Q.No. 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.

Answer:

We have to find out below;

Null Hypothesis (H0): There is significant difference in diameters of cutlets between two units.

Alternative Hypothesis (Ha): There is No difference in diameters of cutlets between two units.

Step 1. Import Necessary Libraries:

In [41]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import statsmodels.api as sm
import matplotlib.pyplot as plt

from scipy import stats
from scipy.stats import norm

import warnings
warnings.filterwarnings('ignore')
```

Step 2. Import Data or Dataset:

In [3]:
```
cutlets_details = pd.read_csv('Cutlets.csv')
cutlets_details.head(5)
```

Out[3]:

	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

Step 3. Data Understanding:

In [4]:
```
cutlets_details.shape
```

Out[4]:
```
(35, 2)
```

In [5]:
```
cutlets_details.dtypes
```

Out[5]:
```
Unit A    float64
Unit B    float64
dtype: object
```

In [6]:
```
cutlets_details.describe() #Just look at Mean and SD
```

Out[6]:

	Unit A	Unit B
count	35.000000	35.000000
mean	7.019091	6.964297
std	0.288408	0.343401
min	6.437600	6.038000
25%	6.831500	6.753600
50%	6.943800	6.939900
75%	7.280550	7.195000
max	7.516900	7.545900

In [7]:
```
cutlets_details.isnull().sum() # Checking any null entries in both datasets
```

Out[7]:
```
Unit A    0
Unit B    0
dtype: int64
```

In [8]:
```
unit_A = cutlets_details['Unit A'] #Taking Unit-A in series
unit_A.head(5)
```

Out[8]:
```
0    6.8090
1    6.4376
2    6.9157
3    7.3012
4    7.4488
Name: Unit A, dtype: float64
```

In [9]:
```
unit_B = cutlets_details['Unit B'] #Taking Unit-B in series
unit_B.head(5)
```

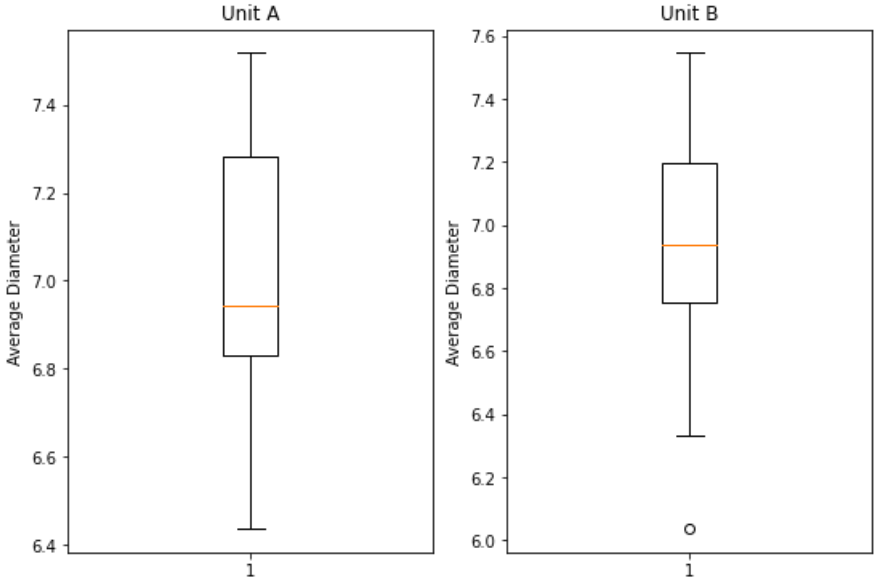
Out[9]:
```
0    6.7703
1    7.5093
2    6.7300
3    6.7878
4    7.1522
Name: Unit B, dtype: float64
```

In [10]:
```
cutlets_details[cutlets_details.duplicated()].shape #Checking duplicate values.
```

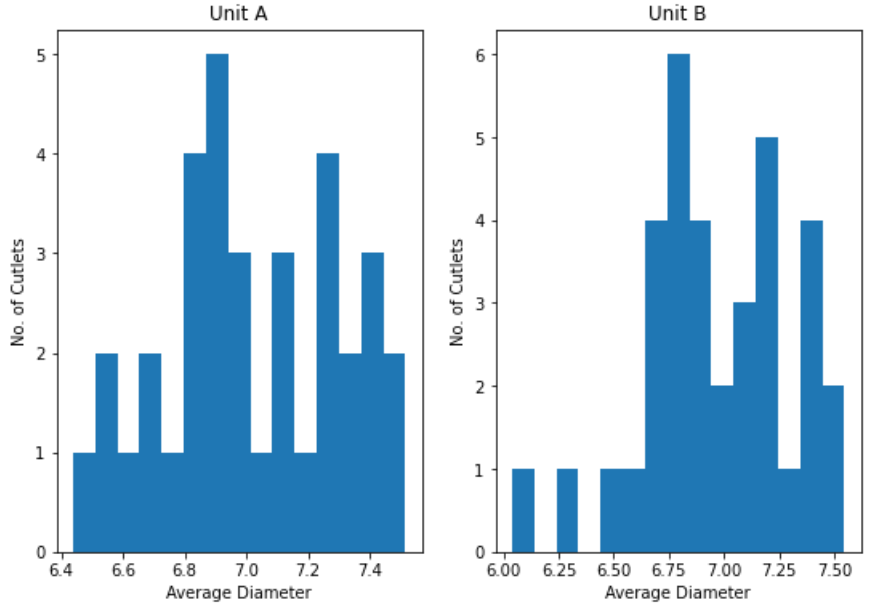
Out[10]:
```
(0, 2)
```

Step 4. Plotting data in different way:

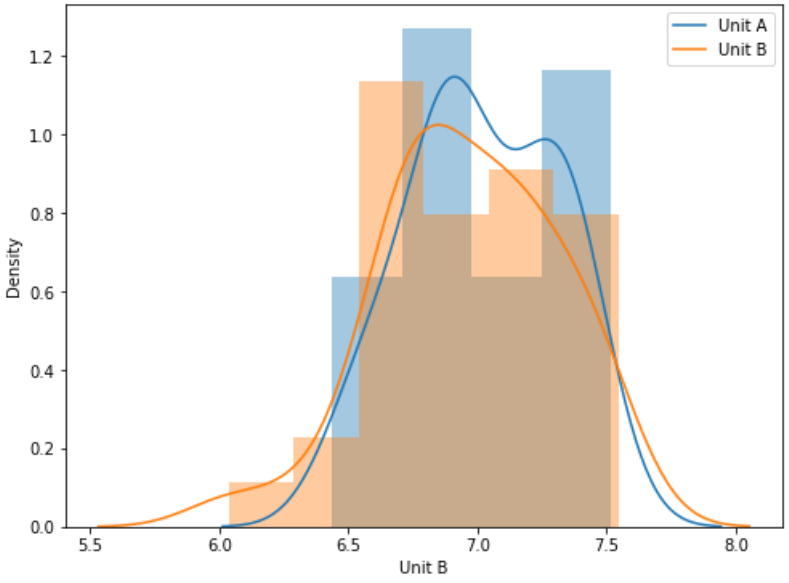
In [11]:
```
plt.subplots(figsize = (9,6))                          # boxplot
plt.subplot(121)
plt.boxplot(cutlets_details['Unit A'])
plt.title('Unit A')
plt.ylabel('Average Diameter')

