**Hypothesis Testing Exercise:**

1. A F&B manager wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured? Analyze the data and draw inferences at 5% significance level. Please state the assumptions and tests that you carried out to check validity of the assumptions.

Minitab File : **Cutlets.mtw**

**Ans:**

import pandas as pd

import numpy as np

from scipy import stats

from scipy.stats import norm

*# Load the dataset*

data**=**pd**.**read\_csv('Database/Cutlets.csv')

data**.**head()

|  | **Unit A** | **Unit B** |
| --- | --- | --- |
| **0** | 6.8090 | 6.7703 |
| **1** | 6.4376 | 7.5093 |
| **2** | 6.9157 | 6.7300 |
| **3** | 7.3012 | 6.7878 |
| **4** | 7.4488 | 7.1522 |

unitA**=**pd**.**Series(data**.**iloc[:,0])

unitA

0 6.8090

1 6.4376

2 6.9157

3 7.3012

4 7.4488

5 7.3871

6 6.8755

7 7.0621

8 6.6840

9 6.8236

10 7.3930

11 7.5169

12 6.9246

13 6.9256

14 6.5797

15 6.8394

16 6.5970

17 7.2705

18 7.2828

19 7.3495

20 6.9438

21 7.1560

22 6.5341

23 7.2854

24 6.9952

25 6.8568

26 7.2163

27 6.6801

28 6.9431

29 7.0852

30 6.7794

31 7.2783

32 7.1561

33 7.3943

34 6.9405

Name: Unit A, dtype: float64

unitB**=**pd**.**Series(data**.**iloc[:,1])

unitB

0 6.7703

1 7.5093

2 6.7300

3 6.7878

4 7.1522

5 6.8110

6 7.2212

7 6.6606

8 7.2402

9 7.0503

10 6.8810

11 7.4059

12 6.7652

13 6.0380

14 7.1581

15 7.0240

16 6.6672

17 7.4314

18 7.3070

19 6.7478

20 6.8889

21 7.4220

22 6.5217

23 7.1688

24 6.7594

25 6.9399

26 7.0133

27 6.9182

28 6.3346

29 7.5459

30 7.0992

31 7.1180

32 6.6965

33 6.5780

34 7.3875

Name: Unit B, dtype: float64

*# 2-sample 2-tail ttest: stats.ttest\_ind(array1,array2) # ind -> independent samples*

p\_value**=**stats**.**ttest\_ind(unitA,unitB)

p\_value

Ttest\_indResult(statistic=0.7228688704678061, pvalue=0.47223947245995)

p\_value[1] *# 2-tail probability*

0.47223947245995

2. A hospital wants to determine whether there is any difference in the average Turn Around Time (TAT) of reports of the laboratories on their preferred list. They collected a random sample and recorded TAT for reports of 4 laboratories. TAT is defined as sample collected to report dispatch.

Analyze the data and determine whether there is any difference in average TAT among the different laboratories at 5% significance level.

Minitab File: **LabTAT.mtw**

**Ans:**

import pandas as pd

import numpy as np

from scipy import stats

from scipy.stats import norm

*# load the dataset*

data**=**pd**.**read\_csv('Database/LabTAT.csv')

data**.**head()

|  | **Laboratory 1** | **Laboratory 2** | **Laboratory 3** | **Laboratory 4** |
| --- | --- | --- | --- | --- |
| **0** | 185.35 | 165.53 | 176.70 | 166.13 |
| **1** | 170.49 | 185.91 | 198.45 | 160.79 |
| **2** | 192.77 | 194.92 | 201.23 | 185.18 |
| **3** | 177.33 | 183.00 | 199.61 | 176.42 |
| **4** | 193.41 | 169.57 | 204.63 | 152.60 |

*# Anova ftest statistics: stats.f\_oneway(column-1,column-2,column-3,column-4)*

p\_value**=**stats**.**f\_oneway(data**.**iloc[:,0],data**.**iloc[:,1],data**.**iloc[:,2],data**.**iloc[:,3])

