Hypothesis Testing -Assignments

In [39]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
from scipy.stats import chi2_contingency
from scipy.stats import chi2
```

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

In [40]:
```
# Load the dataset
data=pd.read_csv('Cutlets.csv')
data.head()
```

Out[40]:

	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 [41]:
```
unitA=pd.Series(data.iloc[:,0]) #calling Series or DataFrame
unitA
```

Out[41]:
```
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
```

In [42]:

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

Out[42]:

```
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
```

In [43]:

```
#Assuming Null hyposthesis as Ho: μ1 = μ2 (There is no difference in diameters of cu

#Thus Alternate hypothesis as Ha: μ1 ≠ μ2 (There is significant difference in diamet
```

In [44]:

```
# 2-sample 2-tail ttest:   stats.ttest_ind(array1,array2)     # ind -> independent s
```

```
p_value=stats.ttest_ind(unitA,unitB)
p_value
```

Out[44]:
```
Ttest_indResult(statistic=0.7228688704678063, pvalue=0.4722394724599501)
```

In [45]:
```
p_value[1]     # 2-tail probability
```

Out[45]:
```
0.4722394724599501
```

In [79]:
```
if p_value[1] <= 0.05:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
Dependent (reject H0)
```

==========

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

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**

In [47]:
```
# load the dataset
data=pd.read_csv('LabTAT.csv')
data.head()
```

Out[47]:

	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 [48]:
```
#Anova ftest statistics: Analysis of varaince between more than 2 samples or columns

#Thus Alternate Hypothesis Ha as It has Variance: Atleast one sample TAT population
```

In [49]:
```
# 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
```

Out[49]:
```
F_onewayResult(statistic=118.70421654401437, pvalue=2.1156708949992414e-57)
```

In [50]:
```
p_value[1]  # comparing with α = 0.05
```

Out[50]:
```
2.1156708949992414e-57
```

In [51]:
```
# failed to reject H0
```

==========

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

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

In [52]:
```
df= pd.read_csv('BuyerRatio.csv')
df.head()
```

Out[52]:

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

In [53]:
```
df_table=df.iloc[:,1:6]
df_table
```

Out[53]:

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

In [54]:
```
val=stats.chi2_contingency(df_table)
val
```

Out[54]:
```
(1.595945538661058,
 0.6603094907091882,
 3,
 array([[  42.76531299,  146.81287862,  131.11756787,   72.30424052],
        [ 442.23468701, 1518.18712138, 1355.88243213,  747.69575948]]))
```

In [55]:
```
type(val)
```

Out[55]:
```
tuple
```

In [56]:
```
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
```

In [57]:
```
Expected_value=val[3]
Expected_value
```

Out[57]:
```
array([[  42.76531299,  146.81287862,  131.11756787,   72.30424052],
       [ 442.23468701, 1518.18712138, 1355.88243213,  747.69575948]])
```

In [58]:
```
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
```

Out[58]:
```
1.5152956451130446
```

In [59]:
```
critical_value=chi2.ppf(0.95,3)
critical_value
```

Out[59]:
```
7.814727903251179
```

In [60]:
```
if chi_square_statestic >= critical_value:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
Independent (fail to reject H0)
```

In [61]:
```
pvalue=1-chi2.cdf(chi_square_statestic,3)
pvalue
```

Out[61]:
```
0.6787446296467897
```

In [62]:
```
if pvalue <= 0.05:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
Independent (fail to reject H0)
```

In [ ]:
```

```

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

In [63]:
```
#Assuming 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)
```

In [67]:
```
custom = pd.read_csv('Costomer+OrderForm.csv')
custom
```

Out[67]:

	Phillippines	Indonesia	Malta	India
0	Error Free	Error Free	Defective	Error Free
1	Error Free	Error Free	Error Free	Defective

	Phillippines	Indonesia	Malta	India
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

In [68]:
```
print(custom['Phillippines'].value_counts(),custom['Indonesia'].value_counts(),custo
```

```
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 [70]:
```
observed=([[271,267,269,280],[29,33,31,20]])
observed
```

Out[70]:
```
[[271, 267, 269, 280], [29, 33, 31, 20]]
```

In [72]:
```
stat, p, dof, expected = chi2_contingency([[271,267,269,280],[29,33,31,20]])
stat
```

Out[72]:
```
3.858960685820355
```

In [73]:
```
p
```

Out[73]:
```
0.2771020991233135
```

In [74]:
```
print('dof=%d' % dof)
print(expected)
```

```
dof=3
[[271.75 271.75 271.75 271.75]
 [ 28.25  28.25  28.25  28.25]]
```

In [75]:
```
alpha = 0.05
prob=1-alpha
critical = chi2.ppf(prob, dof)
print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))
```

```
if abs(stat) >= critical:
        print('Dependent (reject H0),variables are related')
else:
        print('Independent (fail to reject H0), variables are not related')
```

```
probability=0.950, critical=7.815, stat=3.859
Independent (fail to reject H0), variables are not related
```

In [76]:

```
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
significance=0.050, p=0.277
Independent (fail to reject H0)
```

In [77]:

```
# Inference: As (p_value = 0.2771) > (α = 0.05); Accept Null Hypthesis i.e.
#Independence of categorical variables Thus, customer order forms defective % does n
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: