# STATISTICS FOR THE DATA SCIENCE

## Part - 4

- STUDENT T-DISTRIBUTION
- BAYES THEOREM
- CHI – SQUARE TEST
- F – DISTRIBUTION
- ANOVA AND ITS TYPES

#Value_freeContent

@Krishan kumar

## (B) Student t-distribution → Statistical analysis using Z-score we need population standard deviation ($\sigma$).

* How do we perform an analysis when we don't know the population standard deviation.

Ans → Student t-distribution

in Z-score →

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$s$ = Sample std.

* Degree of freedom →

$$dof = n - 1$$

$n \Rightarrow$ Sample size

We 3 people

➤ We can find aera under the curve by using t-value OR dof. for t-distribution where population std (not given) only Sample std (given)

**Problem** → In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has eighter a +ve OR -ve effects on intelligence, or no effect at all. A Sample of 30 participants who have taken the medicine has a mean of 140 with a std of 20. did the medication affect intelligence? C.I. ⇒ 95%.

**Ans** → $\mu = 100$, $\sigma = 20$, $n = 30$, $\bar{x} = 140$, C.I ⇒ 95%., $\alpha = 0.05$

① Null hypothesis ($H_0$) → $\mu = 100$

② Alternate '' ($H_1$) → $\mu \neq 100$ { 2 Tail test }

↓

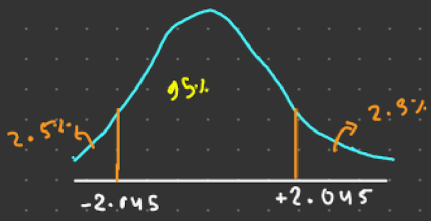
only apply when the value of a mean is eighter ↑↑ OR ↓↓

③ $\alpha = 0.05$

④ Degree of freedom :-

$df \rightarrow 30 - 1 \rightarrow \underline{29}$

⑤ Decision boundry :-




by using t-table,

df = 29

$\alpha = 0.05$

we got ⇒ +2.045

if t-test is less then -2.045 and greater then 2.045, we reject the null hypothesis.

⑥ Calculate t-test statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \Rightarrow \frac{140 - 100}{20/\sqrt{30}} \Rightarrow \frac{40}{3.65} \Rightarrow \underline{10.96}$$

⑦ Conclusion

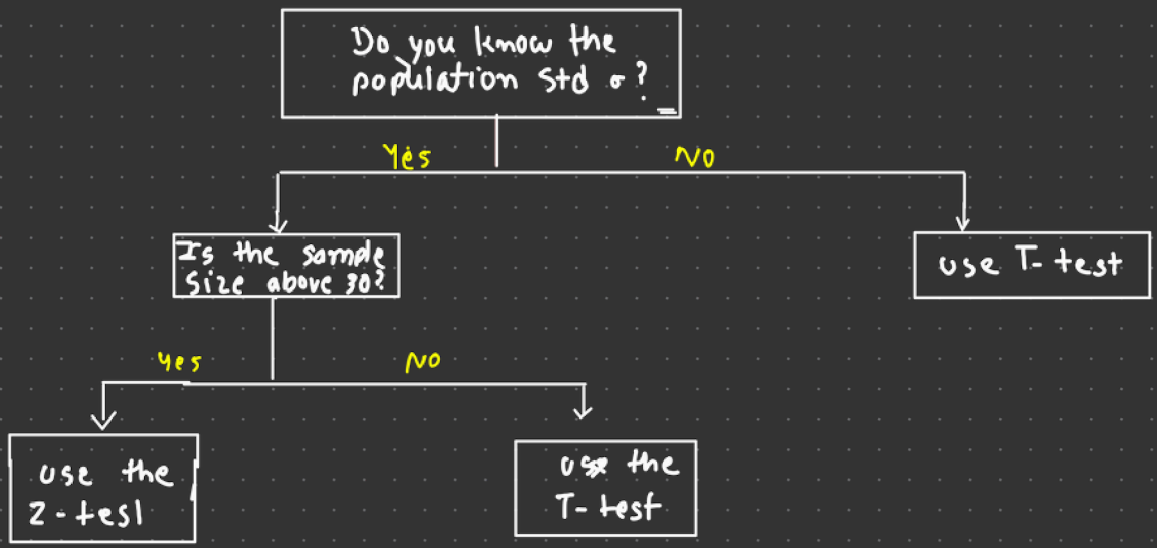
Decision Rule :- if t is less then -2.045 and greater then 2.045 we reject the null hypothesis.

