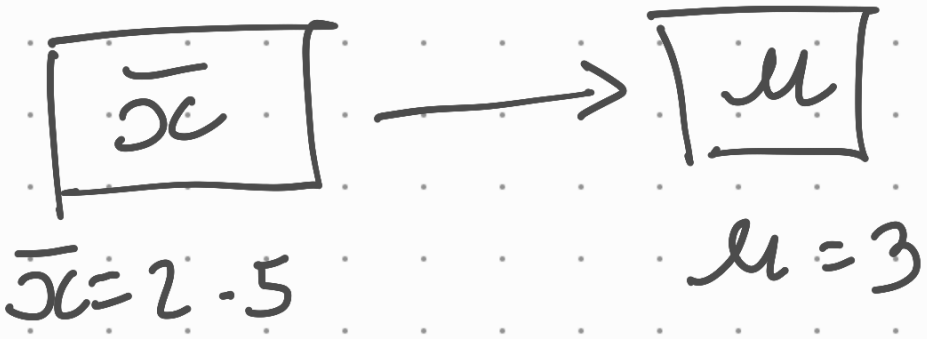
$$P_s(y/x_1,x_2,x_3) = \frac{P_s(y) * P_s(x_1 x_2 x_3 / y)}{P_s(x_1,x_2,x_3)}$$

↳ Bayes' Theorem

Confidence Interval and Margin of Error



Point Estimate



Confidence Interval

Point Estimate $\pm$ Margin of Error

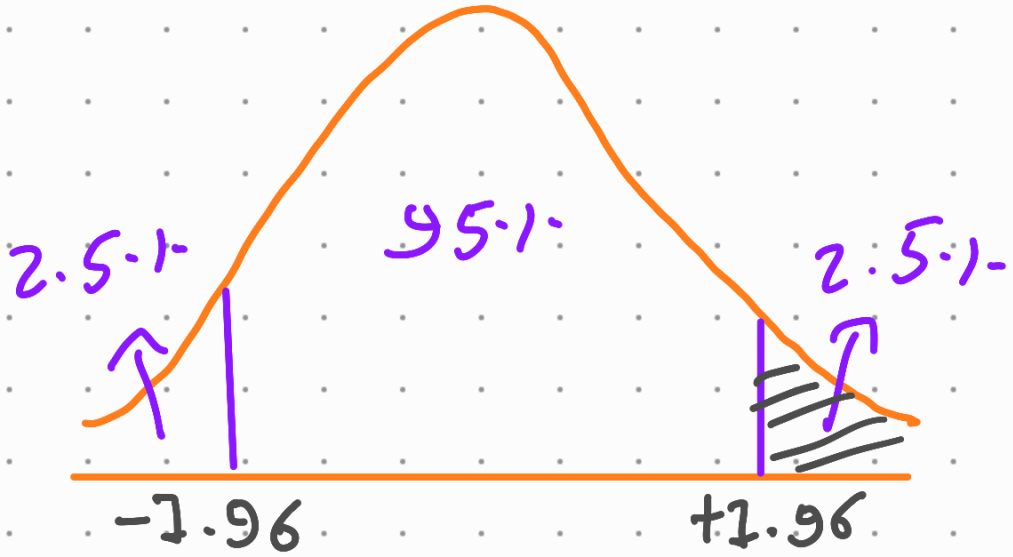
$$Z\text{-Test} \Rightarrow \bar{x} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$T\text{-Test} \Rightarrow \bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

e.g. On the verbal section of exam, the standard deviation is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% C.I. about the mean.