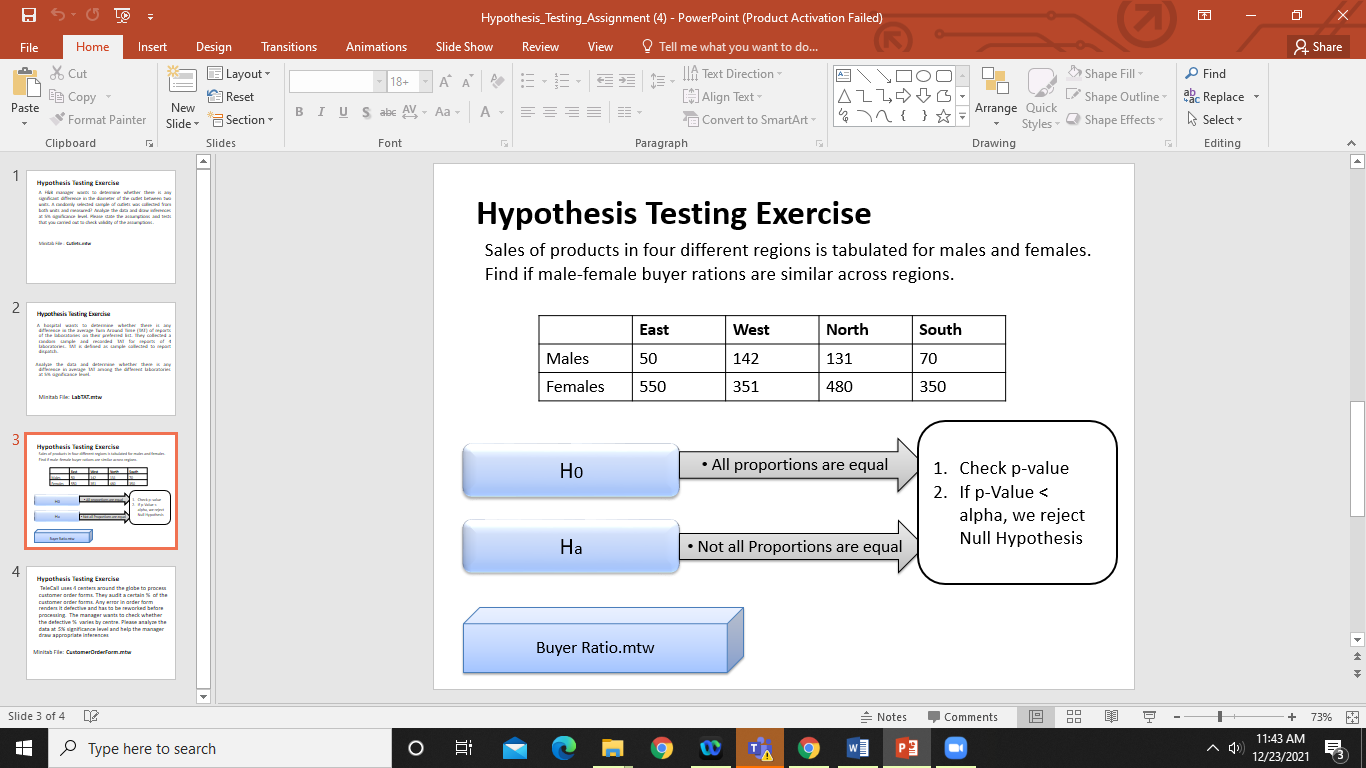
p\_value

F\_onewayResult(statistic=118.70421654401437, pvalue=2.1156708949992414e-57)

p\_value[1] *# compare it with α = 0.05*

2.1156708949992414e-57

3.



Ans:

import pandas as pd

from scipy import stats as stats

import numpy as np

df= pd.read\_csv('C:/Users/Admin/Downloads/BuyerRatio.csv')

df.head()

|  | Observed Values | East | West | North | South |
| --- | --- | --- | --- | --- | --- |
| 0 | Males | 50 | 142 | 131 | 70 |
| 1 | Females | 435 | 1523 | 1356 | 750 |

df\_table=df.iloc[:,1:6]

df\_table

|  | East | West | North | South |
| --- | --- | --- | --- | --- |
| 0 | 50 | 142 | 131 | 70 |
| 1 | 435 | 1523 | 1356 | 750 |

df\_table.values

array([[ 50, 142, 131, 70],

[ 435, 1523, 1356, 750]], dtype=int64)

val=stats.chi2\_contingency(df\_table)

val

(1.595945538661058,

0.6603094907091882,

3,

array([[ 42.76531299, 146.81287862, 131.11756787, 72.30424052],

[ 442.23468701, 1518.18712138, 1355.88243213, 747.69575948]]))

type(val)

tuple

no\_of\_rows=len(df\_table.iloc[0:2,0])

no\_of\_columns=len(df\_table.iloc[0,0:4])

degree\_of\_f=(no\_of\_rows-1)\*(no\_of\_columns-1)

print('Degree of Freedom=',degree\_of\_f)

Degree of Freedom= 3

Expected\_value=val[3]

Expected\_value

array([[ 42.76531299, 146.81287862, 131.11756787, 72.30424052],

[ 442.23468701, 1518.18712138, 1355.88243213, 747.69575948]])

from scipy.stats import chi2

chi\_square=sum([(o-e)\*\*2/e for o,e in zip(df\_table.values,Expected\_value)])

chi\_square\_statestic=chi\_square[0]+chi\_square[1]

chi\_square\_statestic

1.5152956451130446

critical\_value=chi2.ppf(0.95,3)

critical\_value

7.814727903251179

if chi\_square\_statestic >= critical\_value:

print('Dependent (reject H0)')

else:

print('Independent (fail to reject H0)')

Independent (fail to reject H0)

pvalue=1-chi2.cdf(chi\_square\_statestic,3)

pvalue

0.6787446296467897

if pvalue <= 0.05:

print('Dependent (reject H0)')

else:

print('Independent (fail to reject H0)')

Independent (fail to reject H0)

4. TeleCall uses 4 centers around the globe to process customer order forms. They audit a certain % of the customer order forms. Any error in order form renders it defective and has to be reworked before processing. The manager wants to check whether the defective % varies by centre. Please analyze the data at *5%* significance level and help the manager draw appropriate inferences

Minitab File: **CustomerOrderForm.mtw**

**Ans:**

Assume Null Hypothesis as Ho: Independence of categorical variables (customer order forms defective % does not varies by centre) Thus, Alternative hypothesis as Ha Dependence of categorical variables (customer order forms defective % varies by centre)

Inference: As (p\_value = 0.2771) > ( = 0.05); Accept Null Hypthesis i.e. Independence of categorical variables Thus, customer order forms defective % does not varies by centre.