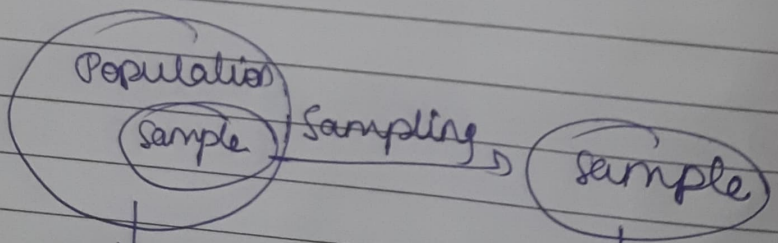
## Population (N)

It includes all the elements from a set of data. The population is the whole set of values or individuals you are interested in. population characteristics are mean ($\mu$), standard deviation ($\sigma$), proportion (P), meadian, percentiles etc. The value of population is fixed. This characteristics are called population Distribution.

* **Sample (n)** → subset of Population Data and the set of values you actually use in your estimation.

This Sample has some quantity computed from values eg: mean ($\bar{x}$), standard deviation (S), Sample proportion etc. this is called Sample distribution




Data under Investigation [very large, difficult to investigate]

Part of Population

* If large dataset then go for Sample
  If small then population

# Statistics

Kisi bhi data ke andar se info nikalne ko upar mathematical analysis using particular methods

## Descriptive

→ Jab ham (entire) population data ke upar analysis karte hai then descriptive ko Statistics :

→ measures of central tendancy (mean, median, mode)
→ measures of variability → Range, MeanAbsolute Deviation, variance, StD
→ Measures of shape → Skewness
→ percentiles → Boxplot
→ covariance and correlation

## Inferential statistics

→ sample ke upar analysis.

imp
→ Central limit theorum
→ Hypothesis testing
→ Z-test
→ T-test
→ Chi-square-test

## ✡ Measurement of central tendancy ✡

### ① Mean / Avg

mean = $\frac{\text{Sum of all data}}{\text{Total no. of data}}$

→ dataset['Age'].mean

→ sns.histplot(x="Age", data=dataset, bins = [i for i in range(0,81,10)])

→ plt.show()

Import numpy as np
Import Pandas as Pd

```
ar = np.array([4,5,6,2,1,8,5,6])
np.sum(ar)/len(ar)
  ⇒ 4.8
np.mean(ar) ⇒ 4.8
```

② median → data → even → $\frac{(n)+(n+1)}{2}$
→ odd → $n/2$

→ First sort

① np.median[dataset["Age"]]
② dataset["Age"].median()

③ mode → most repeated

class → N

→ $\mu = 82$

$\sigma = 20$

→ $n = 81$

$\bar{x} = 90$

⊙ we have to find weather it justifies or not

⊙ here we are given pop - std, mean

⊙ no. of sample data > 30

hence we use z-test

$$z\text{-test} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$\bar{x}$ = sample mean

$\mu$ = pop mean

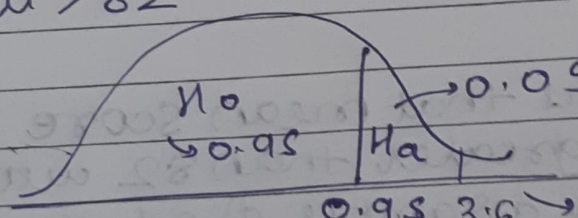
$\sigma$ = pop std

$n$ = no. of sample data

$H_0 = \mu \neq 82$

$H_1 = \mu > 82$

$\alpha = 0.05$




→ right tailed

$z_{cal} = 8.6$

$z_{value} = 1.64$

$$z_{cal} = \frac{90 - 82}{\left(\frac{20}{\sqrt{81}}\right)} = \frac{8}{\left(\frac{20}{9}\right)} = \frac{72}{20} = \frac{3.6}{}$$

