**Hypothesis Testing Exercise**

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**

ANS:

Null Hypothesis = µ1 = µ2

Alternate Hypothesis = µ1 ≠ µ2

Because it is two tailed test alpha/2 = 0.05

Here the Signficance is 5%

And the alpha is 0.05

CODES:

* from google.colab import files                                      uploaded = files.upload () [CODE TO IMPORT THE FILE]
* cutlets = pd.read\_csv("Cutlets.csv") [ CODE TO READ THE FILE]
* cutlets.describe() [ Discriptive stats]

|  | **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 |

* cutlets.isnull().sum()   [ Checking for null values]

Unit A 0 Unit B 0 dtype: int64

* cutlets[cutlets.duplicated()].shape  [checking duplicated values]

(0, 2)

* cutlets.info() [ checking the data type]

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 35 entries, 0 to 34

Data columns (total 2 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

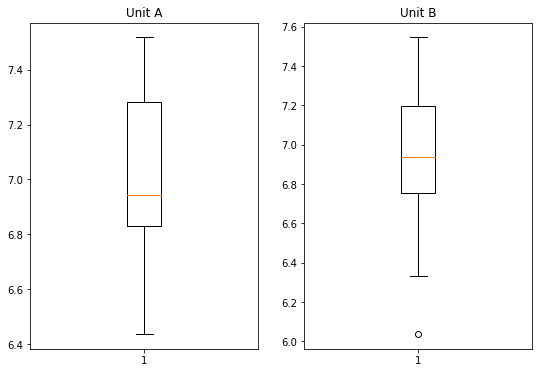
0 Unit A 35 non-null float64

1 Unit B 35 non-null float64

dtypes: float64(2)

memory usage: 688.0 bytes

* plt.subplots(figsize = (9,6))        [ Plotting the data]
* plt.subplot(121)
* plt.boxplot(cutlets['Unit A'])
* plt.title('Unit A')
* plt.subplot(122)
* plt.boxplot(cutlets['Unit B'])
* plt.title('Unit B')
* plt.show()



Compare Evidences with Hypothesis using t-statistics

statistic , p\_value **=** stats**.**ttest\_ind(cutlets['Unit A'],cutlets['Unit B'], alternative **=** 'two-sided')

print('p\_value=',p\_value)

p\_value= 0.4722394724599501

### Compare p\_value with 'α '(Significane Level)

### If p\_value is ≠ 'α ' we failed to reject Null Hypothesis because of lack of evidence

### If p\_value is = 'α ' we reject Null Hypothesis

interpreting p-value

alpha **=** 0.025

print('Significnace=%.3f, p=%.3f' **%** (alpha, p\_value))

**if** p\_value **<=** alpha:

print('We reject Null Hypothesis there is a significance difference between two Units A and B')

**else**:

print('We fail to reject Null hypothesis')

Significnace=0.025, p=0.472

We fail to reject Null hypothesis

Hence, we fail to reject the Null hypothesis because of lack of evidence, because there is no significance difference between them.

1. 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:

Here we can see that there are four independent samples.

So, we will conduct ANOVA TEST.

Null hypothesis : µ1= µ2 =µ3 = µ4

Alternate hypothesis : Atleast one of them is differentiate

Here we can see 5% of Significance and alpha = 0.05

Here we have to prove Null hypothesis

CODES:

* from google.colab import files

uploaded = files.upload () - CODE FOR IMPORTING FILES

* data = pd.read\_csv (‘ ‘) - TO READ THE DATA FILE
* data.describe()  - TO FIND DESCRIPTIVE STATS
* data.isnull().sum () - CHECKING FOR NULL VALUES

Laboratory 1 0

Laboratory 2 0

Laboratory 3 0

Laboratory 4 0

dtype: int64

* data[data.duplicated()].shape - CHECKING FOR DUPLICATIONS

(0, 4)

* data.info() - CHECKING THE DATA TYPE

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 120 entries, 0 to 119

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Laboratory 1 120 non-null float64

1 Laboratory 2 120 non-null float64

2 Laboratory 3 120 non-null float64

3 Laboratory 4 120 non-null float64

dtypes: float64(4)

memory usage: 3.9 KB

* plt.subplots (figsize = (16,9)) - PLOTING THE DATA[BOXPLOT]

plt.subplot(221)

plt.boxplot(data['Laboratory 1'])

plt.title('Laboratory 1')

plt.subplot(222)

plt.boxplot(data['Laboratory 2'])

plt.title('Laboratory 2')

plt.subplot(223)

plt.boxplot(data['Laboratory 3'])

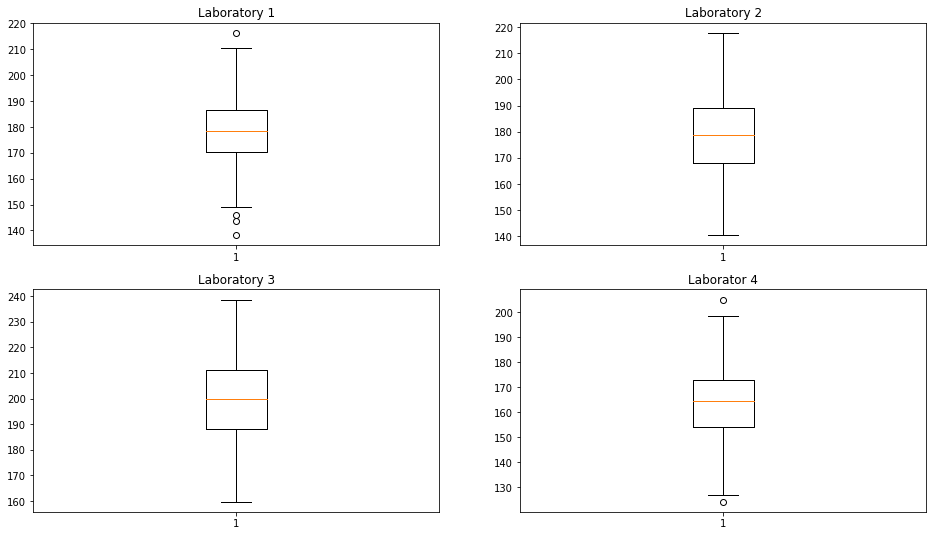
plt.title('Laboratory 3')

