Assignment No.3 : Hypothesis Testing

Name : Samrat Pawar

Branch : Andheri

Hypothesis Testing Exercise

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 :

Step1 : To check whether the diameter of two units are similar or not

Step2 : y and x. So here is y is continuous and x is discrete

Step3 : Here we will use 2-sample t test

Step4 : Find normality of this data

jupyter ExcelR Assignment no.3 (Hypothesis Testing) Last Checkpoint: Last Wednesday at 1:25 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3

Run Code

Q1. 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.

In [2]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
```

In [3]:
```
df = pd.read_csv ("/Users/Computer/Downloads/Cutlets.csv")
```

In [5]:
```
df
```

Out[5]:

	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

In [8]:
```
#null hypothesis : Ho : μ1 = μ2 , there is no difference between diameter of cutlets between two units
#Alternate hypothesis : Ha : μ1 ≠ μ2 , there is significant difference between two units
#as we see there are two means of same group so it is 2 sample 2 tail test
```

In [10]:
```
unitA=pd.Series(df.iloc[:,0])
unitA
```

Out[10]:
```
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
```

In [11]:
```
unitB=pd.Series(df.iloc[:,1])
unitB
```

Out[11]:
```
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
```

In [12]:
```
#p_value : probability value
#ind = independent samples
#unit A = array 1 , unit B = array 2


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

Out[12]:
```
Ttest_indResult(statistic=0.7228688704678063, pvalue=0.4722394724599501)
```

In [14]:
```
p_value[1]
```

Out[14]:
```
0.4722394724599501
```

In [15]:
```
#we have significance level at 5% , α = 0.05 , c = 1-α = 0.95
#Inferenec : By the conclusion , p_value =0.472 > α = 0.05
#Accept null hypoyhesis : Ho : μ1 = μ2 : There is no difference betwn diameter of Cutlets
```

Hypothesis Testing Exercise

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 :

Step1 : In this problem we check , whether there is any difference in average TAT.

Step2 : y and x . So here is 4 labs are input TAT(Turn around time) is outputx is more than 2 discrete and y is continous.

Step3 : Here we will use ANOVA-One wayFind difference between 4 labrotaries with respect to timeX ->4 labratory y -> TAT(Turn around time) .

Step4 : Find normality of this data.

jupyter ExcelR Assignment no.3 (Hypothesis Testing) Last Checkpoint: Last Wednesday at 1:25 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

Q2. 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.

In [16]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
```

In [12]:
```
data = pd.read_csv ("/Users/Computer/Downloads/LabTaT.csv")
```

In [18]:
```
data
```

Out[18]:

	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

In [19]:
```
#As we see there are more than two samplels in the data so here we can apply "Anova Ftest"
#null hypoythesis : Ho : μ1 = μ2 = μ3 = μ4 : there is no differnce in tat reports
#Alternate hypothesis : Ha : atleast one populatuon mean is different
```

In [13]:
```
# Anova ftest statistics:
#f_oneway : accepts multidimensional input arrays.
#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[13]:
```
F_onewayResult(statistic=118.70421654401437, pvalue=2.1156708949992414e-57)
```

In [14]:
```
p_value[1]
```

Out[14]:
```
2.1156708949992414e-57
```

In [15]:
```
#we have significance level of 5% , α = 0.05 , c = 1- 0.05 = 0.95
#Inference : As we see p_value=0 < α=0.05 ,
#so reject null hypothesis : atleast one sample tat population mean  is different.
```

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

H_0

- All proportions are equal

H_a

- Not all Proportions are equal

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

Buyer Ratio.mtw

Ans :

Step1 : To find buyer ratios are similar across region or not

Step2 : y and x x is more than 2 discrete and y is discrete

Step3 : Here we will use Chi-square test

Step4 : Find normality of this data

Q.3.

In [8]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
from scipy.stats import chi2_contingency
```

In [4]:
```
data = pd.read_csv ("C:/Users/Computer/Downloads/BuyerRatio.csv")
```

In [29]:
```
data
```

Out[29]:

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

In [6]:
```
# Make dimensional array
obs=np.array([[50,142,131,70],[435,1523,1356,750]])
obs
```

Out[6]:
```
array([[  50,  142,  131,   70],
       [ 435, 1523, 1356,  750]])
```

```
In [9]: # Chi2 contengency independence test
        chi2_contingency(obs) # o/p is (Chi2 stats value, p_value, df, expected obsvations)

Out[9]: (1.595945538661058,
         0.6603094907091882,
         3,
         array([[  42.76531299,  146.81287862,  131.11756787,   72.30424052],
                [ 442.23468701, 1518.18712138, 1355.88243213,  747.69575948]]))

In [ ]: # Compare p_value with α = 0.05
        #Inference: As (p-value = 0.6603) > (α = 0.05);
        #Accept the Null Hypothesis i.e. Independence of categorical variables
        #Thus, male-female buyer rations are similar across regions and are not related
```

Hypothesis Testing Exercise

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 [36]:
```python
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
from scipy.stats import chi2_contingency
```

In [16]:
```python
data = pd.read_csv("C:/Users/Computer/Downloads/Costomer+OrderForm.csv")
```

In [17]:
```python
data
```

Out[17]:

	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

In [39]:
```
#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)
```

In [42]:
```
data.Phillippines.value_counts()
```

Out[42]:
```
Error Free    271
Defective      29
Name: Phillippines, dtype: int64
```

In [18]:
```
data.Indonesia.value_counts()
```

Out[18]:
```
Error Free    267
Defective      33
Name: Indonesia, dtype: int64
```

In [44]:
```
data.Malta.value_counts()
```

Out[44]:
```
Error Free    269
Defective      31
Name: Malta, dtype: int64
```

In [46]:
```
data.India.value_counts()
```

Out[46]:
```
Error Free    280
Defective      20
Name: India, dtype: int64
```

```
In [45]: # Make a contingency table
         obs=np.array([[271,267,269,280],[29,33,31,20]])
         obs

Out[45]: array([[271, 267, 269, 280],
                [ 29,  33,  31,  20]])

In [47]: #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)

In [48]: # Chi2 contengency independence test
         chi2_contingency(obs)
         # o/p is (Chi2 stats value, p_value, df, expected obsvations)

Out[48]: (3.858960685820355,
          0.2771020991233135,
          3,
          array([[271.75, 271.75, 271.75, 271.75],
                 [ 28.25,  28.25,  28.25,  28.25]]))

In [ ]:

In [49]: # Compare p_value with α = 0.05
         #Inference: As (p_value = 0.2771) > (α = 0.05);
         #Accept Null Hypthesis i.e. Independence of categorical variables Thus, customer order forms defective % does not varies by centr
```