⇒ Inferential statistics Date:- 30/3/22 11:00 AM

* Central Limit Theorm :-

For Continuos Data

Example :-

Temperature @ Deshaipet (My home) @ 6 pm

| Day. | Temperature |
|---|---|
| * Jan 1 | 27. |
| * Jan 2 | 29 |
| * Jan 3 | 34 |
| * Jan 4 | 30 |
| * Jan 5 | 31.5 |
| * Jan 6 | 32 |
| * Jan 7 | ? |

Avg = 30.5 ↑ (What is Jan 7 Temperature)

Avg $(\bar{X})$ = 30.5

* What probability of 30.5 Value For Jan 7

$$P(30.5) = 0, \quad \leq = 1$$

$$P(30.5) = \frac{1}{\text{Infinite}}$$

$$\simeq 0, \quad \text{" "}$$

For Single Value probability is "Zero".

So, we adding Error Both sides,

We have "add" Error

$\bar{X} \pm$ std.error → Standard.error

5 % error.

$\bar{X} = (30.5 \pm 5)$

Center value

$[25.5, 35.5]$ → with standard error

Two side error

$[\bar{X} - \text{stderror}, \bar{X} + \text{std.error}]$

$\therefore$ std.error $= \bar{X} - \frac{\sigma}{\sqrt{n}}$ → Based on Sample collection

$\bar{X} + \frac{\sigma}{\sqrt{n}}$

- Central Limit theorm

Standard deviation

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}}, \bar{X} + \frac{\sigma}{\sqrt{n}}\right]$$

→ Predicting with in Range

Mean

Sample

* For Continuos Data, In order to predict, The Range, we apply Central Limit Theorm.

Not For "Discrete Variable"

↓

Because Discrete data doesn't have standard deviation.

With what Confidence, (or) what probability ?

Confidence Interval :- $\left[\bar{X} - Z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \;,\; \bar{X} + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$

* $Z_{1-\alpha} = Z_{\text{confidence}}$

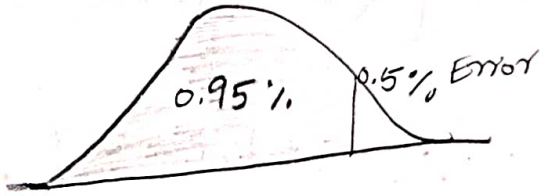
$= Z_{0.95} \Rightarrow$ 5 % Error

$= Z_{0.90} \Rightarrow$ 10 % error

From Z Table , $Z_{95\%} = 1.65$

| Z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| +1.6 | .945 | .946 | .947 | .9484 | .9495 | .9505 |

95 %.




$\therefore \left[\bar{X} - 1.65\frac{\sigma}{\sqrt{n}} \;,\; \bar{X} + 1.65\frac{\sigma}{\sqrt{n}}\right]$ with 95 % Confidence

What is $(1-\alpha)$ in Confidence Interval ?

A :- $\alpha$ = Error

$1-\alpha$ = Confidence ,

Hypothesis testing * → Statistical Testing

We make conslusions ⇒ poplutation data based on Sample data.

* Contionous Variable :-

→ 1 Sample Test
  - 1. Sample ZTest
  - 1 Sample T Test

→ 2 Sample Test
  - 2 Sample Test For Equal Variance
  - 2 sample Test For UnEqual Variance.

→ Anova Test

* Discrete Variable :-

1 proportion Test

2 proportion Test

Chisquare Test

Example :- "SRK" Teaching Feed Back = 5 rating ★★★★★

⇒ 15 students (response) — Sample Data

⇒ Avg. rating = 4.9

⇒ population Data = 50 students → maybe 1.0, 2.0, 2.5 ratings in population

To check by

(Statistical Testing) Hypothesis Testing

(Conclusions / inferences) ⇒ Based on Sample ($\bar{x}$, $s^2$, $s$)

Avg/mean (var°) (stand.

(Estimating population) ($\mu$, $\sigma^2$, $\sigma$)

Pop. Avg, Pop. var, P.op. std.

For Sample, Consider same for population Data

(Sample ki Edi ayethay apply chesthamo, Same population kuda aday Apply cheyali)

Inferential Statistics → Your InFerencing Bgme one, Based on SomeThing.

↓

conducted on "Samples".

→ statement

⇒ Hypothesis testing :-

* it is used To determine whether a statement about "value of a population" parameter should (or) shouldn't be rejected.

Example :-

In Indian Court of Law,

* person is innocent
* person is criminal

In what creteria, we can Judge him,

So, we have.

* Basic Assumption ⇒ Null hypothesis [$H_0$]
* Evidence of crime ⇒ Alternative hypothesis [$H_1$ / $H_a$]

→ Hypothesis testing

- Null Hypothesis [$H_0$] = Person is innocent
- Alternate Hypothesis [$H_1$ / $H_a$] = Person is criminal

⇒ For Population parameter, (Basic assumption) is Null hypothesis ($H_0$)

⇒ The alternative hypothesis ($H_1$/$H_a$) is opposite To Null hypothesis

→ Explanation :-

$H_0$ :- Null hypothesis

$H_1$ :- Alternate hypothesis / Research hypothesis.

Conditions

- $H_0$ :- Status quo / Present situtation is best one. / no difference / no action required

↳ This correct :- No action required.

- $H_1$ :- Researcher is Best and $H_0$ is not Best / difference b/w $H_0$ & $H_1$ / action is Required

↳ This correct :- Take action

Ex :-

$H_0$ : Person is innocent (no action)

$H_1$ : Person is criminal (action required)

Example 2

Q:- A New Manufacturing method is belived to be better than the current method ?

A :- The new method is "No"

* Null hypothesis [$H_0$] :- better Than old method. (No action)

* Alternative hypothesis [$H_1$] :- The new manufacturing method is better. (Action required)

Example 3

Feedback :-

* SRK Sir (Data Science) is Better / Excellent?

* Naresh I.T (Head)

Hypothesis Testing :-

15 students (response). ***** = 5 Rating

$\Rightarrow H_1 - \mu \leq 4.5$ (Less Than or Equal To 4.5/5 rating) (action required)

$\Rightarrow H_0 - \mu > 4.5$ (No action required)

Example 4 :-

Data Science Coarse (Existing coarse)

New coarse → Data

(No action) . $H_0$ – Pintu Data Course (Base) ≤ srk course

(action) . $H_a$ – srk Data Course (new) > Pintu Data Course (Existing coarse)

↓

(Evidence) +8L (package / High Jobs / Cracking certification)

Example 5 :-

Q :- A New bonus plan, That is developed in a attempt to increase Sales

A :-

⇒ Alternative hypothesis [$H_1$] :-

. The new bonus plan increase Sales

* $H_A$ = (Sales new plan > Sales Existing plan)

⇒ Null hypothesis [$H_0$] :-

. New Bonus plan doesn't increases Sales

. If Equal also Don't consider ($H_a$)

* $H_0$ = ( Sales new plan ≤ sales Existing plan)

Example 6 :-

• If sample Data is misleading
Population Data = Error (wrong sampl

Q :- A new Drug ($H_1$) is developed with the goal of lowering cholestrol - Level more than Existing drug ($H_0$)

A :-

* Alternative hypothesis [Ha] :-

• The new drug lowers cholestrol level more than Existing Drug.