$z_{cal}$ is in Ha area hence ⊙ teachers claim that $\mu > 82$ is true

$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ $n > 30$

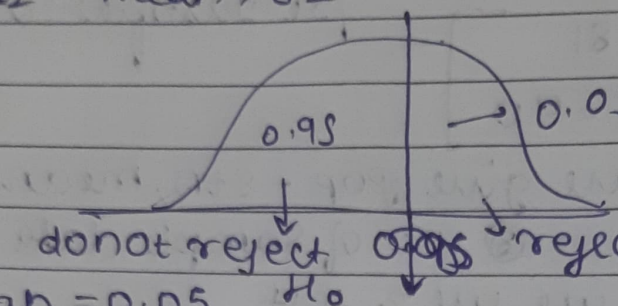
The teacher belive true mean > 82

$H_0$ = mean ≠ 82

$H_1$ = mean > 82

default level of significance = 0.05

mean what is believed to be correct




$Z_{tab}$ = 0.05

→ you look for z-value where area to the left is 95% (0.95)

area - 0.95 → $Z_{tab}$ → 1.64

1.60
0.04
1.64

$Z_{cal}$ = 3.6

which is greater reject $H_0$

hence it true

Q A teacher claims that mean score of student in his class is greater than 82 with a standard deviation 20. If a sample of 81 students was selected with a mean score of 90.

Q2) Imagine you work for an e-commerce company, and your team is responsible for analyzing customer purchase data. you want to ~~find out~~ determine whether a new website design has led to a significant increase in the average purchase amount compared to the old design.
Below is data of random 30 samples of purchase they made on website

- old design data = [45.2, 42.8, 38.9, . . . . ]
- new design data = [48.5, 49.1, 50.2 . . . . ]

Population StD = 2.5

→ $H_0$ : $Web_{new} = Web_{old}$

$H_{01}$ : $Web_{new} > Web_{old}$

$\mu_{new} = \mu_{old}$

$\bar{X}_{new}$, $\bar{X}_{old}$

$\sigma = 2.5$

$n = 30$

$$Z_{test} = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \qquad n \geq 30$$

$$Z_{test} = \left(\frac{\bar{x}_{new} - \bar{x}_{old}}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right) - (\mu_{new} - \mu_{old})$$

$$= \left(\frac{\bar{x}_{new} - \bar{x}_{old}}{\sigma/\sqrt{n}}\right)$$

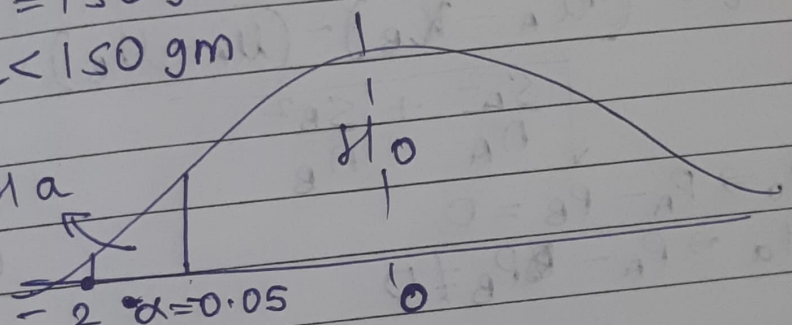
# T - test

Q1) A manufactures claims that the average weight of a bag of potato chips is 150 gm. A sample of 25 bags is taken, and the average weight is found to be 148 gm, with a standard deviation of 5 gm. Test the manufacturers claim using a one-tailed t-test with a significance level of 0.05.

$\mu = 150 \quad \bar{x} = 148 \quad n = 25 \quad s = 5$

$H_0: \mu = 150$ gm

$H_a: \mu < 150$ gm




$df = n - 1 = 24$

$t = -1.711$ (tab) (as on left side)

$t_{cal} = \left(\frac{148 - 150}{5/\sqrt{25}}\right) = \frac{-2}{\left(\frac{5}{5}\right)} = -2$

Hence $t_{cal}$ is in $H_a$ mean $\mu < 150$.

Hence company's claim is wrong

Q27 A company wants to test whether there is difference in productivity between 2 teams. They randomly select 20 employees from each team and record their productivity scores. The mean productivity score for team A is 80 with std of 5, while mean productivity score for Team B is 75 with a standard deviation of 6. Test at 5% level of significance whether there is a difference productivity between 2 teams.