plt.subplot(122)
plt.boxplot(cutlets_details['Unit B'])
plt.title('Unit B')
plt.ylabel('Average Diameter')

plt.show()
```



In [12]:
```
plt.subplots(figsize = (9,6))                          # subplot
plt.subplot(121)
plt.hist(cutlets_details['Unit A'], bins = 15)
plt.title('Unit A')
plt.xlabel('Average Diameter')
plt.ylabel('No. of Cutlets')

plt.subplot(122)
plt.hist(cutlets_details['Unit B'], bins = 15)
plt.title('Unit B')
plt.xlabel('Average Diameter')
plt.ylabel('No. of Cutlets')

plt.show()
```



In [13]:
```
plt.figure(figsize = (8,6))                                    # distplot
labels = ['Unit A', 'Unit B']
sns.distplot(cutlets_details['Unit A'], kde = True)
sns.distplot(cutlets_details['Unit B'],hist = True)
plt.legend(labels)
plt.show()
```



Step 5. Calculating p-value:

Here we have applied 2-Sample 2-Tail Test using t-statistics:

In [14]:
```
statistic,p_value = stats.ttest_ind(a = unit_A, b = unit_B, alternative='two-sided') #Independant
print('p_value is=',p_value)
```

```
p_value is= 0.4722394724599501
```

Step 5. Hypothesis Testing and Interpretation of p-value:

Null Hypothesis (H0): There is significant difference in diameters of cutlets between two units.

Alternative Hypothesis (Ha): There is No difference in diameters of cutlets between two units.

In [83]:
```
# Level of significance = 5% ie, At 5% level of significance, do we reject or not reject?
# alpha = 0.05
# Since it is 2-tailed test we have to divide alpha by 2: 0.05/2 = 0.025

if p_value<=0.025:
    print('We reject the Null Hypothesis and we can claim that there is a significant difference in diameters of cutlets between two units')
else:
    print('We do not reject the Null Hypothesis and we can claim that there is no difference in diameters of cutlets between two units')
```

