Hypothesis Testing

Business problem 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.

File : **Cutlets.mtw**

Solution:

Step 1: Analyzing weather there is any significant difference in the diameter of the cutlet between two units.

Step 2: By comparing the data, we can conclude that the data given is continuous in nature.

Step 3: Check whether the data is following a normal distribution or not

Unit_A:

H_0: Unit_A follows normal distribution.

H_1: Unit_A doesn't follows normal distribution.

By performing ***Shapiro Test***, normality will be decided

Codes :

```
import pandas as pd

import scipy

from scipy import stats

import statsmodels.api as sm


cutlet=pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\Cutlets.csv")

cutlet

cutlet.columns


print(stats.shapiro(cutlet.Unit_A))
```

(0.9649458527565002, 0.3199819028377533)

The p-value for Unit_A is greater than 0.05

Therefore, p high H_0 fly

Therefore Unit_A follows normal distribution.

Unit_B :

H_0: Unit_B follows normal distribution.

H_1: Unit_B doesn't follows normal distribution.

Codes:

```
print(stats.shapiro(cutlet.Unit_B))
```

(0.9727300405502319, 0.5224985480308533)

The p-value for Unit_B is greater than 0.05

Therefore, p high H_0 fly

Therefore Unit_B follows normal distribution.

So we can conclude that the all the populations does follows a normal distribution. So we can proceed to variance equality test.

Step 4: Check whether the external conditions are same or not.

Here the external conditions are different, since the diameter of the cutlets cannot be identical may vary because the persons cooking may not be the same.

Step 5: Check whether the variances are equal or not.

H_0: Unit_A and Unit_B have equal variance.

H_1: Unit_A and Unit_B doesn't have equal variance.

To check variances are equal or not we perform ***Leven's Test.***

Codes:

```
scipy.stats.levene(cutlet.Unit_A, cutlet.Unit_B)
```

LeveneResult(statistic=0.665089763863238, pvalue=0.4176162212502553)

The p value is greater than 0.05, therefore p high H_0 fly.

Therefore Unit_A and Unit_B have equal variance.

Step 6: Check whether Unit_A and Unit_B have equal diameter or not.

H_0: Unit_A and Unit_B don't have equal diameter.

H_1: Unit_A and Unit_B have equal diameter.

To check whether Unit_A and Unit_B have equal diameter or not we perform 2 Sample T-Test.

Codes:

scipy.stats.ttest_ind(cutlet.Unit_A,cutlet.Unit_B)

Ttest_indResult(statistic=0.7228688704678061, pvalue=0.4722394724599501)

The p value is greater than 0.05, therefore p high H_0 fly.

Therefore Unit_A and Unit_B don't have equal Diameter.

```
 7
 8 import pandas as pd
 9 import scipy
10 from scipy import stats
11 import statsmodels.api as sm
12
13 cutlet=pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\Cutlets.csv")
14 cutlet
15 cutlet.columns
16
17
18 ##########Normality Test ############
19 print(stats.shapiro(cutlet.Unit_A))
20
21 print(stats.shapiro(cutlet.Unit_B))
22
23 ##################### Variance Test #############
24 scipy.stats.levene(cutlet.Unit_A, cutlet.Unit_B)
25
26 ######## 2 Sample T test #################
27 scipy.stats.ttest_ind(cutlet.Unit_A,cutlet.Unit_B)
28
```

Business problem 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.

Solution:

Step 1: Analyzing weather there is any difference in the Turn around Time (TAT) in Labs T 5% singnificance level.

Step 2: There are more than 2 populations and by comparing the data, we can conclude that the data given is continuous in nature.

Step 3: Check whether the data is following a normal distribution or not

Lab1: **H_0: Lab1 data follows normal distribution.**

H_1: Lab1 data doesn't follows normal distribution.