* Null hypothesis [Ho] :-

• The new Drug doesn't lower cholestrol - level more than Existing Drug.

Example 7 :-

The Label on Milk bottle states that it Contains 1000 ml.

Null hypothesis ($H_0$) :-

The Label is Correct $\mu \geq 1000$ ml

Alternative hypothesis ($H_a$)

The Label is incorrect $\mu < 1000$ml

* Owner of milk

- $H_0 \doteq \mu = 1000$ ml (No action)
- $H_1 \doteq \mu \neq 1000$ ml (action)

Example 8

Q:- A major hospital in chennai provides one of most compreshensive Emergency medical services in World.

- Operating in a multiple hospital system with approximately 10 mobile medical units, The Service goal is to respond to medical Emergencies with mean time of 8 minutes (or) Less.
- The Director of medical Services wants To Formulate a hypothesis test That could use a sample of Emergency response times to determine whether or not The Service goal of 8 minutes or Less than is being acheived ?

A:- $H_0 : \mu \leq 8$ min (No action) (Mean)

(action) $\Rightarrow H_1 : \mu > 8$ min → Mean of The vechiles

$H_0 : \mu \leq 8$ [The Emergency Service is meeting the response goal; no follow up action required]

$H_1 : \mu > 8$ [The Emergency Service is not meeting the response goal; appropriate Follow-up action is necessary]

$\therefore \mu$ = Mean response Time of Population of medical Emergency requests.

⇒ Confusion Matrix

Example :- Covid patient.

①

| Prediction (Covid Test) Result | Actual (Disease): Person Has Covid | Actual (Disease): Person don't Have Covid |
|---|---|---|
| Person has Covid | TRUE | False (error) |
| Person don't have Covid | False (error) | True |

②

| Prediction (Judge (Evidences)) | Actual: Person is Innocent | Actual: Person is Criminal |
|---|---|---|
| Person is innocent | True | False (Error) |
| Person is Criminal | False (Error) | True |

*

Confusion matrix

Ex : 3

$H_0$ = Students understood class

$H_1$ = Students didn't understood class

Actual (what happend)

| Prediction (Feeling / Evidence) | Student understood | Student didn't understood |
|---|---|---|
| Student understood | True | False (Error) |
| Student didn't understood | False (Error) | True |

ardam ayyendki ani anukuntunna

ardam kaledu ani anukuntunna.

For Example 8

"Hospital Data"

| [Predict] Conclusion | Population Condition [Actual] $H_0$ True ($\mu \leq 8$) | $H_0$ False ($\mu > 8$) |
|---|---|---|
| Accept $H_0$ (Conclude $\mu \leq 8$) | Correct Decision | Type I Error (Incorrect Acceptance) |
| Reject $H_0$ (Conclude $\mu > 8$) | Type II Error (Incorrect Rejection) | Correct Decision |

30/3/22 4:51 Pm

Date:- 31/03/22
1:30pm

Always (=) → Null hypothesis.

* The equility part of the hypothesis always appears in the null hypothesis

Example :- $\mu \geq \mu_0$

$\mu \leq \mu_0$

$\mu = \mu_0$

Mean

⇒ Example 9

* Two Trainers.

Trainer 1

Already Existed

Trainer 2

New Trainer

$H_0$ :- Existing trainer > New Trainer

$H_1$ :- New Trainer > Already Existing Trainer

If Both Trainer good (=)

we Take Null hypothesis

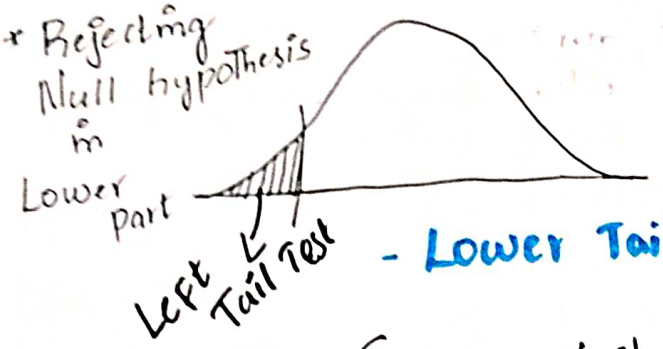
$H_0$ :- Existing trainer $\geq$ New Trainer

Equals To

⇒ One Tailed [lower tailed] test

$H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$ → when ever The (Alternate) (or) Research Hypothesis Has $<$ (Less than) Symbol, Than it is called as "Lower Tailed Test."

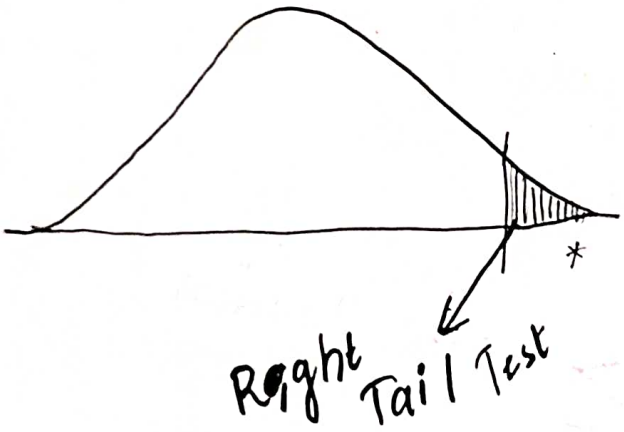



- Lower Tailed Test.

⇒ One Tailed [upper tailed] test

$H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$ → when Ever the (Alternate) or Research Hypothesis Has ($>$ Greater than) Symbol Than it is called as "upper tailed Test".




- Upper Tailed test.

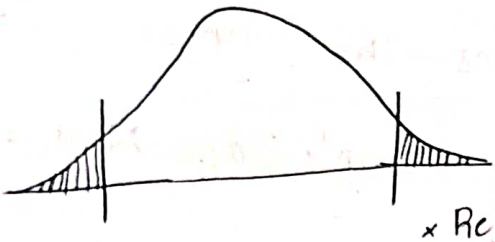
* Two Tailed test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$ →

$H_1$ / $H_a$

when Ever The (Alternate) Has (≠ not Equals to) Symbol, Than it is called as " two tailed Test ".



* Rejecting on Both sides.

- Two Tailed Test

⇒ STeps OF HypoThesis Testing *

STep 1 :- Develop The Null and Alternative Hypothesis

Identify

STep 2 :- Specify The Level of Significance "α"

Here α = Error Rate. To Accept.