```
We do not reject the Null Hypothesis and we can claim that there is no difference in diameters of cutlets between two units
```

Hence, We failed to reject the Null Hypothesis because of lack of evidence, there is no significant difference in diameters of cutlets between the two units.

==========

Q. No. 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.

Answer:

We have to find out below;

Null Hypothesis (H0): There is significant difference in average TAT amongest the different labs.

Alternative Hypothesis (Ha): There is No difference in average TAT amongest the different labs.

Step 1. Import Data or Dataset:

In [16]:
```
lab_details = pd.read_csv('LabTAT.csv')
lab_details.head(5)
```

Out[16]:

	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

Step 2. Data understanding:

In [17]:
```
lab_details.shape
```

Out[17]:
```
(120, 4)
```

In [18]:
```
lab_details.isnull().sum()
```

Out[18]:
```
Laboratory 1    0
Laboratory 2    0
Laboratory 3    0
Laboratory 4    0
dtype: int64
```

In [19]:
```
lab_details.describe()
```

Out[19]:

	Laboratory 1	Laboratory 2	Laboratory 3	Laboratory 4
count	120.000000	120.000000	120.000000	120.00000
mean	178.361583	178.902917	199.913250	163.68275
std	13.173594	14.957114	16.539033	15.08508
min	138.300000	140.550000	159.690000	124.06000
25%	170.335000	168.025000	188.232500	154.05000
50%	178.530000	178.870000	199.805000	164.42500
75%	186.535000	189.112500	211.332500	172.88250
max	216.390000	217.860000	238.700000	205.18000

In [20]:
```
lab_details.dtypes
```

Out[20]:
```
Laboratory 1    float64
Laboratory 2    float64
Laboratory 3    float64
Laboratory 4    float64
dtype: object
```

In [21]:
```
lab_1 = lab_details['Laboratory 1'] #Taking Laboratory 1 in a Series.
lab_1.head(5)
```

Out[21]:
```
0    185.35
1    170.49
2    192.77
3    177.33
4    193.41
Name: Laboratory 1, dtype: float64
```

In [22]:
```
lab_2 = lab_details['Laboratory 2'] #Taking Laboratory 2 in a Series.
lab_2.head(5)
```

Out[22]:
```
0    165.53
1    185.91
2    194.92
3    183.00
4    169.57
Name: Laboratory 2, dtype: float64
```

In [23]:
```
lab_3 = lab_details['Laboratory 3'] #Taking Laboratory 3 in a Series.
lab_3.head(5)
```

Out[23]:
```
0    176.70
1    198.45
2    201.23
3    199.61
4    204.63
Name: Laboratory 3, dtype: float64
```

In [24]:
```
lab_4 = lab_details['Laboratory 4'] #Taking Laboratory 4 in a Series.
lab_4.head(5)
```

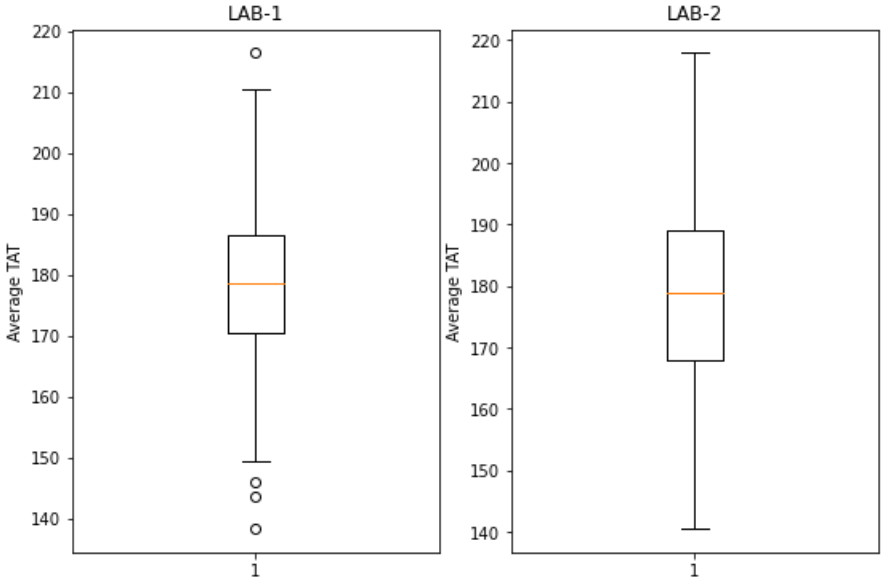
Out[24]:
```
0    166.13
1    160.79
2    185.18
3    176.42
4    152.60
Name: Laboratory 4, dtype: float64
```

In [25]:
```
lab_details[lab_details.duplicated()].shape #Checking duplicate values.
```

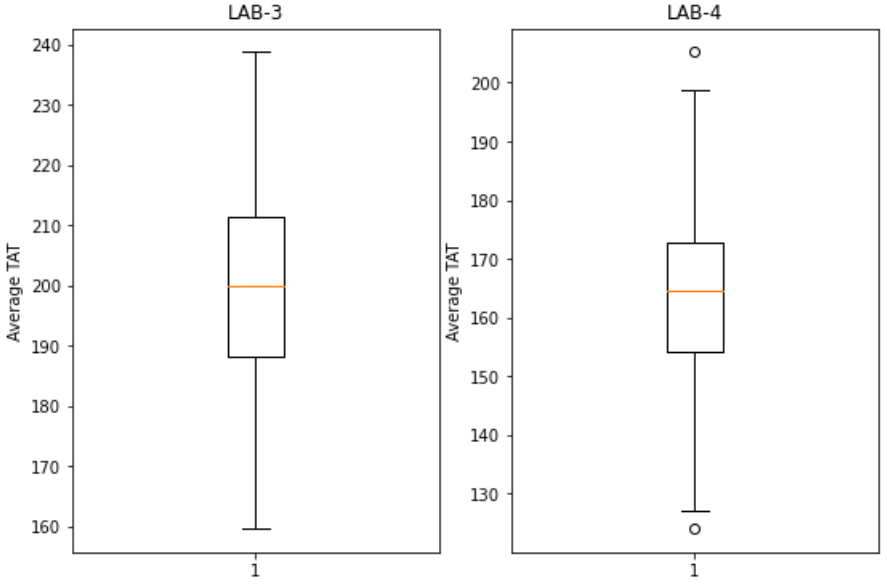
Out[25]:
```
(0, 4)
```

Step 3. Plotting Data in Different way:

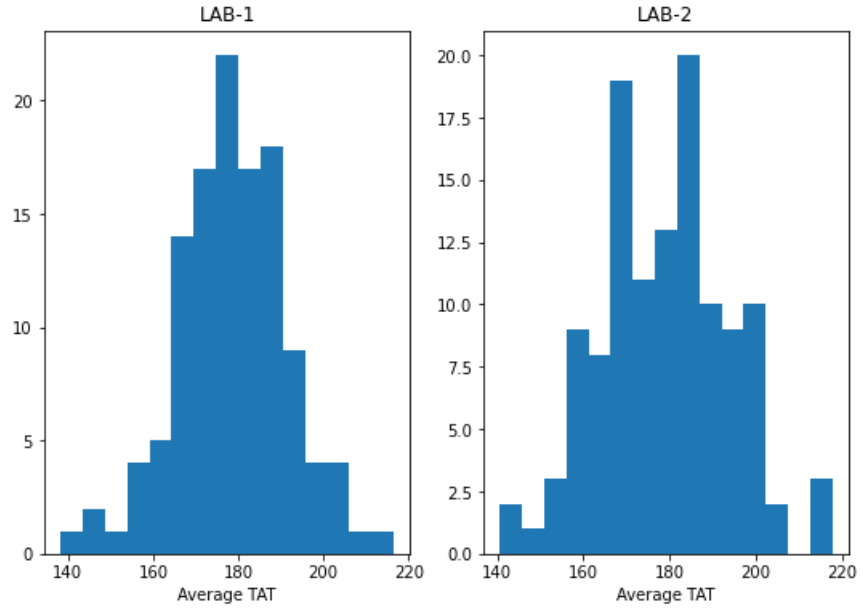
In [26]:
```
plt.subplots(figsize = (9,6))                              # boxplot for LAB-1 and LAB-2
plt.subplot(121)
plt.boxplot(lab_details['Laboratory 1'])
plt.title('LAB-1')
plt.ylabel('Average TAT')