By performing ***Shapiro Test***, normality will be decided

Codes :

```
import pandas as pd

import scipy

from scipy import stats

import statsmodels.api as sm


LabTat = pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\LabTAT.csv")

LabTat

LabTat.columns


print(stats.shapiro(LabTat.Lab1))
```

Out: (0.9901824593544006, 0.5506953597068787)

The p-value for Unit_A is greater than 0.05

Therefore, p high H_0 fly

Therefore Lab1 follows normal distribution.

Lab2: **H_0: Lab2 data follows normal distribution.**

H_1: Lab2 data doesn't follows normal distribution.

print(stats.shapiro(LabTat.Lab2))

Out: (0.9936322569847107, 0.8637524843215942)

The p-value for Unit_A is greater than 0.05

Therefore, p high H_0 fly

Therefore Lab2 follows normal distribution.

Lab3: **H_0: Lab3 data follows normal distribution.**

H_1: Lab3 data doesn't follows normal distribution.

print(stats.shapiro(LabTat.Lab3))

Out: (0.9886345267295837, 0.4205053448677063)

The p-value for Unit_A is greater than 0.05

Therefore, p high H_0 fly

Therefore Lab3 follows normal distribution.

Lab4: **H_0: Lab4 data follows normal distribution.**

H_1: Lab4 data doesn't follows normal distribution.

print(stats.shapiro(LabTat.Lab4))

Out: (0.9913753271102905, 0.6618951559066772)

The p-value for Lab4 is greater than 0.05

Therefore, p high H_0 fly

Therefore Lab4 follows normal distribution.

So we can conclude that the all the populations does follows a normal distribution. So we can proceed to variance equality test.

Step 4: Check whether the external conditions are same or not.

Here the external conditions are different, since the timings may vary and cannot be the specific.

Step 5: Check whether the variances are equal or not.

H_0: Lab1, Lab2, Lab3, and Lab4 have equal variance.

H_1: Lab1, Lab2, Lab3, and Lab4 doesn't have equal variance.

To check variances are equal or not we perform ***Leven's Test.***

Codes:

```
scipy.stats.levene(LabTat.Lab1, LabTat.Lab2, LabTat.Lab3, LabTat.Lab4)
```

```
LeveneResult(statistic=2.599642500418024, pvalue=0.05161343808309816)
```

The p value is greater than 0.05.

Therefore, p high H_0 fly.

Therefore, Lab1, Lab2, Lab3, and Lab4 are having equal variances.

Step 6: Performing 1-way ANOVA test, since the variances of all the populations are same.

```
from statsmodels.formula.api import ols
mod=ols('Lab1~Lab2+Lab3+Lab4',data= LabTat).fit()
table=sm.stats.anova_lm(mod,type=2)
print(table)
```

Out:

	df	sum_sq	mean_sq	F	PR(>F)
Lab2	1.0	332.030416	332.030416	1.940311	0.166299
Lab3	1.0	203.853111	203.853111	1.191271	0.277335
Lab4	1.0	265.614707	265.614707	1.552192	0.215323
Residual	116.0	19850.186366	171.122296	NaN	NaN

Here the vaue of p > 0.05

Fail to reject Ho

Therefore, The average Turn Around Time (TAT) of reports of the laboratories are equal.

Conclusion: There is no any significant difference in the average Turn Around Time (TAT) of reports of the laboratories.