$\alpha = 0.05$

$\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



$1 - 0.025 = 0.9750$

$+1.96 \Rightarrow$ z-table

Lower CI

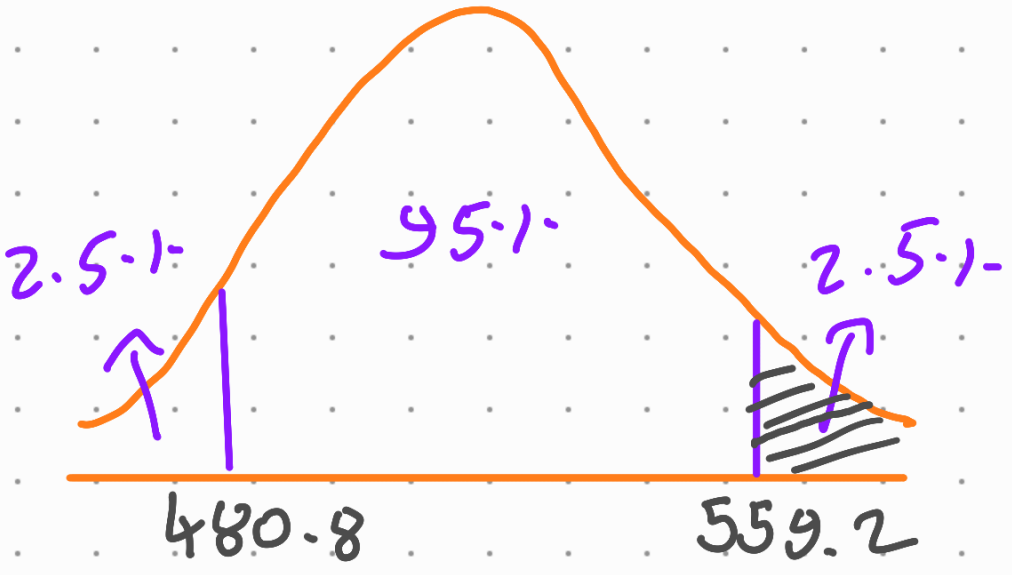
$$= 520 - (1.96) * \frac{100}{\sqrt{25}}$$

$$= 480.8$$

Higher CI

$$= 520 + (1.96) * \frac{100}{\sqrt{25}}$$

$$= 559.2$$



Conclusion: I am 95% confident about the mean score is between 480.8 and 559.2.

CHI Square Test

→ The Chi Square Test for Goodness of Fit

→ This test claims about the population proportions.

→ It is a non-parametric test that is performed on categorical Data

→ Categorical Data [Ordinal and Nominal] data.

e.g. (1) There is a population of mule who likes different Color bikes.

	Theory	sample
Yellow Bike	1/3	22
Red Bike	1/3	17
Orange Bike	1/3	59

Theory categorical Distribution

Observed categorical Distribution

✱ Goodness of Fit Test

e.g. ① In a Science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of people are left handed.

→

	O	E
Left Handed	11	9
Right Handed	64	66
	75	75

② In 2010 census of the city, the weight of the individuals in a small city were found to be the following.

<50kg	50 - 75	>75kg
20%	30%	50%

In 2020, weights of $n=500$ individuals were sampled. Below are the results.