$t = 10.96 > 2.045$ {we Reject the null hypothesis}

Final → Medication has increased the intelligence.

## ⭐ Interview Questions

when to use t-test Vs Z-test




## * Types 1 and Type 2 Errors

Reality : Null hypothesis is true OR $H_0$ is false

Decision : $H_0$ is True OR $H_0$ is false.

Conclusion

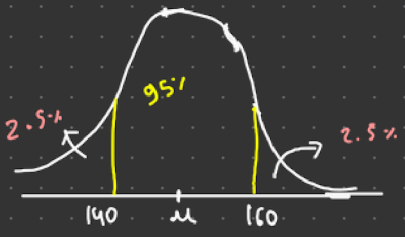
* Outcome : 1 — We reject the $H_0$ (Null hypothesis) when in reality it false → Good

* outcome : 2 — We reject the null hypothesis when in reality it is True → True → Type 1 Error

* outcome : 3 — we retain (accept) the $H_0$, when in reality it is false → Type 2 Error

* outcome 4 : we retain the $H_0$, when in reality it is True → Good

## Confidence Interval and Margin of Error :




$\mu \rightarrow 160$

C.I. → 95%

140 ⟷ 160

Point estimate : A value of any statistics that estimate the value of an unknown population parametter is called "point estimat".

$\bar{x} \rightarrow \mu$

$\bar{x} = 2.95 \quad , \mu = 3$

**Confidence interval:** We construct a C.I. to help the estimate what the actual value of the unknown population mean is :-

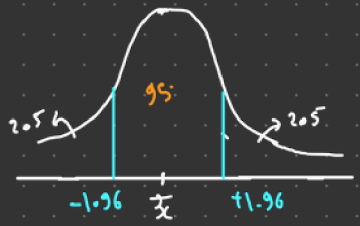
Point estimate ± Margin Error

**Z - test**

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

{ $\alpha$ = Significant value who decided How much C.I. is

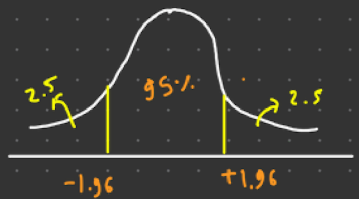
$\alpha = 0.05$




$\Rightarrow$ $Z_{\frac{0.05}{2}}$ $\Rightarrow$ $Z_{0.025}$

**Problem:** On a verbal setion of CAT Exam, the standard deviation is known to be 100. A Sample of 25 test taken has a mean of 520. Construct a 95% C.I. about the mean.

**Ans** $\bar{x} = 520$, $\sigma = 100$, $n = 25$, C.I. = 0.95, $\alpha = 0.05$



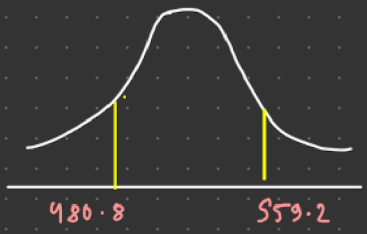

$\Rightarrow$ $1 - 0.25$

$\Rightarrow 0.9750$

$$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

→ Lower C.I. = $520 - (1.96) \times \frac{100}{\sqrt{25}} = 480.8$

→ Higher C.I. = $520 + (1.96) \times \frac{100}{\sqrt{25}} = 559.2$




→ Conclusion : I am 95% sure (confident) that the mean CAT Exam score lies between 480.8 and 559.2

## ★ Bayes theoram

Probability → Independent Event
→ Dependent Event.

### * Independent Event

ex. Rolling a dice
{1,2,3,4,5,6}

P(1) → 1/6
⋮
P(6) → 1/6

Ex → Tossing a coin
P(H) → 1/2
P(T) → 1/2

### * Dependent Event

ex


→ P(R) → 2/5
→ P(B) → 3/4

Here one event effect another events. also.

Ex →

$$P(R \text{ and } B) = P(R) \times P(B/R)$$

$2/5 \times 3/4 \rightarrow 6/20$

$P(A \text{ and } B) = P(B \text{ and } A)$

$P(A) \times P(B/A) = P(B) \times P(A/B)$

$$P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}$$

Bayes theoram

$P(A/B)$ → Conditional Probability

OR

$$P(A/B) = \frac{P(A) \times P(B/A)}{P(B)}$$

A, B → Events

$P(A/B)$ → Probability of A given B is already happ

$P(B/A)$ → Pro. of B given A is true.