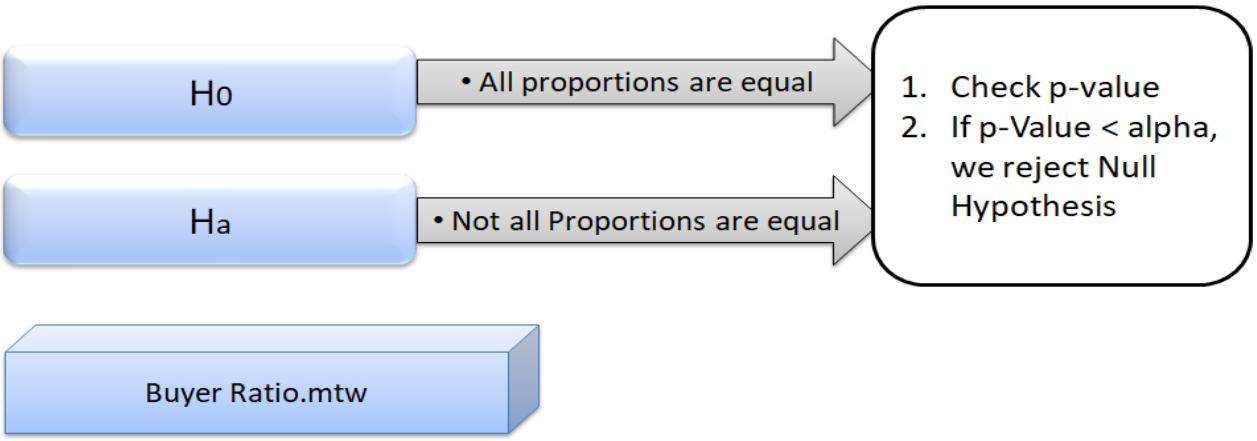
```python
import pandas as pd
import scipy
from scipy import stats
import statsmodels.api as sm

#from plotly.tools import FigureFactory as FF

############2 sample T Test(Marketing Strategy) ##################

LabTat = pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\LabTAT.csv")
LabTat
LabTat.columns

##########Normality Test ############

print(stats.shapiro(LabTat.Lab1))    #Shapiro Test

print(stats.shapiro(LabTat.Lab2))

print(stats.shapiro(LabTat.Lab3))

print(stats.shapiro(LabTat.Lab4))

help(stats.shapiro)

#################### Variance Test #############
scipy.stats.levene(LabTat.Lab1, LabTat.Lab2, LabTat.Lab3, LabTat.Lab4)


#################### One Way ANOVA Test #############

from statsmodels.formula.api import ols
mod=ols('Lab1~Lab2+Lab3+Lab4',data= LabTat).fit()
table=sm.stats.anova_lm(mod,type=2)
print(table)
```

Business problem 3:

Hypothesis Testing Exercise

Sales of products in four different regions is tabulated for males and females. Find if male-female buyer rations are similar across regions.

	East	West	North	South
Males	50	142	131	70
Females	550	351	480	350



Solution:

Step 1: To check whether male-female buyer rations are similar across regions or not.

Step 2 : By comparing the data, we can conclude that the data given is Discrete in nature

Step 3 : Defining Hypothesis.

- **Ho : male-female rations are similar.**
- **Ha : male-female rations are not similar.**

Step3 : Selection of test – Chi-squared test.

- **Ho : male-female rations are similar.**
- **Ha : male-female rations are not similar.**

Codes:

```
import pandas as pd
import numpy as np
Buyerdata= pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\BuyerRatio.csv")

male=[50,142,131,70]
female=[435,1523,1356,750]

buyr=np.array([male,female])
buyr
```

Out:
```
array([[  50,  142,  131,   70],
       [ 435, 1523, 1356,  750]])
```

```
from scipy import stats
chi2_stat,p_val,dof,ex=stats.chi2_contingency(buyr)
print(chi2_stat)
```

Out: 1.595945538661058

```
print(p_val)
```

Out: 0.6603094907091882

```
print(dof)
```

Out: 3

```
print(ex)
```

Out:
```
[[  42.76531299  146.81287862  131.11756787   72.30424052]
 [ 442.23468701 1518.18712138 1355.88243213  747.69575948]]
```

Here the value of p > 0.05

Fail to reject Ho

Therefore, Male-female rations are similar.

Conclusion: In four different regions male-female buyer ratios are similar across regions.

```python
import pandas as pd
import numpy as np
from scipy import stats


Buyerdata= pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\BuyerRatio.csv")

male=[50,142,131,70]
female=[435,1523,1356,750]

buyr=np.array([male,female])

################Chi-Square Test ################

chi2_stat,p_val,dof,ex=stats.chi2_contingency(buyr)
print(chi2_stat)
print(p_val)
print(dof)
print(ex)

```

Business problem 4:

Tele Call 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

Solution:

Step 1: To check whether the defective percentage varies by centre or not.

