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**

To determine whether there is any significant difference in the diameter of the cutlet between two units

In [24]:
```
import pandas as pd
import numpy as np
import matplotlib as plt
import scipy
from scipy import stats
from statsmodels.stats.proportion import proportions_ztest
import statsmodels.api as sm
from numpy.random import seed
from numpy.random import randn
from scipy.stats import shapiro
import statistics
from statistics import variance
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.formula.api import ols
import seaborn as sns
```
executed in 32ms, finished 15:56:18 2021-01-22

Ho=There is no significance difference in the diameter of the cutlet between two units

H1= There is a significance difference in the diameter of the cutlet between two units

In [25]:
```
cutlet=pd.read_csv("Cutlets.csv")
```
executed in 16ms, finished 15:56:21 2021-01-22

In [26]:
```
cutlet
```
executed in 32ms, finished 15:56:23 2021-01-22

Out[26]:

	Unit A	Unit B
0	6.8090	6.7703
1	6.4376	7.5093
2	6.9157	6.7300

```
cutlet.isnull().sum()
```
executed in 16ms, finished 15:56:29 2021-01-22

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

```
cutlet.describe()
```
executed in 47ms, finished 15:56:33 2021-01-22

	Unit A	Unit B
count	35.000000	35.000000
mean	7.019091	6.964297
std	0.288408	0.343401
min	6.437600	6.038000

```
#normality test
stat,p=shapiro(cutlet)
```
executed in 16ms, finished 15:56:37 2021-01-22

```
print('statistics=%.3f.p=%.3f'%(stat,p))
```
executed in 31ms, finished 15:56:41 2021-01-22

```
statistics=0.976.p=0.204
```

As we got P=0.204, p>0.05 hence we can accept null hypothesis. Here Null Hypothesis is saying that the cutlet data is **Normal**

To find the Variance for the given sample data with the help of variance Tests

```
cut=pd.DataFrame(cutlet)
```
executed in 15ms, finished 15:56:43 2021-01-22

```
cutlet_a=cut.rename(columns={"Unit A":"sampleA","Unit B":"sampleB"})
cutlet_a
```
executed in 31ms, finished 15:56:47 2021-01-22

	sampleA	sampleB
0	6.8090	6.7703
1	6.4376	7.5093

```
: print("Variance of sample setA = %f" %(statistics.variance(cutlet_a.sampleA)))
```
executed in 16ms, finished 15:56:52 2021-01-22

```
Variance of sample setA = 0.083179
```

```
: print("Variance of sample setB = %f" %(statistics.variance(cutlet_a.sampleB)))
```
executed in 31ms, finished 15:56:54 2021-01-22

```
Variance of sample setB = 0.117924
```

Compute Test Statistic

Using Two Tailed t test we can compute that weather to accept the null hypothesis or not

```
|: X=cutlet_a.sampleA
   Y=cutlet_a.sampleB
```
executed in 15ms, finished 15:56:58 2021-01-22

```
|: #normality test
   #cut=stat.ttest_ind(cutlet_a.sampleA,cutlet_a.sampleB)
   #print(cut)
```
executed in 27ms, finished 22:52:11 2021-01-21

```
|: stats.f_oneway(X,Y)
```
executed in 15ms, finished 15:57:01 2021-01-22

```
|: F_onewayResult(statistic=0.5225394038913945, pvalue=0.4722394724599509)
```

Interpretation

As the obtained p value is more than alpha value(0.05),P>0.05 we can accept Null hypothesis. As we observed there is no significant difference in the diameter of the cutlet between 2 units.

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.