If 5 % Error = 0.05

10 % Error = 0.1

20 % Error = 0.2 ....

STep3 :- Collect The Sample data and Compute The Test statistic "P-Value".

Step 4 :- Compute The p-Value, Using Value of test static

* The p-value is The probability, Computed using Test statistic, that measures the Support (or Lack of Support) provided by sample of Null hypothesis.

* If the p-Value is Less than (or) Equal to the Level of significance $\alpha$, the Value of Test Static is in Rejection region.

Step 5 :- Reject $H_0$ if, The p-value $\leq \alpha$ (Null hypothesis)

$P \leq \alpha \Rightarrow$ Reject $H_0$

$P > \alpha \Rightarrow$ Accept $H_0$

$\Rightarrow$ "p" Low "Null" Go → Reject

$\Rightarrow$ "p" High "Null" Fly → Accept

Probability

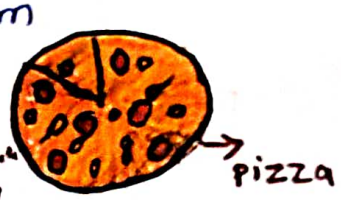
Level of significance

* Case Study

## One Tailed Tests about population When ($\sigma$) Known.

### Example 1

- The Mean Response Times For a Random Sample of "30 pizzas" Deliveries in 32 minutes.




From sample we can conclude = Avg. val = 32, it is not meeting criteria

- The Population Standard deviation is belived To be 10 minutes
- The pizza delivery Services director wants to Perform a hypothesis test, with $\alpha = 0.05$ Level of significance, to determine whether the Service goal of 30 minutes or Less is being acheived.

- Answer :

But, we are not working on sample Data.

We have To Estimate population.

- No. of pizzas ($n$) = 30
- Mean Sample ($\bar{x}$) = 32 minutes
  ($\mu$) = 30 min
- Standard deviation ($\sigma$) = 10 min
- Level of significance ($\alpha$) = 0.05 Level (Error)

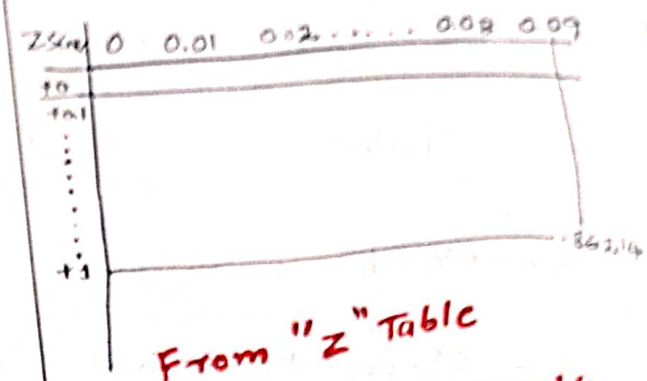
where, 30 minutes goal is Acheived or not, So, 95% Accuracy

So, where, 30 : 32, we have to check with 95%.

Step 1 :-

$H_0 : \mu \leq 30$

$H_1 : \mu > 30$

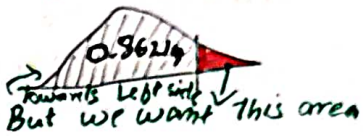



From "Z" Table

1.09 = 0.86214

Step 2 :-

Level of Significance = 0.05




$P = 1 - 0.86214$

$P = 0.1378$

Step 3 :-

Upper Tail test

$$Z_{score} = \frac{X - \mu}{\sigma}$$

$$Z_{Test} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{32 - 30}{\frac{10}{\sqrt{30}}} = 1.09$$

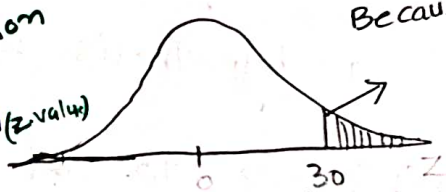
(Z value)

Sample mean

Estimation

Standard deviation

no. of Samples




Because,

If 1, 2, 3, 4, 5 = ..... it's ok, But, it should not cross 30 min

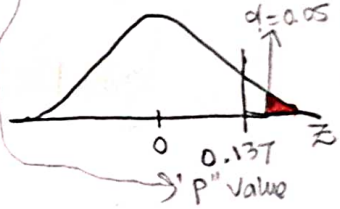
As per requirements in Question

If Alternative hypothesis is

$H_1 : \mu > 30$

greater

called "upper" Tail Test




***

| "Z" - distribution | "T" distribution |
|---|---|
| * Population standard deviation ($\sigma$) is known is called "Z" Distribution | If population standard deviation ($\sigma$) is unknown Than it is called as "T" distribution. |

$Z = \frac{X - \mu}{\sigma}$ → Mean; σ → Standard deviation

"Z" Distributing

Ex :-

| Marks |
|---|
| 90 |
| 60 |
| 58 |
| 83 |
| 98 |
| 32 |

(90) $\frac{X - \text{Avg/mean}}{\sigma}$

$\frac{60 - \text{Avg/mean}}{\sigma}$

Ex :- $\frac{90 - 60}{10} = +3.$ → "Z" score

So, 90 is Converted To Equi. +3. by Z score. From Z-Table

⇒ Conversion of Each of The Value To corrsponding "Z" score is called "Z" distrubution also called as "Standard Normal Distribution"

Z Test ⇒ $Z_{Test} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

* Step 4 :-

Probability of Z score

P (Z score) =

$t_{n-1} = \frac{X - H}{S}$ → Mean; S → Sample Standard deviation

n-1 → degree of Freedom

* Bessel correction :

For sample deviation, Sample Variance, we have To consider (n-1).

⇒ Conversion of Every Value To "T" score is called as T Distribution

$T_{Test} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

* Step 4 :-

Probability of "T" score

P (T score) = ——

Z Distribution and T Distribution are Equa Than,

If $n > 1000$, $Z = T$ → Sample = 1000 Than,

Ex: 0.738, 0.78 → upTo 3,4 Decimals same answer

$n > 30$, $Z \approx T$

Ex: 0.74 ≈ 0.72, 0.73 → upTo First Digit Equal

* Step 5

Minimum Sample Should be 30, Because, Z and T distribution almost be "Equal"

## "P" Value approach

**Step 4 :-** Compute The P-Value

For $Z = 1.09$

P-Value $= 0.137$

**Step 5 :-** Determine whether To reject $H_0$

Because, [P. value $= 0.137 > \alpha = 0.05$]

Here

P. value > Level Of significance ($\alpha$)

So, we Accept "Null Hypothesis" ($H_0$)

we Reject "Alternative Hypothesis" ($H_1$)

* There are not Sufficient Statistical Evidence to infer that pizza Delivery Services is not meeting The response goal of "30 min".