$P(A)$, $P(B)$ → The independent Pro. of A and B

## * Uses of Bayes theoram in ML : (ex)

Dataset / Pradict the price → (Y = ?)

| Size of the House | No. of Room | Location | Price |
|---|---|---|---|
| $x_1$ | $x_2$ | $x_3$ | $y$ |

By Bayes theoram →

$$P\left(\frac{y}{(x_1, x_2, x_3)}\right) = \frac{P(y) \times P\left(\frac{x_1, x_2, x_3}{y}\right)}{P(x_1, x_2, x_3)}$$

Analysis by using bayes theoram → Bayesian Statistics

# (1) CHI SQUARE-TEST :-

It is non parametric test that performed categorical {ordinal, nominal} data.

→ CHI SQUARE TEST for Goodness of fit test claims about population proportions [categorical variance].

ex → ①

| | Theory | Sample |
|---|---|---|
| yellow bike | 1/3 | 22 |
| orange bike | 1/3 | 17 |
| Red bike | 1/3 | 59 |

Applyed Goodness fit

Theory categorical distribution

Observed categorical distribution

ex → ② Goodness of fit test : In a student's class of 100 student where 30 are right hand. Does this class fit the theory 12% of people are right handed.

↳

| | O | E |
|---|---|---|
| Right handed | 30 | 12 |
| left handed | 700/100 | 88/100 |

O → Observed value

E → Expected value

12/100 x100

Sample info

Theory categorical distribution

**Problem:** In 2010 Cancus of city, the weight of the individuals in a small city were found to be the following.

| $\leq 50$kg | 50 - 75 | >75 |
|---|---|---|
| 20 % | 30 % | 50 % |

In 2020, age of $n = 500$ individuals were sampled. Below are the Result.

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

using $\alpha = 0.05$ would you conclude the population difference of weights has changed in the last 10 Years.

**Ans:**

Expected value

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

(2010)

$n = 500$

observed value

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

(2020)

| <50 | 50-75 | >75 |
|---|---|---|
| $0.2 \times 500$ $\Rightarrow 100$ | $0.3 \times 500$ $\Rightarrow 150$ | $0.5 \times 500$ $\Rightarrow 250$ |

20 % of 500 $\rightarrow \frac{20 \times 500}{100}$

30 % of 500 $\rightarrow 0.3 \times 500$

50 % of 500 $\rightarrow 0.5 \times 500$

① Null hypothesis: $H_0$: The Data meet the expectation

Alternate Hypothesis: $H_1$: The data does not meet the expectation.

② $\alpha = 0.05$, C.I. → 95%.

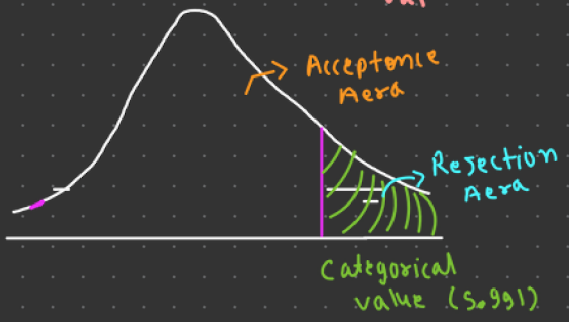
③ Degree of freedom

$$dof = k-1$$

$k \rightarrow$ No. of category

$3-1 \rightarrow 2$

④ Decision Boundry : We used CHI-Square table for the categorical Val value.




denote :

$x^2 \Rightarrow$ CHI Square (Sumbal)

If $x^2$ is greater than 5.991, Reject $H_0$

else:

We fail to Reject the $H_0$

⑤ Calculate CHI-SQUARE Test-Statistic :

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

O → observed, E → expected value

$$\Rightarrow \left(\frac{40}{100}\right)^2 + \left(\frac{10}{150}\right)^2 + \left(\frac{-50}{250}\right)^2$$

$$\Rightarrow \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$\Rightarrow 16 + 0.66 + 10 \Rightarrow 26.66$$

E value

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

O value

| <50 | 50-75 | >75 |
|---|---|---|
| 100 | 150 | 250 |

Final: If $\chi^2$ greater than 5.99 Reject the $H_0$, else fail to reject the $H_0$.

So, $\chi^2 = 26.66 > 5.99$ { we Reject the Null Hypothesis }

Cons: The weights of 2020 population are different than those expected on the 2010 population.

# F-distribution :-

In probability theory and statistic the F-distribution or F-ratio also known ass snedecor's F-dis. or the Fisher-snedecor distribution is continous probubility distribution that aries frequently as the null dis. of a test test statistic, most notebly in the analysis of variance (ANOVA) and other F-test.