plt.subplot(224)

plt.boxplot(data['Laboratory 4'])

plt.title('Laborator 4')

plt.show()



Comparing using t staistics

* test\_statistic, p\_value = stats.f\_oneway(data.iloc[:,0],data.iloc[:,1],data.iloc[:,2], data.iloc[:,3])

print('p\_value = ', p\_value)

p\_value = 2.1156708949992414e-57

### Compare p\_value with 'α '(Significane Level)

### If p\_value is ≠ 'α ' we failed to reject Null Hypothesis because of lack of evidence

### If p\_value is = 'α ' we reject Null Hypothesis

* alpha = 0.05

print('Significnace=%.3f, p=%.3f' % (alpha, p\_value))

if p\_value <= alpha:

     print('We reject Null Hypothesis there is a significance

difference between TAT of reports of the laboratories')

else:

     print('We fail to reject Null hypothesis')

Significnace=0.050, p=0.000

We reject Null Hypothesis there is a significance difference between TAT of reports of the laboratories.

Hence, we fail to reject null hypothesis because of lack of evidence, there is no significant difference between the samples.

### 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 |

### Buyer Ratio.mtw

**ANS:**

**Null Hypothesis:** There is no association or dependency between the gender based buyer rations across regions

**Alternative Hypthoesis :** There is a significant association or dependency between the gender based buyer rations across regions

### Significance is 5% , and alpha = 0.05

### It is one tailed test, alpha = 1-0.95 = 0.05

### CODES:

* from google.colab import files

uploaded=files.upload()

* data=pd.read\_csv("BuyerRatio.csv")

### Applying Chi-Square χ2 contingency table to convert observed value into expected value

* stat, p, dof, exp = stats.chi2\_contingency(data)

print(stat,"\n", p,"\n", dof,"\n", exp)

* observed = np.array([50, 142, 131, 70, 435, 1523, 1356, 750])

expected = np.array([42.76531299,  146.81287862,  131.11756787, 72.30424052, 442.23468701, 1518.18712138, 1355.88243213, 747.69575948])

COMPARISION

* statistics, p\_value = stats.chisquare(observed, expected, ddof = 3)

   print("Statistics = ",statistics,"\n",'P\_Value = ', p\_value)

### Compare p\_value with 'α '(Significane Level)

### If p\_value is ≠ 'α ' we failed to reject Null Hypothesis because of lack of evidence

### If p\_value is = 'α ' we reject Null Hypothesis

* alpha = 0.05

print('Significnace=%.3f, p=%.3f' % (alpha, p\_value))

if p\_value <= alpha:

    print('We reject Null Hypothesis there is a significance difference between TAT of reports of the laboratories')

else:

    print('We fail to reject Null hypothesis')

We fail to reject null hypothesis. Therefore there is no association and dependency between male and female buyer rations and similar across regions so its independent samples.

### 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

**ANS:**

Null Hypothesis: μ1=μ2 = μ3 = μ4

Alternative Hypthosis: Atleast One of them is Differentiate

Here the significance is of 5 % and the alpha = 0.05

CODES:

* from google.colab import files

uploaded = files.upload ()

* data=pd.read\_csv('Costomer+OrderForm.csv')
* data.describe()
* data.isnull().sum()

Phillippines 0

Indonesia 0

Malta 0

India 0

dtype: int64

* data[data.isnull().any(axis=1)]
* data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 300 entries, 0 to 299

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Phillippines 300 non-null object

1 Indonesia 300 non-null object

2 Malta 300 non-null object

3 India 300 non-null object

dtypes: object(4)

* print(data['Phillippines'].value\_counts(),'\n',data['Indonesia'].value\_counts(),'\n',data['Malta'].value\_counts(),'\n',data['India'].value\_counts(),'\n')

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

* contingency\_table = [[271,267,269,280],

                    [29,33,31,20]]

print(contingency\_table)

* stat, p, df, exp = stats.chi2\_contingency(contingency\_table)

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]]

* observed = np.array ([271, 267, 269, 280,29, 33, 31, 20])

expected = np.array ([271.75,271.75,271.75,271.75,28.25,28.25,28.25,28.25])

* 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

### Compare p\_value with 'α '(Significane Level)

### If p\_value is ≠ 'α ' we failed to reject Null Hypothesis because of lack of evidence

### If p\_value is = 'α ' we reject Null Hypothesis

* alpha = 0.05

print('Significnace=%.3f, p=%.3f' % (alpha, p\_value))

if p\_value <= alpha:

print('We reject Null Hypothesis there is a significance difference between TAT of reports of the laboratories')

else:

Significnace=0.050, p=0.425

Hence we fail to reject Null hypothesis.

### 