* CODING

# import numpy as np

# From scipy import stats

⇒ Solution

- Null hypothesis = Mean Time <= 30
- Alternative hypothesis = Mean Time > 30
- Alpha = 0.5
- Right Tail z test → Right Tail Rejection.
- z calculated = [observed mean − Population Mean] / Std. Error
- $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

↙

# Z_test = (32-30) / (10 / np.sqrt(30))

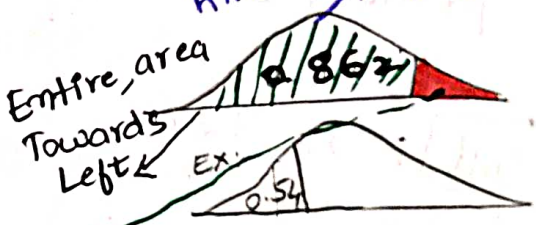
print (Z_test)

Out :- 1.0954451150

# stats.norm.cdf (1.09) → Normal Distribution → Cumulative probability Distribution / Function

Out :- 0.862143 → p value

Ans :- Always Left side




For Right Tail Test ←

# 1 − stats.norm.cdf (1.09)

Out :- 0.137 → Probability it is rejected

* 0.137 > 0.05

So, p High Null Fly / Accept)

**Ex:-**

If, Sample = 30, (Mean Time) 28 = minutes

So, it is Satisfying The requirement Less Than 30. (No problem)

But,

If Sample = 30, Mean Time = 36

So, already we done with 32,

We change 32 To 36. and Calculate the "Z" Table, Z value,

```
# z_test = (36-30) /(10/np.sqrt(30))
  print (z_test)
```

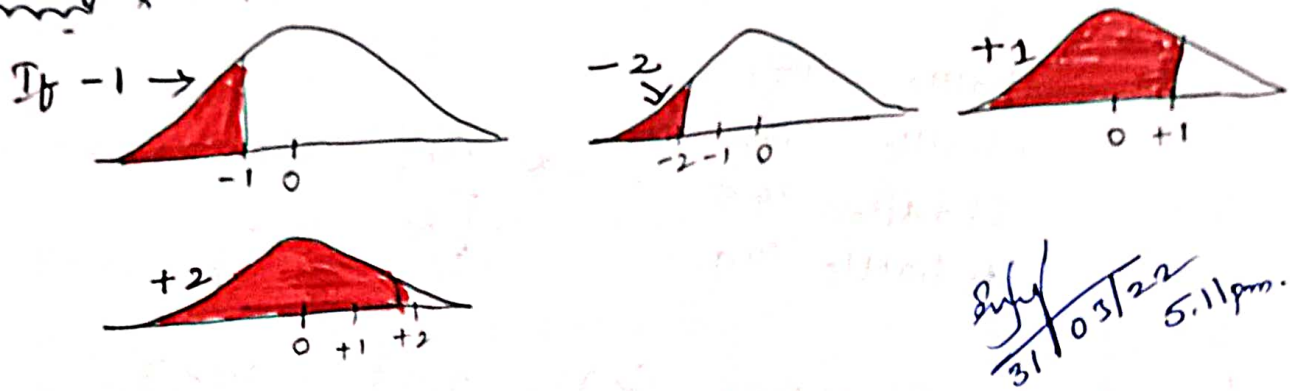
Out : 3.2826

```
# 1- stats.norm.cdf(z_test)
```

Out : 0.00507 ↘ P < 0.05

0.005 < 0.05 ⟹ Reject Null Hypothesis.

* Always towards Left side (Area)




Sufyan 31/03/22 5.11pm.

31/03/22
9:30pm

## Case Study 2

## Left Tail test (Lower Tail)

### Example 2

A Machine is producing perfume bottles wi
The long term Average of 150 ml and standard deviation of 2 ml. Four bottles were picked and average Volume was Found to be 149 ml. Has Volume m̊ of 150 ml ? Test with The average
95 % Confidence level.


⟹ Perfume Bottle

Ans :-

Mean ($\mu$) = 150 ml

Standard deviation ($\sigma$) = 2 ml

$n = 4$

Assume, $n = 4$

↓

1 bottle = 151
2 bottle = 150
3 bottle = 148
4 bottle = 149

$\Rightarrow$ Avg = 149

$\boxed{\bar{x} = 149}$

* level of Confidence ($\alpha$) = 0.05 (95%)

STep 1 :-

Alternate ($H_1$) :- $\mu < 150$ (action)

Null ($H_0$) : $\mu \geq 150$ (no action)

STep 2 :-

$\alpha = 0.05$

$(1 - 0.95)$


If less Than 150, we have To reject

STep 3 :-

$$Z_{Test} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{149 - 150}{\frac{2}{\sqrt{4}}} = \frac{-1}{1} \Rightarrow -1$$

$\bar{X}$ : sample mean/Avg

$\sigma$ : standard deviation

$n$ : no. of bottles

STep 4 :-

For Probability (-1) $\Rightarrow$

From "z" Table.

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| -0 | | | | | | |
| -0.1 | | | | | | |
| -0.9 | | | | | | |
| -1 | .15866 | | | | | |

From Zscore $\doteq$ 0.15866

$P(-1) = 0.15866$ 

Step 5 :-

So, it is Left Tail Test, we got Left Tail Value as 0.15866, So we don't need To [1 - stats.norm.cdf (z_test)]

$P(-1) = 0.15866 > 0.05$

"p" Value is Greater Than "$\alpha$" value

So, 'p' High → Null Fly (Accept The Null)

⇒ CODEING

```
Import numpy as np.
From scipy import stats.
# z_cal = (149 - 150) // (2 / np.sqrt(4))
```

(division)

z_cal → z' Calculated Value.

Out: -1.0

# Left Tail → stats.norm.cdf (z-cal)

# stats.norm.cdf (-1.0)

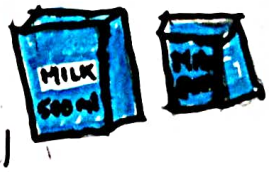
Out :- 0.158655

## Case study 3

2 Tailed Test

Example 3 :- Milk Carton

- Assume that a Sample of 30 milk Carton provides a Sample mean of 505 ml
- Population Standard Deviation 10 ml
- Perform Hypothesis test, at 0.03 Level of Significance, Population mean 500 ml. and to help determine whether the Filling process Should Continue operating or be stopped and corrected.




Ans:-

* $n = 30$
* Sample mean $(\bar{x}) = 505$ ml
* Standard deviation $(\sigma) = 10$ ml

* Level of significance ($\alpha$) = 0.03
* population mean ($\mu$) = 500ml

quality checker
Not Less Than 500ml

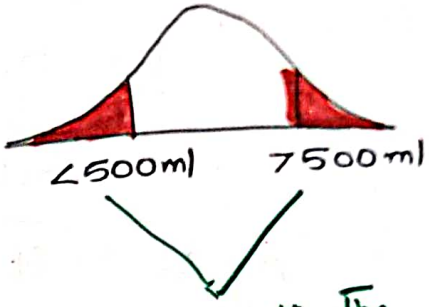
Step 1 :-

(Owner Perceptive)

$H_0$ :- $\mu = 500$ mL

don't Exceed more Than 500 ml

$H_1$ : $\mu \neq 500$ mL




Both The side Rejectio

Step 2 :-

$\alpha = 0.03$ [Error]

$[1 - 0.97] = 97\%$ Accuracy.