<50kg	50-75	>75kg
140	160	200

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

→ 2020 Expected

<50kg	50-75	>75kg
20%	30%	50%

2020 Observed $n=500$

<50kg	50-75	>75kg
140	160	200

	< 50	50 - 75	> 75
Expected	0.2 × 500 = 100	0.3 × 500 = 150	0.5 × 300 = 250

(a) Null Hypothesis H_0 = The data meets the expectations

(b) Alternate Hypothesis H_1 = The data does not meet the expectations

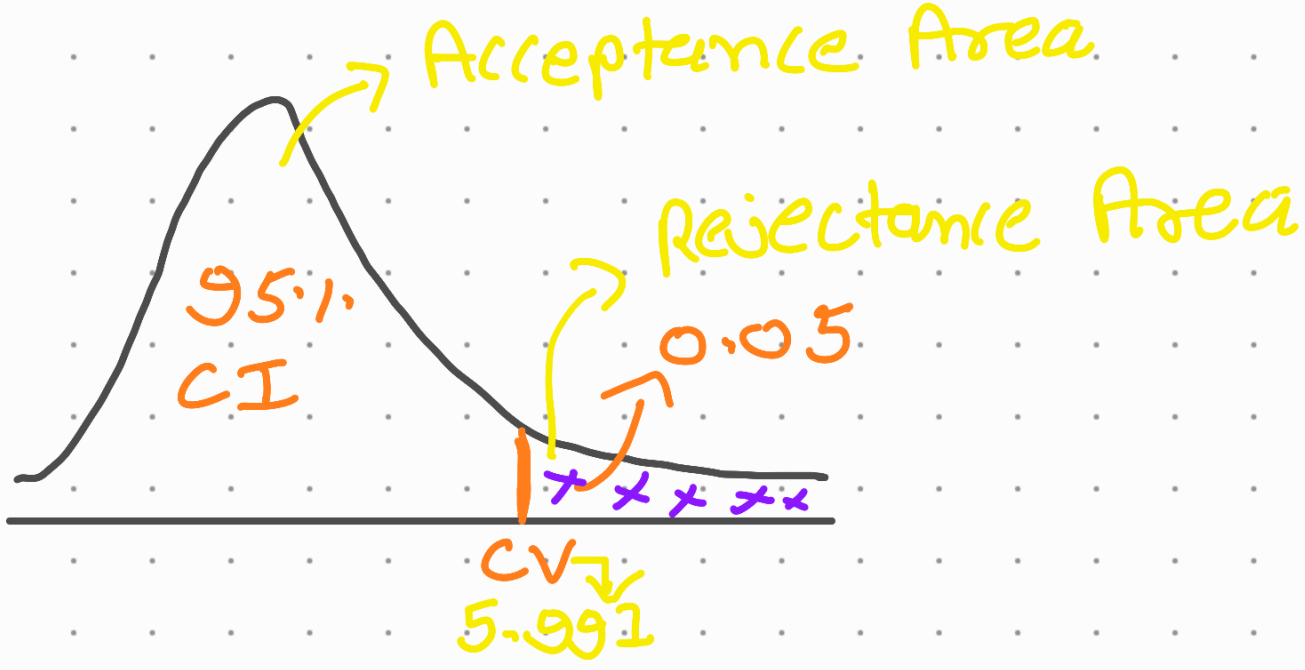
(c) $\alpha = 0.05$ CI = 95%

(d) degree of freedom

$= k - 1 = 3 - 1 = 2$

k = number of categories

(e) Decision Boundary



→ If χ^2 is ~~greater~~ than 5.991, Reject the Null Hypothesis (H_0)

→ Else we fail to reject the Null Hypothesis (H_1)

(f) Calculate χ^2-Test stats

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$= (140-100)^2 + (160-150)^2 + (200-250)^2$$

$$= \frac{1600}{100} + \frac{100}{250} + \frac{2500}{250}$$

$$= 16 + 0.66 + 12$$

$$\boxed{\chi^2 = 26.66} > 5.99, \text{ Reject}$$

the Null Hypothesis H_0.

Conclusion

→ The weights of 2020 population are different than those expected in 2010.

Analysis of Variance (ANOVA)

→ ANOVA is a statistical method used to compare the means of 2 or more groups.

→ ANOVA

① Factors (variable)

② Levels

e.g. ① Medicine (Factor)

[Dosage] 5mg 10mg 15mg

Levels

② Mode of Payment

[Levels] GPay, PhonPe, Imps, NEFT

→ Assumptions In Anova

① Normality of Sampling Distribution of Mean

→ The distributio of sample mean is normally distributed.

② Absence of Outliers
→ Outling score need to be removed from the dataset.

③ Homogenity of Variance
→ Population Variance in different levels of each independent variable are equal

$$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$$

④ Samples are independent and random.

Types of ANOVA (3Types)

① One Way ANOVA : One factor with atleast 2 Levels. These levels are independent.

e.g ① Doctor wants to test new medication to decrease headache.
→ They split the participants in 3 conditions [10mg, 20mg, 30mg]
→ Doctor ask the participants to rate the headache between [1-10]

Medication → Factor

10 mg	20 mg	30 mg → [Levels]	
5	7	2	} Rating to Headaches
3	4	6	
-	-	-	
-	-	-	

② Repeated Measures ANOVA : One factor with atleast 2 levels. These levels are dependent.

e.g.

Running → Factor

Day-1	Day-2	Day-3 → Levels
8	5	4
7	4	9
-	-	-
-	-	-

③ Factorial ANOVA: Two or more factors each of which with atleast 2 levels. These levels can be independent or dependent.

e.g:

Running → Factor

Gender (Factor)	Levels	Day-1	Day-2	Day-3 (Levels)
	Male	8	5	4
		9	4	3
	Female	2	4	6
		7	8	3

Hypothesis Testing In ANOVA

(Partitioning of Variance In The ANOVA)

(a) Null Hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 \dots \mu_k$

(b) Alternate Hypothesis H_1: Atleast one of the sample mean is not equal

$\times$ $\mu_1 \neq \mu_2 \neq \mu_3 \dots \mu_k$

Test Statistics

$$F = \frac{\text{Variance between sample}}{\text{Variance within sample}}$$

Variance Between Sample

Variance within Samples

x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4
$\bar{x}_1 = 3$	$\bar{x}_2 = 29/5$	$\bar{x}_3 = 4$

$H_0 = \bar{x}_1 = \bar{x}_2 = \bar{x}_3$

H_1 = Atleast one sample mean is not equal

One Way ANOVA

→ One factor with atleast 2 levels are independent

e.g:

① Doctors want to test a new medication which reduces headache. They split the participant into 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache between [1-10]. Are there any differences between the 3 conditions using alpha = 0.05?

