STATISTICS - 4 - Assignment

By Prakash Ghosh

In [1]:

```
# import libraries
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
import math
```

Problem Statement 1:

Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Question: Are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

In [2]:

```
# Create DataFrame from the given Data
lst_qualification = ['High School','Bachelors','Masters','PHD']
lst_female = [60,54,46,41]
lst_male = [40,44,53,57]
df=pd.DataFrame({'Qualification':lst_qualification,'Count_F': lst_female ,'Count_M': lst_male})
df
```

Out[2]:

	Qualification	Count_F	Count_M
0	High School	60	40
1	Bachelors	54	44
2	Masters	46	53
3	PHD	41	57

Solution-1: Using Z Score and p Value

In [3]:

```
# Add column in the Dataframe for Mean, Standard Deviation, Z Score
# and P Values for Female(F) and Male (M)

df['Mean_F']=df['Count_F'].mean()
df['Mean_M']=df['Count_M'].mean()

df['Std_Dev_F']=df['Count_F'].std()
df['Std_Dev_M']=df['Count_M'].std()

df['Z_F']=stats.zscore(df['Count_F'])
df['Z_M']=stats.zscore(df['Count_M'])

df['p_F']=[stats.norm.cdf(pval) for pval in stats.zscore(df['Count_F'])]
df['p_M']=[stats.norm.cdf(pval) for pval in stats.zscore(df['Count_M'])]
df
```

Out[3]:

	Qualification	Count_F	Count_M	Mean_F	Mean_M	Std_Dev_F	Std_Dev_M	Z_F	Z_M	p_F	p_M
0	High School	60	40	50.25	48.5	8.421203	7.852813	1.336903	-1.249865	0.909373	0.105674
1	Bachelors	54	44	50.25	48.5	8.421203	7.852813	0.514193	-0.661693	0.696442	0.254084
2	Masters	46	53	50.25	48.5	8.421203	7.852813	-0.582752	0.661693	0.280030	0.745916
3	PHD	41	57	50.25	48.5	8.421203	7.852813	-1.268344	1.249865	0.102338	0.894326

In [5]:

```
print('Conclutions from the above table pvalue of Male and Female (more than 5%, there is a relationship \n' \
      'between the gender of an individual and the level of education that they have obtained.\n')

print('Female populations is more at High School and Bachelors')
print('Female populations is less at Masters and PHD\n')

print('Male populations is less at High School and Bachelors')
print('Male populations is more at Masters and PHD')
```

```
Conclutions from the above table pvalue of Male and Female (more than 5%, there is a relationship
between the gender of an individual and the level of education that they have obtained.

Female populations is more at High School and Bachelors
Female populations is less at Masters and PHD

Male populations is less at High School and Bachelors
Male populations is more at Masters and PHD
```

Solution-2: Using Chi- Square Test

In [ ]:

```
# redefine the dataset
df=df[['Qualification','Count_F','Count_M']]
```

In [17]:

```
N = 395          # Sample Size
df['Count_Total']=df.Count_F+df.Count_M

# Expected frequency = ((row total×column)/total sample size
df['ef_F']=(df.Count_F.sum()*df.Count_Total)/N
df['ef_M']=df.Count_Total-df.ef_F

# Chi Sqaure value χ2=∑(Observe freq-Expected Freq)2/Expected Freq
df['chi_F']=[(math.pow((df.Count_F.values[i]-df.ef_F.values[i]),2))/df.ef_F.values[i] for i in range(df.Count_F.count
())]
df['chi_M']=[(math.pow((df.Count_M.values[i]-df.ef_M.values[i]),2))/df.ef_M.values[i] for i in range(df.Count_M.count
())]
df
```

Out[17]:

	Qualification	Count_F	Count_M	Count_Total	ef_F	ef_M	chi_F	chi_M
0	High School	60	40	100	50.886076	49.113924	1.632345	1.691244
1	Bachelors	54	44	98	49.868354	48.131646	0.342311	0.354663
2	Masters	46	53	99	50.377215	48.622785	0.380331	0.394054
3	PHD	41	57	98	49.868354	48.131646	1.577107	1.634012