Step 3 :-

$$Z\,Test = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{505 - 500}{\frac{10}{\sqrt{30}}} = +2.7386$$

$\bar{X}$ : Sample mean

$\mu$ : Population mean / Estimate

$\sigma$ : Standard Deviation

$n$ : no. of milk Bartons

Step 4

$P(2.7386)$ = Let us consider it as +2.74

| Z | 0 | 0.01 | 0.02 | 0.03 | 0.04 |
|---|---|---|---|---|---|
| +2 | | | | | |
| +2.1 | | | | | |
| ⋮ | | | | | |
| +2.7 | | | | | 0.9968 |

$P(2.7) = 0.99693$ → Left Tail Test

For "+" value Always do (1 − "P" value)

So, For Right Tai Test

$1 - 0.99693$

$= 0.003$

If "z" value Have "−" symbol

Direct 2 * (z score) Multiply with "2"

Again, its 2 Tail Test. So For Two Tail Test

$2 \times [1 - 0.9963]$ → Calculator lo Vastholed But, Jupyter lo Vasthundi.

$P = 0.00616$ →

Steps

$P = 0.00616 < 0.03$

So, "P" Low, Null Go (Reject)

Rejecting Null hypothesis

→ There is no Sufficient Statistical Evidence to infer that The Null hypothesis is True.

(i.e, The mean Filling quantity is not 500 ml)

CODing

* Z_cal = (505 - 500) / (10 / np.sqrt(30))

print (Z_cal)

Out :- 2.738612

# Pvalue = 2 * [1 - stats.norm.cdf (Z_cal)]

pvalue

Out 0.0061698

31/03/22
11:30 pm.

Dt:- 1/04/22
12:40 pm.

Case Study 4

## Fabric Data (1 Sample "Z" Test)

Question :4
The Length of 25 Samples of a Fabric are Taken at random. Mean and Standard deviation From the historic 2 years Study are 150 and 4 respectively. Test if the Current mean is greater than its historic mean, Assume $\alpha$ to be 0.05 → If They don't give

Ans :- Data is Given in Excel Sheet, Alpha value, it has Only one column So, it is one Sample. Take always $\alpha = 0.05$ (Standard Val

$\sigma = 4$

$\alpha = 0.05$, So, it is "z" test.

Statistic Test
- Parametric Test (Entire given Data is normal)
- Non parametric Test

* Steps involved in Continous Data (Only For) (Hypothesis Testing)

Step 1 :- Formulate, $H_0$ and $H_1$

Step 2 :- Level of significance ($\alpha$)

Step 3 :- check For Normality of given Data [For Continous Data only]

For Continous data only we have Histogram not For Discrete Data.

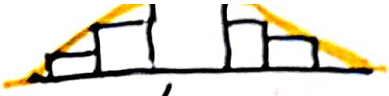
Step 4 :- Select Statistical Test and Calculate "p" Value

Step 5 :- Based on "p" value, Conclude hypothesis Test.

## * How To Find "Normality" in The Data

- Option 1 :- Skewness
- Option 2 :- Density Curve
- Option 3 :- Shapiro Test

## Option 1 [Skewness]



→ If the skewness is between -1 to +1 [normal Distribution]

→ if the skewness is Less Than -1 [Left Skewed]

→ if the skewness is greater Than +1 [right skewed]

## Option 2 [Density Curve]

→ If the density curve is Symmetrical [normal Distribution]

→ If the density curve is non Symmetrical [Skewed distribution]

## *** Option 3 [Shapiro test] For checking normality

Hypothesis Test

→ $H_0$ : Data is normal [if, $P > 0.05$ [Normal]]

→ $H_1$ : Data is not normal [If $P \leq 0.05$ [Not normal]]

---

EX:- Raymond Company. 150cm (sold) Said and

Shirt Fabric Sample - 147 cm ( But it having)

So, They want check samples.

* $H_0: \mu \geq 150$ [Fabric having 150 (or) more) No action

* $H_1: \mu < 150$ [Fabric having Less metres) action

Step : 2

Level of significance $\alpha = 0.05$

Step 3 Check the normality.

# option 1 (skewness)

# fabric ["Fabric_Length"].skew()

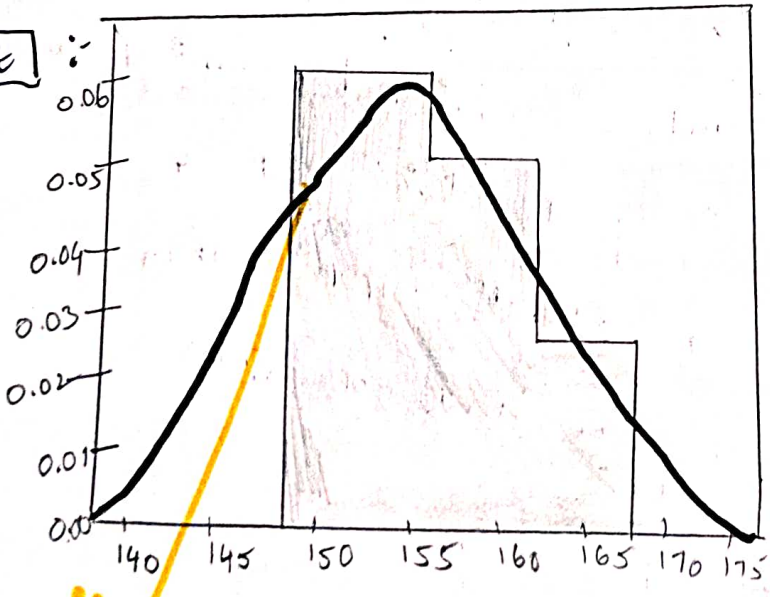
Out :- 0.296

# option 2 (Density curve)

# sns.displot(fabric["Fabric_Length"])

plt.show()

Out :




Here curve is Little Bended : How can we say it is "normal"

So, we take this as "Last option" in Find normal Distributions. (")

-1
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
1
+1.0
+0.1
+0.2
+0.3
+0.4
+0.5
+0.6
+0.7
+0.8
+0.9

# Option 3 [Shapiro Test]

# Stats.Shapiro (fabric["Fabric_Length"])

Out :- Shapiro result (statistic = 0.9397.., Pvalue = 0.1460)

# p value > 0.05, (Normal)

* **Doubt**

Going Back To question

$\mu = 150$

$\mu != 150$ → Ela kuda Thesukovachu Kadha ?