plt.subplot(122)
plt.boxplot(lab_details['Laboratory 2'])
plt.title('LAB-2')
plt.ylabel('Average TAT')

plt.show()
```



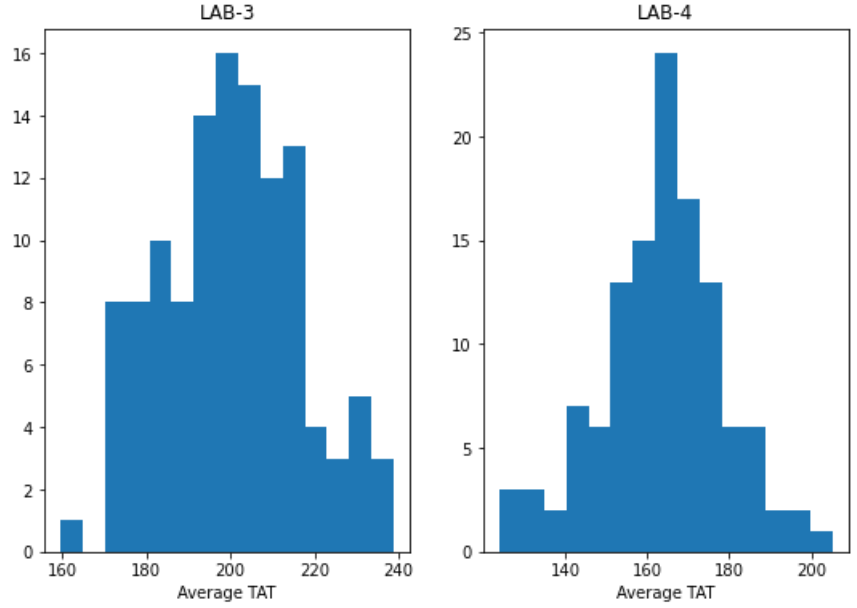
In [27]:
```
plt.subplots(figsize = (9,6))                          # boxplot for LAB-3 and LAB-4
plt.subplot(121)
plt.boxplot(lab_details['Laboratory 3'])
plt.title('LAB-3')
plt.ylabel('Average TAT')