Ho=There is a difference in average TAT among the different laboratories

H1=There is no difference in average TAT among the different laboratories

```
: Data=pd.read_csv("LabTAT.csv")
```
executed in 31ms, finished 15:57:05 2021-01-22

```
: Data.info
```
executed in 31ms, finished 15:57:08 2021-01-22

```
: <bound method DataFrame.info of      Laboratory 1  Laboratory 2
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
```

```
Data.shape
```
executed in 15ms, finished 15:57:11 2021-01-22

```
(120, 4)
```

```
Data.isnull().sum()
```
executed in 31ms, finished 15:57:14 2021-01-22

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

```
Data.describe()
```
executed in 47ms, finished 15:57:18 2021-01-22

	Laboratory 1	Laboratory 2	Laboratory 3	Laboratory 4
count	120.000000	120.000000	120.000000	120.00000

```
stat,p = shapiro(Data)
print('statistics=%.3f,p=%.3f' %(stat,p))
```

executed in 15ms, finished 15:57:22 2021-01-22

```
statistics=0.995,p=0.118
```

From the above snippet of code, we see that the p-value is >0.05 i.e,0.118>0.05 for all density groups. Hence, we can conclude that they follow the Gaussian Distribution.

Homogeneity of Variance Assumption check

```
#Levene variance test,Method 2
stats.levene(Data["Laboratory 1"],Data["Laboratory 2"],Data["Laboratory 3"],Data["Laboratory 4"])
```

executed in 32ms, finished 15:57:25 2021-01-22

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

We see that p-value >0.05 for all density groups. Hence, we can conclude that groups have equal variances.

```
df = pd.melt(Data.reset_index(),id_vars=['index'],
             value_vars=['Laboratory 1','Laboratory 2','Laboratory 3','Laboratory 4'])
df
```

executed in 46ms, finished 15:57:31 2021-01-22

	index	variable	value
0	0	Laboratory 1	185.35

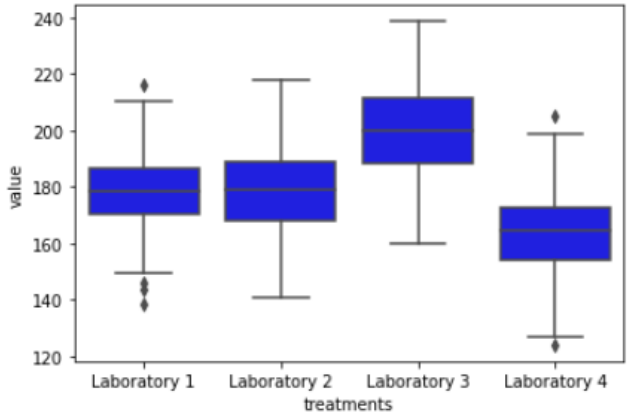
```
df.columns = ['index','treatments','value']
df.columns
```

executed in 31ms, finished 15:57:35 2021-01-22

```
Index(['index', 'treatments', 'value'], dtype='object')
```

Generate a boxplot to see the data distribution by treatments. Using boxplot, we can easily detect the differences between different treatments

```
: ax = sns.boxplot(x='treatments',y='value',data=df,color='blue')
executed in 703ms, finished 15:57:39 2021-01-22
```



One-Way ANOVA Test using stats models module

```
stats.f_oneway(Data["Laboratory 1"],Data["Laboratory 2"],Data["Laboratory 3"],Data["Laboratory 4"])
executed in 16ms, finished 15:57:43 2021-01-22
```

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

We see that p-value <0.05. Hence, we can reject the Null Hypothesis – there are no differences among different groups P value=2.1156708949992414e-57,p<0.05,P low , fail to accept null hypothesis , what alternate hypothesis says "Average turn around time of all 4 laboratories are not same"

When we conduct an ANOVA, we are attempting to determine if there is a statistically significant difference among the groups. So what if we find statistical significance?

If we find that there is a difference, we will then need to examine where the group differences lay. So, we'll use the Tukey HSD test to identify where the difference lies

```
Multi_Comp=sm.stats.multicomp.MultiComparison(df['value'],df['treatments'])
Multi_Comp_Results=Multi_Comp.tukeyhsd()
print(Multi_Comp_Results)
```

executed in 141ms, finished 15:57:46 2021-01-22

```
       Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================================
   group1       group2    meandiff p-adj  lower    upper   reject
---------------------------------------------------------------------
Laboratory 1 Laboratory 2   0.5413   0.9  -4.4468   5.5294  False
Laboratory 1 Laboratory 3  21.5517 0.001  16.5636  26.5398   True
Laboratory 1 Laboratory 4 -14.6788 0.001 -19.6669  -9.6907   True
Laboratory 2 Laboratory 3  21.0103 0.001  16.0222  25.9984   True
Laboratory 2 Laboratory 4 -15.2202 0.001 -20.2083 -10.2321   True
Laboratory 3 Laboratory 4 -36.2305 0.001 -41.2186 -31.2424   True
---------------------------------------------------------------------
```