* F-test used to campare variance of mean between two group. F-test also called 'variance ratio test'

parametter: $d_1, d_2 > 0$ (degree of freedom)

Support: $x \in (0, +\infty)$

F-distribution with $d_1$ and $d_2$ degree of freedom is the dis. of.

$$\chi = \frac{S_1/d_1}{S_2/d_2}$$

- $S_1 \rightarrow$ Independent random variable { CHI Square distribution }
- $S_2 \rightarrow$ " " "
- $d_1 \rightarrow$ Degree of freedom
- $d_2 \rightarrow$ Degree of freedom

## ⇒ F-test:

**Problem:** The following data show the no. of bulbs produced daily for same day by 2 worker A and B.

| A | B |
|---|---|
| 40 | 39 |
| 30 | 38 |
| 38 | 41 |
| 41 | 33 |
| 38 | 32 |
| 35 | 39 |
| | 40 |
| | 34 |

Can be consider based on the data worker B is more stable and effiecient

$\alpha = 0.05$

**Ans →**

① Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2$

② calculation of variance:

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

**A**

| $x_1$ | $\bar{x}$ | $(x_1 - \bar{x})^2$ |
|---|---|---|
| 40 | 37 | 9 |
| 30 | 37 | 49 |
| 38 | 37 | 1 |
| 41 | 37 | 16 |
| 38 | 37 | 1 |
| 35 | 37 | 4 |
| $\bar{x} = 37$ | | $\Sigma (x_1 - \bar{x})$ ⇒ 80 |

$S_1^2 \Rightarrow \frac{(x_1 - \bar{x})^2}{n-1}$

$S_1^2 \Rightarrow \frac{80}{6-1} \Rightarrow \frac{80}{5} \Rightarrow 13$

**B**

| $x_2$ | $\bar{x}$ | $(x_2 - \bar{x})^2$ |
|---|---|---|
| 39 | 37 | 4 |
| 38 | 37 | 1 |
| 41 | ' | 36 |
| 33 | ' | 16 |
| 32 | ' | 25 |
| 39 | ' | 4 |
| 40 | ' | 9 |
| 34 | ' | 9 |
| $\Sigma = 37$ | | $\Sigma = (x_1 - \bar{x})^2$ ⇒ 84 |

$S_2^2 = \frac{(x_2 - \bar{x})^2}{n-1}$

$S_2^2 = \frac{84}{8-1} \Rightarrow \frac{84}{7} \Rightarrow 12$

→ Calculation of variance Ratio (F-test):

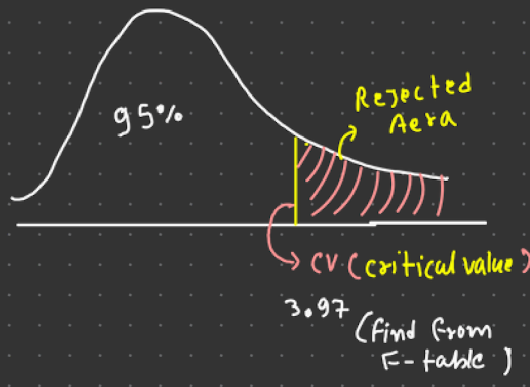
$$F = \frac{S_1^2}{S_2^2} \Rightarrow \frac{13}{12} \Rightarrow 1.33$$

③ Decision Rule:

$df_1 \rightarrow 6-1 \rightarrow 5$

$df_2 \rightarrow 8-1 \rightarrow 7$

$\alpha \rightarrow 0.05$




If F-test greater than 3.97, Reject the Null Hypothesis

1.33 < 3.97 , {so, we Reject the Ho}

Final conclusion :- Worker B is not efficient when compare to worker A.

## (ANOVA) Analysis of variance :

Anova is a statistical method to compared the mean of 2 or more group.

ANOVA :

① factors (variable)

② levels

Ex ① factors = Medicine

level = 5mg, 10 mg, 15mg [Dosage]

Ex ① Mode of Payment [Factors]

G-Pay  Phone-Pe  IMPS  NEFT  [levels]

## ⇒ Assumptions in ANOVA:

① Normality of Sampling Distribution of mean The dis. of Sample mean is normaly distributed.

② Absense of outlier.

Outlier score need to be Remove from dataset.

③ Homogenity of variance

Each one of the population has same variance

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

Population variance in different levels of each independe variable are equal.

④ Sample are independent and Random.

## ⇒ Types of ANOVA:

① One way ANOVA
② Repeated Measured ANOVA
③ Factorial ANOVA

① One way ANOVA: One factor with at least 2 levels, there levels are independent.

② **Repeated Measured anova:** One factor with at least 2 levels, levels are dependent.