But in question,

We have Current Mean is "Greaterthan"

So, we Take

$H_0 - \mu \geq 150$

$H_1 - \mu < 150$

Step 4 Stastical Test. (One Sample - ZTest)

We have,

$$Z_{Test} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\mu = 150$

$\sigma = 4$

$n = 25$

$\bar{X} = ?$

Mean ↑ To Know ↓ x value

# fabric["Fabric_Length"].mean()

Out :- 155.063

# $Z_{Test} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

# Z_Test = (155 - 150) / (4 / np.sqrt(25))

Print Z_Test

Out :- 6.325 ⟶ ZTest = P(6.325)

Zscore = P(0.999)

# stats.norm.cdf(z_test)

Out :- 0.999


Left Tail Test, so we don't need (-1) to add.

Step 5 :- Based on P.value, (Accept/Reject)

P = 0.999 > α = 0.05

"P" High Null Fly (Accept)

So, we don't reject Null hypothesis.

Mean >= 150 (No action required)

To import :- Excel data

fabric = Pd.read_excel("D:\\data science\\Shivaramakrishna\\5. inferential statistics\\Fabric data.xlsx")

Case Study 5

Example 5 **BOLT Diameter (1 Sample T Test)**

The mean Diameter of Bolt manufactured should be 10 mm to be able to fit into nut. 20 Samples are Taken at random from production line by a quality inspector. Conduct a' Test to check with 95% confidence that the mean is not different from the Specific Value.

Answer :- Given, Data Set of 20 Samples.
Here ($\sigma$) is unknown so it is "T" test.

Step 1 :-

$H_0 : \mu = 10$ mm (No action)

$H_1 : \mu \neq 10$ mm (action)

Step 2 :-

$\alpha = 0.05$

$(1 - 0.95)$

Step 3 :- Check the normality.

```
# bolt ["Diameter"].skew()
```

out :- 0.3809

Step 4 :- Stastical Test. (1 sample "T" Test)

$$T_{Test} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

In coding

# Stats.. ttest_1Samp (bolt.Diameter, 10) ← mean vundha ledha?

Out :- Ttest_1Samp Result (stastic = 4.114 , pvalue = 0.0005)

Step5 :- Based on "p" value, Accept, Reject $H_0$

$P = 0.005 < \alpha = 0.05$

So, "p" Low = Null go (Reject)

Reject "Null hypothesis (Avg. bolt diameter is not equal to 10) Take action.

1 Sample = Ztest (If n<30, we have To Ztest)

1 Sample = TTest (If n>30, we Take TTest)

In Real life, we Take TTest mostly.

Case Study 6 (2 Sample Equal Variance)

Example 6 **MARKETING STRATEGY**

A Financial anaylst at a Financial institute wants to Evaluate a recent Credit Card promotion. after this promotion. 500 Card holders were Randomly Selected. Half received an ad promoting a Full waiver of intrest rate on purchases mader over the next three months, half received a standard christmas advertisment. Did the ad promoting Full intrest rate waiver, Increase purchases ?

* Understanding Question / with Different Example.

Ex:- FlipKart

365 Days

5 % on axis Bank

(On Festivals)
10 % Discount
on HDFC, SBI

So, it is Giving good Results

They are implementing Every Festival.

By doing Hypothesis Testing.

* $H_0$ : 5 % (365 Days) $\geq$ 10 % dis (Fstvl)
* $H_1$ : 5 % $<$ 10 %

```
# import numpy as np
# import matplot.pylot as plt
# import seaborn as sns
# From Scipy import stats
# import pandas as pd
```

Importing Data

```
# Promotion = pd.read_excel ("D:\\datascience\\Shiva rama Krishna class\\5. inferential statics\\Promotion.xlsx")
Promotion.head()
```

2 Sample Test.

Out :-

| Intrest Rate Waiver | Standard promotion |
|---|---|
| 1989.10 | 1272.35 |
| 1808.38 | 1250.38 |
| 1153.75 | 1474.78 |
| 1745.46 | 2064.89 |
| 1008.24 | 2030.87 |

Step 1 :-

$\mu$ (No action)

$H_0$ :- Avg. Purchases made by Full Intrest Waves $\leq$ Avg. Standard

$H_1$ : Avg. purchase by FIW $>$ Avg. Standard promotion (action)

Step 2

Level of significance = $\alpha$ (0.05)

$(1 - 0.95) \Rightarrow 1$

Step 3 (check Normality) → Test For "2" Samples.

# print (stats. shapiro (promotion . Intrest Rate Waiver))

Out :- Shapiro Result (Static = 0.992, p Value = 0.2245)
Greater Than 0.05
So Normal Distribution

# print (stats. Shapiro (promotion. Standard promotion)

Out :- Shapiro Result (stastitics = 0.991 , p value = 0.1915)
Greater Than 0.05
0.191 > 0.05

## Variance Test / Levene Test

Fixed [ $H_0$ :- Variances are Equal ✓
$H_1$ :- Variences are not Equal.

# stats. Levene (Promotion . Interest Rate Waiver, Promotion . Standard Promotion)

Out :- Levene Result (statistic = 1.13 , p value = 0.287)
Greater Than 0.05
So, Variences are Equal.

Here, Variences are Equal,
2 Samples are normal
we apply,

**2 Sample Test For Equal Variance**

Step 4 Select statistical Test & calculate 'p' value.

No ($\sigma$) value, so, we Take "T" Test

- 2 Sample "T" Test For Equal variance.

# stats.ttest_ind (promotion.InterestRateWaiver, Promotion.
Standard Promotion, equal_var = True)

Out :- TTest Result (Statistics = 2.260, Pvalue = 0.024)

$P = 0.024 < 0.05$

Step 5 :- Based on "p" value, Accept / Reject $H_0$

$P = 0.024 < 0.05$

Here
"p" value is Less Than "$\alpha$" value,

So,
"p" Low, Null Go (Reject)

→ Ha = Avg. of Purchases made by "Full Interest waiver" is 'Greater than' Avg. Purchases made by standard promotion.

1/4/22 4:30pm

Dt :- 8/2/2022 (12.57 Am 1/4/22 night)

Revising

Problem

$H_0$ : FIW $\leq$ SC
$H_1$ : FIW $>$ SC
} 2 sample "T" Test $\rightarrow$ $P < \alpha$

(i) Levene Test [Equal Variance]

$H_0$ :- Variences are Equal

$H_1$ :- Variences are not Equal

$\therefore$ $P \geq \alpha \Rightarrow$ Equal

$P \leq \alpha \Rightarrow$ not equal

(ii) Shapiro Test [Normality Test]

$H_0$ : Data is normal

$H_1$ : Data is not normal

$\therefore$ $P > \alpha \Rightarrow$ Normal

$P < \alpha \Rightarrow$ not norm

Flowchart :-

2 Samples
↓
Data is normal ——No——> Convert Skew To normal
↓ yes
Equal Varience ? ——No——> 2 sample "T" Test fo UnEqual Varience
↓ yes
2 Sample "T" Test For Equal Varience

Case Study 7 [ANOVA / FTest]

Example 7 **Contract Renewal**

A marketing organization out sources their back-office operations to three different Suppliers. The Contracts are up For renewal and CMO (Manager) wants to determine whether they should renew Contracts with all suppliers (or) any Specify Supplier. CMO want to renew The Contract of Supplier with Least Transaction time. CMO will renew all Contracts if the performance of all suppliers are Similar.