Tuckey HSD test clearly says that there's a significant difference between Group1 – Group2

Interpretation: Hence statistically Proved that Average turn around time of all 4 laboratories are not same

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

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

executed in 31ms, finished 15:57:50 2021-01-22

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

```
c=br['Observed Values']=br['Observed Values'].map({'Males':0,'Females':1})
c
```

executed in 31ms, finished 15:57:58 2021-01-22

```
0    0
1    1
Name: Observed Values, dtype: int64
```

```
df=br.T
df
```

executed in 31ms, finished 15:58:02 2021-01-22

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

```
scipy.stats.chi2_contingency(df)
```

executed in 31ms, finished 15:58:06 2021-01-22

```
(1.6929696469183673,
 0.7919942975413565,
 4,
 array([[8.81561238e-02, 9.11843876e-01],
        [4.27557201e+01, 4.42244280e+02],
        [1.46779946e+02, 1.51822005e+03],
        [1.31088156e+02, 1.35591184e+03],
        [7.22880215e+01, 7.47711978e+02]]))
```

```
chi,pval,dof,exp = scipy.stats.chi2_contingency(br)
chi,pval,dof,exp
```

executed in 16ms, finished 15:58:09 2021-01-22

```
(1.6929696469183673,
 0.7919942975413565,
 4,
 array([[8.81561238e-02, 4.27557201e+01, 1.46779946e+02, 1.31088156e+02,
         7.22880215e+01],
        [9.11843876e-01, 4.42244280e+02, 1.51822005e+03, 1.35591184e+03,
         7.47711978e+02]]))
```

```
print('p-value is:', pval)
significance = 0.05
p = 1-significance
critical_value = scipy.stats.chi2.ppf(p,dof)
print('chi=%.6f, critical value= %.6f\n' %(chi, critical_value))
if chi > critical_value:
    print("""At %.2f level of significance,we reject the null hypothesis and accept H1.
    Not all proportions are equal.""" %(significance))
else:
    print("""At %.2f level of significance,we accept the null hypothesis. all proportions are equal"""
          %(significance))
```

executed in 31ms, finished 15:58:13 2021-01-22

```
p-value is: 0.7919942975413565
chi=1.692970, critical value= 9.487729

At 0.05 level of significance,we accept the null hypothesis. all proportions are equal
```

INTERPRETATION:-

Here from the above data if u observe P-value>0.05 we Accept Null Hypothesis Proportion Of Buyersratio Across all the Regions Should Be Same Statistically Proved

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

Ho: The defective % not varies by centre.

H1: The defective % varies by centre.

```
customer=pd.read_csv("Costomer+OrderForm.csv")
customer
```

executed in 31ms, finished 15:58:18 2021-01-22

	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

```
customer.isnull().sum()
```

executed in 16ms, finished 15:58:22 2021-01-22

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

```
customer['Phillippines']=customer['Phillippines'].map({'Error Free':0,'Defective':1})
customer['Indonesia']=customer['Indonesia'].map({'Error Free':0,'Defective':1})
customer['Malta']=customer['Malta'].map({'Error Free':0,'Defective':1})
customer['India']=customer['India'].map({'Error Free':0,'Defective':1})
```

executed in 31ms, finished 15:58:26 2021-01-22

```
customer_a=pd.DataFrame([customer.Phillippines,customer.Indonesia,customer.Malta,customer.India])
df_a=customer_a.T
df_a
```

executed in 63ms, finished 15:58:29 2021-01-22

	Phillippines	Indonesia	Malta	India
0	0	0	1	0
1	0	0	0	1

```
stat,p = shapiro(customer)
print('Statistics=%.3f,p=%.3f' %(stat,p))
```