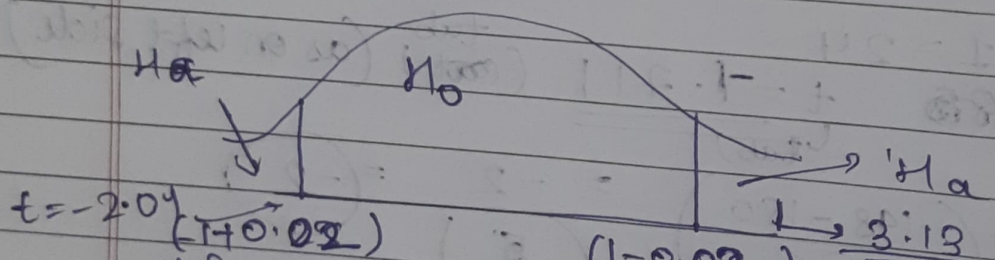
→

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

$H_0 \Rightarrow P_A - P_B = 0$

$H_a \Rightarrow P_A - P_B \neq 0$

$\frac{0.05}{2} = 0.025$




$df = 20+20-2 = 38$

$$t_{cal} = \frac{(80-75)-0}{\sqrt{\frac{25}{20} + \frac{36}{20}}} \quad \frac{(5)}{\sqrt{\frac{61}{20}}} = 3.13$$

∴ Ha is right ; Ho is rejected.

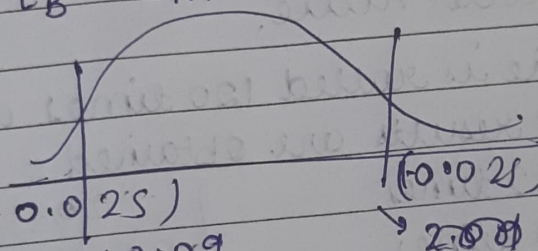
Here we have taken degree of freedom (df) as n-1 because typing speed of same sample was recorded previous question same diff.

Q3) A company wants do test weather the new training program improve the typing speed of the employees. The typing speed of 20 employees was recarded before and after the training program. The data is given below. Test at 5% level of significance whether the training program has an effect on the typing speed of the employees.

- Before: 50, 60, 45, 65, 55, 70, 40, 75, .... 65
- After: 60, 70, 55, 75, 80, 50 . . . . .

→ $H_0 = t_B - t_A = 0$

$H_a = t_B - t_A \neq 0$




2.09 2.09

$df = n - 1 = 19$

$$t = \frac{\mu_A - \mu_B}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{n}}}$$

t_table = St.t.ppf (1-0.025, 19)

t_table

⇒ 2.09302

Before = np.arr([50, 60, 45 . . . 65])

after = np.arr([60, 70, 55, 75, 65 . . .])

std_a = np.STD (After)

Std_b = np.STD (Before)

mean_a = np.mean (After)

mean_b = np.mean (Before)

t_cal = (mean_a - mean_b) / (np.sqrt ((std_a * std_b) / len(after)) + ((std_b * std_b) / len(before)))

t_cal

→ 2.061200s

$t_{cal} < t_{tab}$

$\therefore$ $H_0$ is true

## Chi-square test

→ goodness

→ ~~relationship~~ Independence

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$\chi^2$ → chi-squared

$O_i$ = observed value

$E_i$ = Expected value.

(Q1) A fair die is rolled 120 times and the following results are obtained:

face 1 : 22 times → 20

face 2 : 17 times → 20

face 3 : 20 times → 20

face 4 : 26 times → 20

face 5 : 22 times → 20

face 6 : 13 times → 20

→ Ideal $\frac{120}{6} = 20$

Test at a ~~level~~ 5% level of significance weather the die is fair

→ $H_0$ → die is fair

$H_a$ → die is not fair

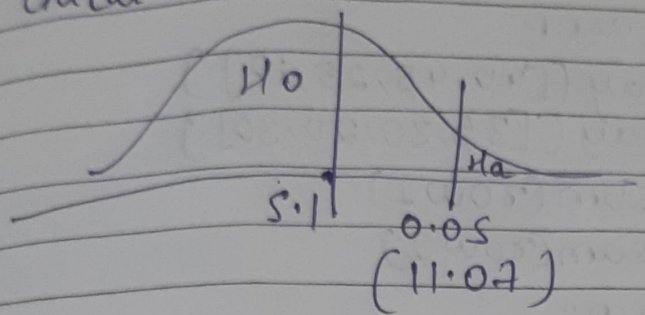
→ $df = n - 1 (6 - 1) = 5$

$chi_{tab} = 11.07$

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

```
import numpy as np
ob = np.array([22,17,20,26,22,13])
ex = np.array([20,20,20,20,20,20])
np.sum(np.square(ob-ex)/ex)
= 5.0000...
```