Renewal of Contracts

From question → 1. All are Equal _ if not
2. Less Time

Given Data, FTest ← ANOVA Test one way

3 Samples

Because, we have more than 2 Samples in the Data.

Supplier A Supplier B Supplier C

Import The Data.

Load Libraries :-

```
# import numpy as np
# import pandas as pd
# import matplot.pylot as plt
# import Seaborn as sns
# From Scipy import Stats
```

Load the data:

I always Draw Data from where it stored.

So, I write Total address of the Data.

Mostly, we save in "pwd". But I store in another Folder. So, I write Total address

```
#Contract = pd.read_excel("D:\\data science \\ Shiva Rama Krishna \\ 5. inferential statistics \\ Contract Renewal.xlsx")
```

```
Contract.head()
```

Out :-

| Supplier A | Supplier B | Supplier C |
|---|---|---|
| 6.15 | 7.87 | 7.41 |
| 6.22 | 5.21 | 3.61 |
| 6.76 | 7.94 | 7.23 |
| 4.29 | 7.36 | 5.53 |
| 7.08 | 6.17 | 3.97 |

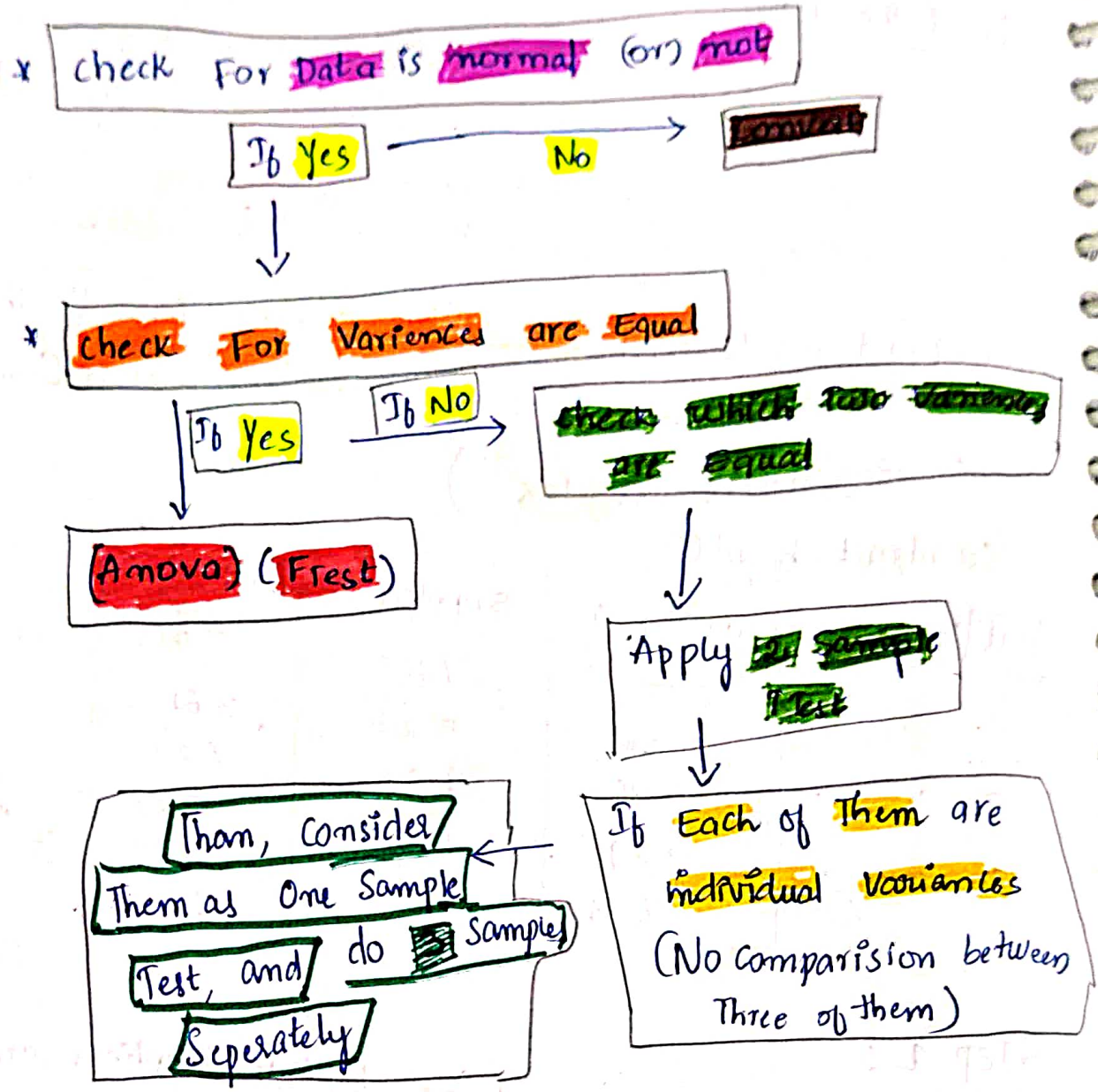
Step 1 :-

$H_0$ : Average Time by all Suppliers are Equal

$H_1$ : Average Time by all Suppliers not Equal

Step 2 :- Level of Significance $\alpha = 0.05$

Step 3 :- Check for normality




Normality :-

```
# Print ( stats.Shapiro (contract . Supplier A))
, Print ( Stats. Shapiro . (contract. Supplier B))
Print ( Stats. Shapiro. (Contract. Supplier C))
```

Out

```
Shapiro Result ( P.value = 0.89 )
"              [ P value = 0.64 ]  normal
"              ( P.value = 0.57 )
```

# Variance Test

# stats. Levene (Contract. Supplier A, Contract. Supplier B, Contract. Supplier C) 3 inputs

Out :- Levine Result (pvalue = 0.77)

p value > 0.05 (α)

Variences are Equal

Step 4 (Anova Test / F Test)

# stats. f_oneway (Contract. Supplier A, Contract. Supplier B, Contract. Supplier C)

Out :- F_one way Result (p value = 0.103)

P value > 0.05 (α)

From "p" value, P is Greater Than α,

'p' high, Null Fly (Accept)

We accept Null hypothesis.

= All the 3 suppliers have Equal mean Transaction Time.

