Business Problem:Is the male-female buyer rations are similar across regions

Importing Libraries

In [1]:

```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
```

Importimg Data

In [2]:

```
buyerratio_data = pd.read_csv('BuyerRatio.csv')
buyerratio_data
```

Out[2]:

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

In [10]:

```
buyerratio_data.shape
```

Out[10]:

```
(2, 5)
```

In [4]:

```
buyerratio_data.isna().sum()
```

Out[4]:

```
Observed Values    0
East               0
West               0
North              0
South              0
dtype: int64
```

In [5]:

```
buyerratio_data.dtypes
```

Out[5]:

```
Observed Values    object
East                int64
West                int64
North               int64
South               int64
dtype: object
```

In [28]:

```
buyerratio_data.describe()
```

Out[28]:

	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 [31]:

```
# Since there are more than 2 variable we will perform Chi-square test
chi2_score,p_val,dof,expected_table=stats.chi2_contingency([buyerratio_data['East'],buyerra
print('chi - Squared value : ',chi2_score)
print('P - value           : ', p_val)
print('Degree of Freedom   : ', dof)
print('Expected Table      :\n',expected_table )
```

```
chi - Squared value :  1.5959455386610577
P - value           :  0.6603094907091882
Degree of Freedom   :  3
Expected Table      :
 [[  42.76531299  442.23468701]
 [ 146.81287862 1518.18712138]
 [ 131.11756787 1355.88243213]
 [  72.30424052  747.69575948]]
```

In [32]:

```
if p_val<0.05:
    print('We Reject the Null Hypothesis')
else:
    print('We Accept the Null Hypothesis')
```

```
We Accept the Null Hypothesis
```

The Male - Female buyer rations are Similar across regions

Costomer+OrderFrom - Problem Statement : TeleCall Uses 4Centers around the globe to Process customer order froms. They audit a certain % of the customer order froms. Any error in order from renders it defective and has to be reworked before processing. The manager Wants to check whether the defective % nvaries by centre. Please analyze the data at 5% Significance level and help the manager draw appropriate inferences.

Problem : Does the defective % varies significantly by centre ?

Importing data

In [33]:

```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
```

In [35]:

```
customer_data = pd.read_csv('Costomer+OrderForm.csv')
customer_data
```

Out[35]:

	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
...	...	...	...	...
295	Error Free	Error Free	Error Free	Error Free
296	Error Free	Error Free	Error Free	Error Free
297	Error Free	Error Free	Defective	Error Free
298	Error Free	Error Free	Error Free	Error Free
299	Error Free	Defective	Defective	Error Free

300 rows × 4 columns

Initial Analysis

In [36]:

```
customer_data.shape
```

Out[36]:

```
(300, 4)
```

In [37]:

```
customer_data.head()
```

Out[37]:

	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 [38]:

```
customer_data.dtypes
```

Out[38]:

```
Phillippines    object
Indonesia       object
Malta           object
India           object
dtype: object
```

In [39]:

```
customer_data.isna().sum()
```

Out[39]:

```
Phillippines    0
Indonesia       0
Malta           0
India           0
dtype: int64
```

In [40]:

```
customer_data.describe()
```

Out[40]:

	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 [41]:

```
stats.chi2_contingency([customer_data['Phillippines'].value_counts(),customer_data['Malta']
```

Out[41]:

```
(2.826219512195122,
 0.24338523637117,
 2,
 array([[273.33333333,  26.66666667],
        [273.33333333,  26.66666667],
        [273.33333333,  26.66666667]]))
```

In [43]:

```python
if p_val<0.05:
    print('we reject the null hypothesis')
else:
    print('we accept the null hypothesis')
```

```
we accept the null hypothesis
```

Cutlets - Problem Statement : 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 Assumption and tests you carried out to check validity of the assumptions

Problem : is threre significant differance in the diamter of the cutlet?