plt.subplot(122)
plt.boxplot(lab_details['Laboratory 4'])
plt.title('LAB-4')
plt.ylabel('Average TAT')

plt.show()
```



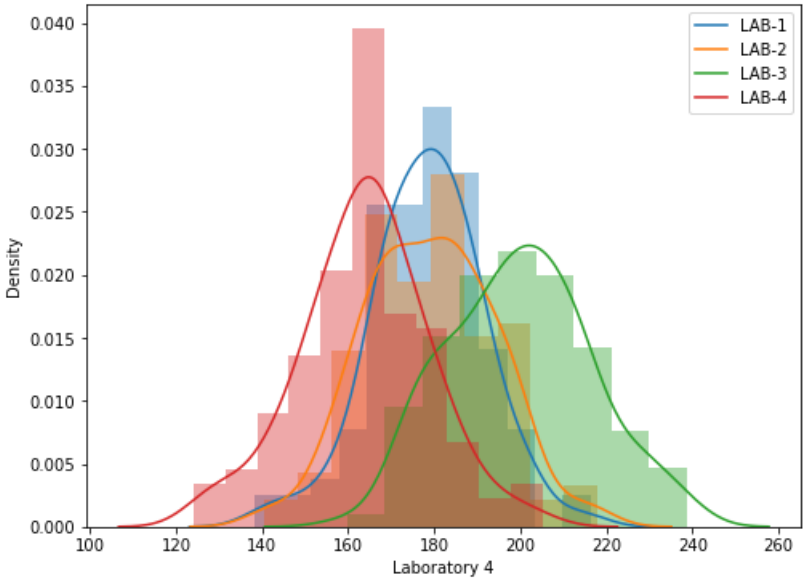
In [28]:
```
plt.subplots(figsize = (9,6))                          # subplot for LAB-1 and LAB-2
plt.subplot(121)
plt.hist(lab_details['Laboratory 1'], bins = 15)
plt.title('LAB-1')
plt.xlabel('Average TAT')

plt.subplot(122)
plt.hist(lab_details['Laboratory 2'], bins = 15)
plt.title('LAB-2')
plt.xlabel('Average TAT')

plt.show()
```



In [29]:
```
plt.subplots(figsize = (9,6))                          # subplot for LAB-3 and LAB-4
plt.subplot(121)
plt.hist(lab_details['Laboratory 3'], bins = 15)
plt.title('LAB-3')
plt.xlabel('Average TAT')

plt.subplot(122)
plt.hist(lab_details['Laboratory 4'], bins = 15)
plt.title('LAB-4')
plt.xlabel('Average TAT')

plt.show()
```



In [30]:
```
plt.figure(figsize = (8,6))                                    # distplot for LAB-1, 2, 3, 4.
labels = ['LAB-1', 'LAB-2', 'LAB-3', 'LAB-4']
sns.distplot(lab_details['Laboratory 1'], kde = True)
sns.distplot(lab_details['Laboratory 2'],hist = True)
sns.distplot(lab_details['Laboratory 3'],hist = True)
sns.distplot(lab_details['Laboratory 4'],hist = True)
plt.legend(labels)
plt.show()
```



Step 4. Calculating p-value:

Here we have applied ANOVA Test using t-statistics:

The one-way ANOVA tests the null hypothesis that two or more groups have the same population mean. The test is applied to samples from two or more groups, possibly with differing sizes.

In [31]:
```
test_statistic , p_value = stats.f_oneway(lab_1,lab_2,lab_3,lab_4) # Here's one-way ANOVA performed.
print('p_value is =',p_value)
```