→

15 mg	30 mg	45 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

(a) Null and Alternate Hypothesis

$H_0: \mu_{15} = \mu_{30} = \mu_{45}$

H_1: Not all means are equal

(b) $\alpha = 0.05$ $CI = 0.95$

(c) Degree of freedom

$N = 21$ $n = 7$ $a = 3$

$dof_{between} = a - 1 = 3 - 1 = 2$

$dof_{within} = N - a = 21 - 3 = 18$

$dof_{total} = N - 1 = 20$

Range of value will be between

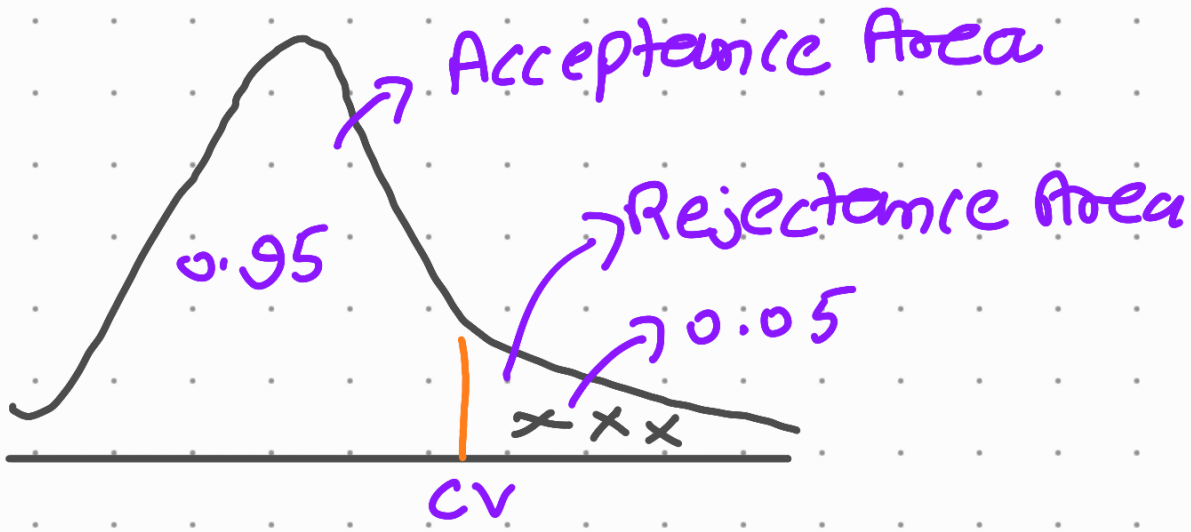
$(2, 18) \Rightarrow$ F-table

df_1 df_2

$\alpha = 0.05$

Critical Value

(d) Decision Boundary



Decision Rule

→ If the F value is greater than 3.5546, reject the null hypothesis H_0.

(e) F-Test Stats

$$F = \frac{\text{Variance between samples}}{\text{Variance within sample}}$$

	SS	df	MS (SS/df)	F
Between	98.67	2	49.34	86.56
Within	10.29	18	0.54	
Total	108.96	20	49.88	

① $SS_{between} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$

15 mg = 9+8+7+8+9+8 = 57

30 mg = 7+6+6+7+8+7+6 = 47

45 mg = 4+3+2+3+4+3+2 = 21

$= \frac{(57)^2 + (47)^2 + (21)^2}{7} - \frac{[57^2 + 47^2 + 21^2]}{21}$

= 98.67

(2) $SS_{within} = \Sigma y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}$

$\Sigma y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + \cdots$

$= 853$

$= 853 - \frac{[57^2 + 47^2 + 27^2]}{7}$

$= 10.29$

$F\text{-}Test = \frac{MS_{between}}{MS_{within}}$

$= \frac{49.34}{0.54} = 86.56$

→ If F is greater than 3.5546, Reject the H_0

→ 86.56 > 3.5546, Reject the null Hypothesis.