Chical = 5.1




∴ $H_0$ → die is fair

Q A study was conducted to investigate whether there is a relationship between gender and the preferred genre of music. A sample of 235 people was selected, and data collected is shown below. Test at 5% level of significance whether there is a significant association between gender and music preference

| | POP | Hiphop | classical | Rock |
|---|---|---|---|---|
| male | 40 | 45 | 25 | 10 |
| female | 35 | 30 | 20 | 30 |

$H_0$ : No association  $\alpha = 0.05$

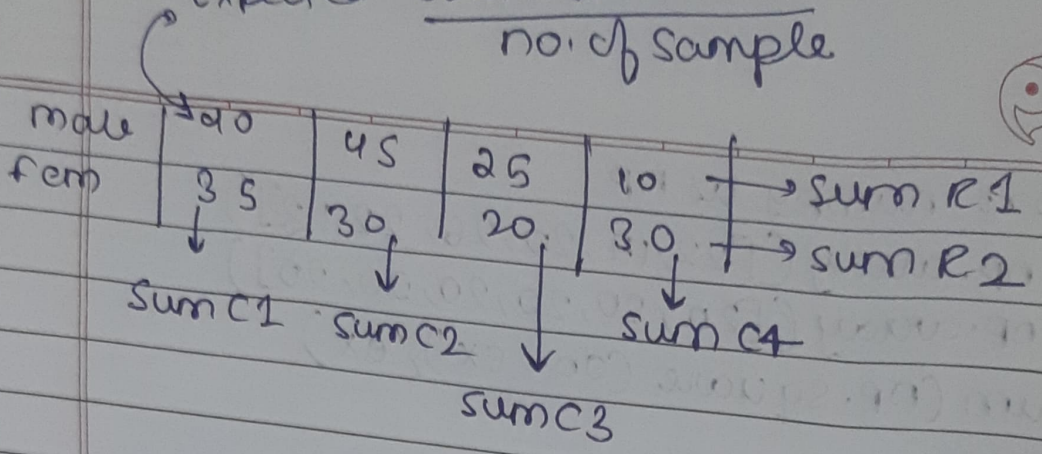
$H_a$ : association  $df = (row-1)(col-1)$

$(2-1)(4-1)$

$= 3$

$Chi_{tab} = 7.815$

Expected = $\frac{\text{sum C1} \times \text{sum R1}}{\text{no. of sample}}$



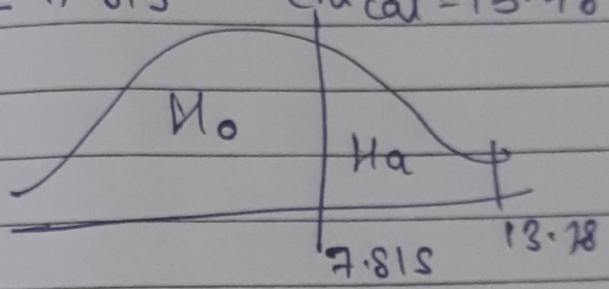

```
mport numpy as np
row1 = np.array([40,45,25,10])
row2 = np.array([35,30,20,30])
sum_r1=np.sum(row1)
sum_r2=np.sum(row2)
sum_r1,sum_r2
(120,115)
sum_cal = row1+row2
sum_cal
array([75,75,45,40])
sum_row = np.array([sum_r1,sum_r2])
sum_row
[120,115]
exp=[]
for i in sum_row:
    for j in sum_cal:
        print(i,j)
        value = (i*j)/235
        exp.append(value)
exp
[38.297,
 38.2978,
 .
 .
 .
 19.5744]
```

obj = np.array([40,45,25,10,35,30,20,30])

np.sum((np.square(obj - exp))/exp)

= 13.788747998

Chi tab = 7.815 Chi cal = 13.78




∴ Ho is wrong