```
p_value is = 2.1156708949992414e-57
```

Step 5. Hypothesis Testing and Interpretation of p-value:

Null Hypothesis (H0): There is significant difference in average TAT amongest the different labs.

Alternative Hypothesis (Ha): There is No difference in average TAT amongest the different labs.

In [34]:
```
# Level of significance = 5% ie, At 5% level of significance, do we reject or not reject?
# alpha = 0.05

if p_value<=0.05:
    print('We reject the Null Hypothesis and we can claim that there is significant difference in average TAT amongest the different labs')
else:
    print('We do not reject the Null Hypothesis and we can claim that there is No difference in average TAT amongest the different labs')
```

```
We reject the Null Hypothesis and we can claim that there is significant difference in average TAT amongest the different labs
```

Hence, We failed to reject the Null Hypothesis because of lack of evidence, there is no significant difference in average TAT amongest the different labs.

==========

Q. No. 3

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

Check p-value p-Value; If p-Value < alpha, we reject Null Hypothesis.

Ho = All proportions are equal.

Ha = Not all proportions are equal.

Answer:

We have to find out below;

Null Hypothesis (H0): There is no association between the gender based buyer rations across regions.

Alternative Hypothesis (Ha): There is a significant association between the gender based buyer rations across regions.

Step 1. Import Data or Dataset:

In [35]:
```
buyer_details = pd.read_csv('BuyerRatio.csv')
buyer_details.head(5)
```

Out[35]:

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

In [67]:
```
buyers_table = [[50,142,131,70], #Creating Contingency table
                [435,1523,1356,750]]
print(buyers_table)
```

```
[[50, 142, 131, 70], [435, 1523, 1356, 750]]
```

Step 2. Data understanding:

In [38]:
```
buyer_details.describe()
```

Out[38]:

	East	West	North	South
count	2.000000	2.000000	2.000000	2.000000
mean	242.500000	832.500000	743.500000	410.000000
std	272.236111	976.514465	866.205807	480.832611
min	50.000000	142.000000	131.000000	70.000000
25%	146.250000	487.250000	437.250000	240.000000
50%	242.500000	832.500000	743.500000	410.000000
75%	338.750000	1177.750000	1049.750000	580.000000
max	435.000000	1523.000000	1356.000000	750.000000

In [39]:
```
buyer_details.dtypes
```

Out[39]:
```
Observed Values    object
East                int64
West                int64
North               int64
South               int64
dtype: object
```

Step 3. Conduct a Test of Independence using Chi-Square test with Contingency table

In [46]:
```
stats.chi2_contingency(buyers_table)
```

Out[46]:
```
(1.595945538661058,
 0.6603094907091882,
 3,
 array([[  42.76531299,  146.81287862,  131.11756787,   72.30424052],
        [ 442.23468701, 1518.18712138, 1355.88243213,  747.69575948]]))
```

In [47]:
```
obs_value = np.array([50, 142, 131, 70, 435, 1523, 1356, 750])
exp_value = np.array([42.76531299,146.81287862,131.11756787,72.30424052,442.23468701,1518.18712138,1355.88243213,747.69575948])
```

Step 4. Checking Hypothesis statement (One way Chi-Square test):

In [57]:
```
statistics, p_value = stats.chisquare(obs_value, exp_value, ddof = 3)
print("Statistics = ",statistics,"\n",'P_Value = ', p_value)
```

```
Statistics =  1.5959455390914483
 P_Value =  0.8095206646905712
```

Step 5. Interpreting p-value

Compare p_value with alpha (Significane Level)

In [59]:
```
alpha = 0.05
print('Significnace=%.3f, p=%.3f' % (alpha, p_value))
if p_value <= alpha:
    print('We reject the Null Hypothesis and we can claim that there is a significant association between the gender based buyer rations across
else:
    print('We do not reject the Null Hypothesis and we can claim that there is no significant association between the gender based buyer rations
```