③ **Factorial ANOVA:** Two or more factor (each of which with after 2 level), levels can be eighter dependent or independent.

## ⇒ Hypothesis testing in ANOVA:

Null Hypothesis: $H_0$: $\mu_1 = \mu_2 = \mu_3, \ldots \mu_n$

$H_1$: At least one of the mean is not equal

$$\times \boxed{\mu_1 = \mu_2 = \mu_3 \ldots \mu_n}$$

## Test statistic:

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

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| 1 | 6 | 5 |
| 2 | 7 | 6 |
| 4 | 3 | 3 |
| 5 | 2 | 2 |
| 3 | 1 | 4 |
| $\Sigma X_1 = 15$ | $\Sigma X_2 = 14$ | $\Sigma X_3 = 20$ |
| $\bar{X} = 15/3 \Rightarrow 3$ | $\bar{X_2} = 14/5$ | $\bar{X_3} = \frac{20}{5} \Rightarrow 4$ |

$H_0$: $\bar{X_1} = \bar{X_2} = \bar{X_3}$

$H_1$: at least one sample mean is not equal

* **One way anova :** One factor at least 2 levels, levels are independent

**Problem :** Doctor want to test a new medication which reduce headach. They split the participant into 3 condition [15mg, 30mg, 45mg]. later on the doctor ask the patient to rate the headach between [1 → 10] are there any difference between 3 condition using Alpha ($\alpha$) → 0.05 ?

**Ans:**

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

① **Define null Hypothesis :**

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

$H_1$ : not all $\mu$'s are equal

② **State significant value :**

$\alpha = 0.05$

C.I. → 0.95

③ **Calculate degree of freedom :**

$N = 21$, $a = 3$, $n = 7$ (a → Number of Sample)

$df_{between}$ : $a-1 \rightarrow 3-1 \Rightarrow 2$

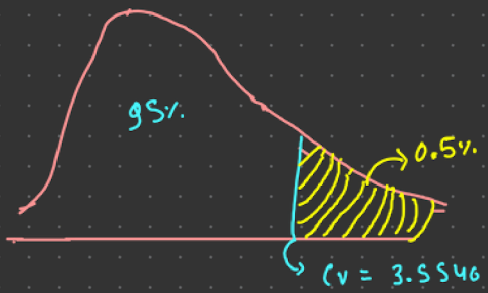
$df_{within}$ : $N-a \rightarrow 21-3 \Rightarrow 18$

$df_{total}$ : $N-1 \rightarrow 21-1 \Rightarrow 20$

(2, 18)

↳ super useful to finding critical value in table

F-Table




Critical Value → 3.5546

④ **State decision Rule :**

If F is greater than 3.5546, we Reject the $H_0$

⑤ **Calculate test Statistic :**

SS = Sum of Square

$SS_{between}$ $SS_{within}$ $SS_{total}$

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

⑥ $$SS_{between} = \frac{\Sigma (\Sigma a_i)^2}{n} - \frac{T^2}{N}$$

15 mg → 9+8+7+8+8+9+8

30 mg → 7+6+6+7+8+7+6

45 mg → 4+3+2+3+4+3+2

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

$$\Rightarrow 98.67$$

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

{ $y^2$ ⇒ Given table value में हर element का square करना }

$$\Rightarrow \Sigma y^2 - \left[\frac{57^2+47^2+21^2}{7}\right]$$

$\Sigma y^2 \Rightarrow 9^2+8^2+7^2+8^2+\ldots\ldots+3^2+2^2 \Rightarrow 853$

$$\Rightarrow 853 - \left[\frac{57^2+47^2+21^2}{7}\right]$$

$$\Rightarrow 10.29$$

(3) $SS_{Total}: \Sigma y^2 - \frac{T^2}{N}$

$$\Rightarrow 853 - \frac{125^2}{21}$$

$$\Rightarrow 108.95$$

| | SS | df | Ms | F |
|---|---|---|---|---|
| Between | 98.67 | 2 | 49.34 | |
| within | 10.29 | 18 | 0.54 | |
| Total | 108.95 | 20 | 49.88 | |

{ $MS = SS/df$ }

F ⇒ $\frac{Ms\ between}{Ms\ Within}$ is equal to ⇒ $F = \frac{\text{Variation between sample}}{\text{variation within sample}}$

$$\Rightarrow \frac{49.34}{0.54} \Rightarrow 86.56 \Rightarrow F = 86.56$$ (1)

**Final:**

If F is greater than 3.5546, we Reject the Ho

86.57 > 3.5546, So we Reject $H_0$.