Importing data

In [44]:

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

In [47]:

```
order_data = pd.read_csv('Cutlets.csv')
order_data
```

Out[47]:

	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
5	7.3871	6.8110
6	6.8755	7.2212
7	7.0621	6.6606
8	6.6840	7.2402
9	6.8236	7.0503
10	7.3930	6.8810
11	7.5169	7.4059
12	6.9246	6.7652
13	6.9256	6.0380
14	6.5797	7.1581
15	6.8394	7.0240
16	6.5970	6.6672
17	7.2705	7.4314
18	7.2828	7.3070
19	7.3495	6.7478
20	6.9438	6.8889
21	7.1560	7.4220
22	6.5341	6.5217
23	7.2854	7.1688
24	6.9952	6.7594
25	6.8568	6.9399
26	7.2163	7.0133
27	6.6801	6.9182
28	6.9431	6.3346
29	7.0852	7.5459
30	6.7794	7.0992
31	7.2783	7.1180
32	7.1561	6.6965
33	7.3943	6.5780

	Unit A	Unit B
34	6.9405	7.3875

In [48]:

```
order_data.shape
```

Out[48]:

```
(35, 2)
```

In [49]:

```
order_data.head()
```

Out[49]:

	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

In [50]:

```
order_data.isna().sum()
```

Out[50]:

```
Unit A    0
Unit B    0
dtype: int64
```

In [51]:

```
order_data.describe()
```

Out[51]:

	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 [52]:

```
order_data.dtypes
```

Out[52]:

```
Unit A    float64
Unit B    float64
dtype: object
```

In [53]:

```
stats.shapiro(order_data['Unit A'])
```

Out[53]:

```
ShapiroResult(statistic=0.9649458527565002, pvalue=0.3199819028377533)
```

P value for unit A==0.32>α

In [54]:

```
stats.shapiro(order_data['Unit B'])
```

Out[54]:

```
ShapiroResult(statistic=0.9727300405502319, pvalue=0.5224985480308533)
```

P value for unit B == 0.522 > α

HO is Accepted. That is both Y1 and Y2 are normal, Thus we can perform Paired T Test

In [55]:

```
stats.ttest_rel(order_data['Unit A'],order_data['Unit B'])
```

Out[55]:

```
Ttest_relResult(statistic=0.7536787225614314, pvalue=0.4562300768038412)
```

Null Hypothesis is Accepted. So Mean of Y1 and Y2 are equal

LabTAT : Problem Statement : A hospital wants to determine whether there is any difference in the avrage 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

Problem : is there a significant difference in the average Turn Around Time between laboratories?

In [56]:

```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
```

In [58]:

```
lab_data = pd.read_csv('LabTAT.csv')
lab_data
```

Out[58]:

	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
...	...	...	...	...
115	178.49	170.66	193.80	172.68
116	176.08	183.98	215.25	177.64
117	202.48	174.54	203.99	170.27
118	182.40	197.18	194.52	150.87
119	182.09	215.17	221.49	162.21

120 rows × 4 columns

In [60]:

```
lab_data.head()
```

Out[60]:

	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

In [61]:

```
lab_data.dtypes
```

Out[61]:

```
Laboratory 1    float64
Laboratory 2    float64
Laboratory 3    float64
Laboratory 4    float64
dtype: object
```

In [62]:

```
lab_data.isna().sum()
```

Out[62]:

```
Laboratory 1    0
Laboratory 2    0
Laboratory 3    0
Laboratory 4    0
dtype: int64
```

In [63]:

```
lab_data.describe()
```

Out[63]:

	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

H0 = Y1, Y2, and Y4 are normal H1 = Y1, Y2, Y3 and Y4 are not normal

In [65]:

```
stats.shapiro(lab_data['Laboratory 1'])
```

Out[65]:

```
ShapiroResult(statistic=0.9901824593544006, pvalue=0.5506953597068787)
```

P value for laboratory 1 == 0.5506 > α

In [66]:

```
stats.shapiro(lab_data['Laboratory 2'])
```

Out[66]:

```
ShapiroResult(statistic=0.9936322569847107, pvalue=0.8637524843215942)
```

P value for laboratory 2 == 0.8637 > α

In [67]:

```
stats.shapiro(lab_data['Laboratory 3'])
```

Out[67]:

```
ShapiroResult(statistic=0.9886345267295837, pvalue=0.4205053448677063)
```

P value for laboratory 3 == 0.4205 > α

In [68]:

```
stats.shapiro(lab_data['Laboratory 4'])
```

Out[68]:

```
ShapiroResult(statistic=0.9913753271102905, pvalue=0.6618951559066772)
```

P value of Variance test is == 0.6618 > α

In [70]:

```
stats.stats.f_oneway(lab_data['Laboratory 1'],lab_data['Laboratory 2'],lab_data['Laboratory
```

Out[70]:

```
F_onewayResult(statistic=118.70421654401437, pvalue=2.1156708949992414e-57)
```

P value of the One way Anova test is == 2.1156e-57 < α

Thus H1 is Accepted.

Main TAT for 4 laboratories not equal (There is a Significance)

In [ ]:

```

```