And, If we don't have 3 supplies Equal variance Then check, what 2 samples Having Same Variance

# Print (stats. Levene (Contract. Supplier A, Contract Supp B))
# Print (" . " (Contract. Supplier B, Contract. Supplier C))
# Print (" . " (Contract. Supplier C, Contract. Supplier A))

* Discrete Variable

Case study 1

Example 1 **Foot ball Coach (1 proportion Test)**

The people carry out a poll to find the acceptability of a new foot ball coach. it was decided that if the Support rate for the coach for Entire population was Truly less than 25 %. The coach would be Fired - 2000 people participated and 482 people Supported the new coach. Conduct a Test to check If new coach should be Fired with 95 % level of confidence

question Explanation :- Sample people = 2000 (n)

$n = 2000$

(not Supporting) 1518

482 (Supporting)

But, it is Sample ⇒ If less Than 25 %. Fired.

May be in population $n = 10,000$ ← 7000

3000 = $\frac{3000}{7000}$ = 30 %.

$\frac{482}{1518} = 24.1$ From only we can conclude it is not reaching 25 %.

Another Similar Example :

Population

| Male |
| --- |
| Female |
| M |
| M |
| F |
| M |
| M |
| F |
| F |
| F |
| M |
| F |
| M |
| F |
| M |

$n = 15$

Male = 8

Female = 7

Sample Data

M, M, F, F

From the can, Conclude

$M = F \ [2 = 2]$

By sample Data,

But

Actual Population Data is

$M \neq F$

$8 \neq 7$

Step 1 :-

$H_0$ : Coach not be Fired [No action] / No Difference

$H_1$ : Coach to be Fired [action]

Step 2 :-

$\alpha = 0.05$

Step 3 :- check For normality / check varience/

We Don't have these in Discrete Data

Step 4 :- Select Statistical Test, Calculate "p" value.

* 1 proportion Test

Because, we Have Only One column

Support → Total → Less than 25%

# Stats. binom_test (482, 2000, 0.25)

(No data xlsx given)

Directly written

OUT :0.36615

"p" value 0.366 > 0.05

"p" High Null Fly (accept)

Do not reject null hypothesis.

Case Study 2

Example 2 **Johny Talkers (2 Proportion Test)**

Johny Talker Soft drinks division Sales manager has been planning To Launch a new Sales incentive Program For their Sales Executives. The Sales Executives Felt that adults (>40yrs) won't buy, children will & hence requested Sales manager not to Lunch Programs. Analyze data & determine whether there is evidence at 5 % Significance Level support the hypothesis

```
# df = pd.read_excel ("Johny Tallbers.xlsx")
df.head()
```

written As it is class video But, Address must change.

Out :-

| | Person | Drinks |
|---|---|---|
| 0 | Adults | Did not purch |
| 1 | Adults | did not purch |
| 2 | Adults | did n't purchas |
| 3 | Adults | did n't purch |
| 4 | Adults | did n't purc |

```
# df["Person"].value_counts()
```

Out: children 740
Adults 480

Total = 1220 → 740 (C), 480 (A)

```
# df["Drinks"].value_counts()
```

Out :- did n't purchase 1010
purchase 210

Total (1220) → 1010 (d.P), 210 (P)

#CrossTable

```
# pd.CrossTab (df["Person"], df["Drinks", margins=True)
```

margins → total

Out :-

| Drinks | Did n't Purchase | Purchased | All |
|---|---|---|---|
| Person | | | |
| → Adults | 422 | 58 | 480 |
| → Children | 588 | 152 | 740 |
| → All | 1010 | 210 | 1220 |

By, 12, 20. are Equal (or) not By seeing Sample Data we say it is not Equal,

But what about → Here Adults (purchased) / 480 = 12 %, childrens (Purchased) / 740 = 20 % population data! Equal (or) not

Step 1 :-

$H_0$ : Proportion of Adults $\geq$ Proportion of Children

$H_1$ : Proportion of Adults $<$ Proportion of Children

Step 2

$\alpha = 0.05$

Step 3 :- No Normal / No variance — Discrete data

Step 4 :- 2 proportion Test

```
From statsmodels.stats.proportion import proportions_ztest

pur = np.array([58,152]) # How many childrens are purchas.
total = np.array([480, 740]) # Total no. of adults/childrn
                               are there.

proportions_ztest(pur,total, alternative = "two-sided")
```

Equal (or) not, $=, \neq,$

Smaller $<$

Larger $>$

Out :- (-3.82, 0.00013)

$p = 0.00013 < 0.05$

Step 5 :- "p" Low Null Go (Reject)

Sufyl

2 | 4 | 22

4:20 Am

1 2 3 ⇒ 3 proportions

| Person | Drinks | class |
|---|---|---|
| Adult | Purchase | Low |
| Adult | not. P | middle |
| Adult | not. P | upper middle |
| Children | not. P | High |
| Adult | not. P | Low |
| Children | Purchase | Low |

In
* Person → Adult / children
* Drinks → Purchased / not purchased
* class → Low / middle / u.m / High.

Dt:- 24/4/22
12:00pm

## Case Study 3 (chi square Test)

Example 3

## Bahaman ReSearch

Baba Man Tech Research Company uses 4 regional Centers in South Asia [India, China, Srilanka, Bangladesh] To input data of questionnaire responses. They audit a Certain % of questionnaire response versus data entry. Any Error in data entry renders it defective. The chief data scientist wants to check whether the defective % Varies by Country. analyze data 5 % Level of signif help the manager draw appropriate inferences.

["1" means not deffective, '0' means defectiv

# Load data set

```
# Bahaman = pd.read_excel("Bahaman.xlsx")

Bahaman.head()
```

Out :-

| defective | Country |
|---|---|
| 0 | India |
| 0 | India |
| 0 | India |
| 0 | India |
| 1 | India |

```
# Count = pd.crosstab(Bahaman["Defective"], Bahaman["Count"])

Count
```

| Country / Defective | Bangladesh | China | India | Srilanka |
|---|---|---|---|---|
| 0 | 183 | 179 | 175 | 178 |
| 1 | 17 | 21 | 25 | 22 |
| | 200 | 200 | 200 | 200 |

17/200 , 21/200 , 25/200 , 22/200 = Equal or not

Step 1 :

$H_0$ : No difference in proportion between Countries

$H_1$ : difference b/w countries

Step 2

Level of significance $\alpha = 0.05$

Step 3

No Normality / No Variance → Because it is Discrete data

Step 4 Chisquare Test (because, we have more Than 2proportion) so, we Take Chisquare Test in Discrete data.

Contingency

# stats. chi2 — contingency (count) → Cross Tab Table.

OUT :- p value (0.63)

$p(0.63) > 0.05 (\alpha)$

"p" high Null Fly (Accept)

we accept The hypothesis.

— coding Completed ✓

Inferential Statistics Completed ,

Suf 2/4/22

3:06pm