executed in 15ms, finished 15:58:34 2021-01-22

```
Statistics=0.330,p=0.000
```

```
stats.f_oneway(customer['Phillippines'],customer['Indonesia'],customer['Malta'],customer['India'])
```

executed in 31ms, finished 15:58:37 2021-01-22

```
F_onewayResult(statistic=1.286168556089167, pvalue=0.2776780955705948)
```

```
cust_ord = pd.melt(customer.reset_index(),id_vars=['index'],
            value_vars=['Phillippines','Indonesia','Malta','India'])
cust_ord
```

executed in 47ms, finished 16:06:07 2021-01-22

	index	variable	value
0	0	Phillippines	0
1	1	Phillippines	0

```
cust_ord.columns = ['index','treatments','value']
cust_ord.columns
```

executed in 16ms, finished 16:06:11 2021-01-22

```
Index(['index', 'treatments', 'value'], dtype='object')
```

```
mc = sm.stats.multicomp.MultiComparison(cust_ord['value'],cust_ord['treatments'])
mc_result=mc.tukeyhsd()
print(mc_result)
```

executed in 109ms, finished 16:06:15 2021-01-22

```
    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================================
  group1     group2    meandiff p-adj   lower  upper  reject
---------------------------------------------------------------------
    India    Indonesia   0.0433 0.2658  -0.018 0.1047  False
    India        Malta   0.0367 0.4168 -0.0247  0.098  False
    India Phillippines     0.03 0.5792 -0.0314 0.0914  False
Indonesia        Malta  -0.0067    0.9  -0.068 0.0547  False
Indonesia Phillippines  -0.0133    0.9 -0.0747  0.048  False
    Malta Phillippines  -0.0067    0.9  -0.068 0.0547  False
---------------------------------------------------------------------
```

Interpretation:-

p>0.05 we accept null hypothesis, Acording to the data There is no defect varies by a center and the Tukeyhsd states that there is a significant different among the groups

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.

Ho: % of males versus females walking in to the store not differ based on day of the week

H1: % of males versus females walking in to the store differ based on day of the week

]:
```
fantloon = pd.read_csv("Faltoons.csv")
fantloon
```
executed in 31ms, finished 15:58:54 2021-01-22

]:

	Weekdays	Weekend
0	Male	Female
1	Female	Male
2	Female	Male

```
fantloon['Weekdays']=fantloon['Weekdays'].map({'Male':0,'Female':1})
fantloon['Weekend']=fantloon['Weekend'].map({'Male':0,'Female':1})
fantloon.head(2)
```
executed in 31ms, finished 15:58:58 2021-01-22

	Weekdays	Weekend
0	0	1
1	1	0

```
fant=pd.crosstab(fantloon.Weekdays,fantloon.Weekend)
fant
```
executed in 47ms, finished 15:59:01 2021-01-22

Weekend	0	1
Weekdays		
0	47	66
1	120	167

```
count= np.array([47,66])#how many men and women are females walking in the store
nobs= np.array([120,167])#total number of people coming
```
executed in 15ms, finished 15:59:04 2021-01-22

```
stat, pval_a=proportions_ztest(count,nobs,alternative='two-sided')
stat,pval_a
```
executed in 31ms, finished 15:59:07 2021-01-22

```
(-0.06059497248502743, 0.9516817775441105)
```

INTERPRETATION:- p = 0.951,p>0.05, P high Null fly, Acccept Null hypothesis, what null Hypothesis says is "equal Proportions"

CONCLUSION : sales manager commit to start fantaloons sales at weekends and weekdays both , why Because Both The days sales with respect to male Vs Female are walking in to the store not differ based on day of the week, statistically proved

NOTE:

Alternative The alternative hypothesis can be either two-sided or one of the one-sided tests smaller means that the alternative hypothesis is prop < value
larger means prop > value.

two. sided -> means checking for equal proportions of Adults and children under purchased p-value < 0.05 accept alternate hypothesis.