Step 2: Discrete data, Comparing more than two population with each other.

Step 3: Test selection : Chi-Square Test.

Ho : Defective percentage is same.

Ha : Defective percentage varies.

Codes:

```
import pandas as pd
import numpy as np

COF = pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\Costomer+OrderForm.csv")

import scipy
from scipy import stats

count = pd.crosstab([COF.Phillippines,COF.Indonesia],[ COF.Malta, COF.India])
count
```

Out:

Malta		Defective	Error Free	
India		Error Free	Defective	Error Free
Phillippines	Indonesia			
Defective	Defective	0	0	2
	Error Free	4	2	21
Error Free	Defective	9	2	20
	Error Free	18	16	206

```
Chisquares_results = scipy.stats.chi2_contingency(count)
Chisquares_results
```

Out:

```
(14.909686613258446,
 0.020970901162272855,
 6,
 array([[2.06666667e-01, 1.33333333e-01, 1.66000000e+00],
        [2.79000000e+00, 1.80000000e+00, 2.24100000e+01],
        [3.20333333e+00, 2.06666667e+00, 2.57300000e+01],
        [2.48000000e+01, 1.60000000e+01, 1.99200000e+02]]))
```

Here the value of p > 0.05

Therefore, Fail to reject Ho

Therefore, Defective percentage is same

Conclusion: The Defective percentage is same

```
import pandas as pd
import numpy as np

COF = pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\Costomer+OrderForm.csv")



################Chi-Square Test ################

import scipy
from scipy import stats

count = pd.crosstab([COF.Phillippines,COF.Indonesia],[COF.Malta,COF.India])
count


Chisquares_results = scipy.stats.chi2_contingency(count)
Chisquares_results

```

Business problem 5:

Fantaloons Sales managers commented that % of males versus females walking in to the store differ based on day of the week. Analyze the data and determine whether there is evidence at 5 % significance level to support this hypothesis.

Solution:

Step 1: % of males versus females walking in to the store differ or same based on day of the week.

Step 2: Discrete data, Comparing more than two population with each other.

Step 3: test selection: 2 sample Test.

Step 4: Defining Hypothesis.

Case 1 :

Ho: % of males is equal to females.

Ha: % of males is not equal to females.

Codes:

```
import pandas as pd
import numpy as np
import scipy as stats
import statsmodels.api as sm

FL= pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\Faltoons.csv")

count = pd.crosstab(FL["Weekdays"],FL["Weekend"])

n1=113
113/400
p1=0.2825

n2=167
167/400
p2=0.4175

population1 = np.random.binomial(1, p1, n1)
population2 = np.random.binomial(1, p2, n2)
sm.stats.ttest_ind(population1, population2)
```

Out: (-3.2919245374040362, 0.0011236561392755 87, 278.0)

Here p-value < 0.05

Reject Null Hypothesis.

Therefore, % of males is not equal to females.

Case 2 :

Ho: % of males is > females

Ha: % of males is < females

sm.stats.ttest_ind(population1, population2,alternative = "smaller")

Out: (-3.2919245374040362, 0.0005618280696377935, 278.0)

Here p-value < 0.05

Reject Null Hypothesis.

Therefore, % of males is < females.

Conclusion :

% of males versus females walking is different . Therefore comments made by sales manager is right.

```
import pandas as pd
import numpy as np
import scipy as stats
import statsmodels.api as sm

FL= pd.read_csv("E:\Data\Assignments\i made\hypothesis testing\Faltoons.csv")

count = pd.crosstab(FL["Weekdays"],FL["Weekend"])

n1=113
113/400
p1=0.2825

n2=167
167/400
p2=0.4175

population1 = np.random.binomial(1, p1, n1)
population2 = np.random.binomial(1, p2, n2)
sm.stats.ttest_ind(population1, population2)


sm.stats.ttest_ind(population1, population2,alternative = "smaller")

```