```
Significnace=0.050, p=0.810
We do not reject the Null Hypothesis and we can claim that there is no significant association between the gender based buyer rations across r
egions
```

Hence, We failed to reject the Null Hypothesis because of lack of evidence, there is no significant association between the gender based buyer rations across regions.

==========

Q. No. 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.

Answer:

We have to find out below;

Null Hypothesis (H0): There is no defective % variation amongest 4 centres around the globe.

Alternative Hypothesis (Ha): There is a significant defective % variation amongest 4 centres around the globe.

Step 1. Import Data or Dataset:

In [61]:
```
cu_form_details = pd.read_csv('Costomer+OrderForm.csv')
cu_form_details.head(5)
```

Out[61]:

	Phillippines	Indonesia	Malta	India
0	Error Free	Error Free	Defective	Error Free
1	Error Free	Error Free	Error Free	Defective
2	Error Free	Defective	Defective	Error Free
3	Error Free	Error Free	Error Free	Error Free
4	Error Free	Error Free	Defective	Error Free

In [64]:
```
cu_form_details.describe()
```

Out[64]:

	Phillippines	Indonesia	Malta	India
count	300	300	300	300
unique	2	2	2	2
top	Error Free	Error Free	Error Free	Error Free
freq	271	267	269	280

In [65]:
```
cu_form_details.isna().sum()
```

Out[65]:
```
Phillippines    0
Indonesia       0
Malta           0
India           0
dtype: int64
```

In [66]:
```
cu_form_details.dtypes
```

Out[66]:
```
Phillippines    object
Indonesia       object
Malta           object
India           object
dtype: object
```

Checking the value counts in dataset

In [76]:
```
print(cu_form_details['Phillippines'].value_counts(),'\n',cu_form_details['Indonesia'].value_counts(),'\n',cu_form_details['Malta'].value_counts
```

```
Error Free    271
Defective      29
Name: Phillippines, dtype: int64
 Error Free    267
Defective      33
Name: Indonesia, dtype: int64
 Error Free    269
Defective      31
Name: Malta, dtype: int64
 Error Free    280
Defective      20
Name: India, dtype: int64
```

In [68]:
```
contingency_table = [[271,267,269,280], #Creating Contingency table
                     [29,33,31,20]]
print(contingency_table)
```

```
[[271, 267, 269, 280], [29, 33, 31, 20]]
```

In [77]:
```
stat, p, df, exp = stats.chi2_contingency(contingency_table) #Calculating Expected Values for Observed data
print("Statistics = ",stat,"\n",'P_Value = ', p,'\n', 'degree of freedom =', df,'\n', 'Expected Values = ', exp)
```

```
Statistics =  3.858960685820355
 P_Value =  0.2771020991233135
 degree of freedom = 3
 Expected Values =  [[271.75 271.75 271.75 271.75]
 [ 28.25  28.25  28.25  28.25]]
```

In [78]:
```
observed = np.array([271, 267, 269, 280, 29, 33, 31, 20]) #Defining Expected values and observed values
expected = np.array([271.75, 271.75, 271.75, 271.75, 28.25, 28.25, 28.25, 28.25])
```

Comparing with Hypothesis using t-statictic:

In [79]:
```
test_statistic , p_value = stats.chisquare(observed, expected, ddof = df)
print("Test Statistic = ",test_statistic,'\n', 'p_value =',p_value)
```

```
Test Statistic =  3.858960685820355
 p_value = 0.4254298144535761
```

In [80]:
```
#Compare p_value with 'Alpha'(Significane Level)
```

interpreting p-value

In [82]:
```
alpha = 0.05
print('Significnace=%.3f, p=%.3f' % (alpha, p_value))
if p_value <= alpha:
    print('We reject the Null Hypothesis & we can claim that there is a significant defective % variation amongest 4 centres around the globe')
else:
    print('We do not reject the Null Hypothesis & we can claim that there is no defective % variation amongest 4 centres around the globe')
```

```
Significnace=0.050, p=0.425
We do not reject the Null Hypothesis & we can claim that there is no defective % variation amongest 4 centres around the globe
```

Hence, We failed to reject the Null Hypothesis because of lack of evidence, there is no defective % variation amongest the 04 centres around the globe.

==========THE END==========