In [18]:

```
chi_sq_stat =df.chi_F.sum() + df.chi_M.sum()
print("Chi-Square Test Statstic value:\t", chi_sq_stat)
dof = 3       # Degree of Freedom - here dof =3

# Calculate P value from chi_square_stat and degree of freedom using cdf function
p_val = 1 - stats.chi2.cdf(chi_sq_stat,dof)
print("Chi-Square P value\t\t", p_val)

α =0.05  # significance level, confidence level 95%

#Calculate chi-square crtical value
chi_critical= stats.chi2.ppf(0.95,dof)
print("Chi-Square Test Critical value:\t", chi_critical)

print('\nAs Chi-Square Test Statstic value (8.006) greater than Chi-Square Test Critical value (7.815)' \
      '\nby Null hypothesis, it can be concluded Education level depends on gender (at 5% significance level)')
```

```
Chi-Square Test Statstic value:  8.006066246262538
Chi-Square P value               0.04588650089174717
Chi-Square Test Critical value:  7.814727903251179

As Chi-Square Test Statstic value (8.006) greater than Chi-Square Test Critical value (7.815)
by Null hypothesis, it can be concluded Education level depends on gender (at 5% significance level)
```

Problem Statement 2:

Using the following data, perform a oneway analysis of variance using α=.05. Write up the results in APA format.

[Group1: 51, 45, 33, 45, 67]
[Group2: 23, 43, 23, 43, 45]
[Group3: 56, 76, 74, 87, 56]

In [19]:

```
# Create DataFrame from the given Data
lst_group1 = [51, 45, 33, 45, 67]
lst_group2 = [23, 43, 23, 43, 45]
lst_group3 = [56, 76, 74, 87, 56]
df=pd.DataFrame({'Gr1':lst_group1,'Gr2': lst_group2 ,'Gr3': lst_group3})
df
```

Out[19]:

	Gr1	Gr2	Gr3
0	51	23	56
1	45	43	76
2	33	23	74
3	45	43	87
4	67	45	56

In [21]:

```
p_Val=stats.f_oneway(df['Gr1'],df['Gr2'],df['Gr3']).pvalue
F_Val=stats.f_oneway(df['Gr1'],df['Gr2'],df['Gr3']).statistic

α = 0.05                    # Significance level, confidence level 95%

print('Null Hypothesis: \t Group1=Group2=Group3')

print('\nHypothesis testing with 5% significance')

print('\nHere p Value greater than α , so Null Hypothesis(Group1=Group2=Group3) can be Accepted. ')

print('\nWriting up the results in APA format:')

print('\t Significance level:\t', round(α,4))
print('\t F Value:\t\t', round(F_Val,4))
print('\t p Value:\t\t', round(p_Val,4), ' <', round(α,4) , '(Significance level)' )
print('\t So, Accept Null Hypothesis: \t Group1=Group2=Group3' )
```

```
Null Hypothesis:         Group1=Group2=Group3

Hypothesis testing with 5% significance

Here p Value greater than α , so Null Hypothesis(Group1=Group2=Group3) can be Accepted.

Writing up the results in APA format:
         Significance level:     0.05
         F Value:                9.7472
         p Value:                0.0031  < 0.05 (Significance level)
         So, Accept Null Hypothesis:     Group1=Group2=Group3
```

Problem Statement 3:

Calculate F Test for given 10, 20, 30, 40, 50 and 5,10,15, 20, 25.
For 10, 20, 30, 40, 50:

In [22]:

```
# Create DataFrame from the given Data
lst_group1 = [10,20,30,40,50]
lst_group2 = [5,10,15, 20, 25]

df=pd.DataFrame({'Gr1':lst_group1,'Gr2': lst_group2})
df
```

Out[22]:

	Gr1	Gr2
0	10	5
1	20	10
2	30	15
3	40	20
4	50	25

In [23]:

```
# Add column in the Dataframe for Mean, Standard Deviation and Variance

df['Mean_Gr1']=df['Gr1'].mean()
df['Mean_Gr2']=df['Gr2'].mean()

df['Std_Dev_Gr1']=df['Gr1'].std()
df['Std_Dev_Gr2']=df['Gr2'].std()

df['Var_Gr1']=df['Gr1'].var()
df['Var_Gr2']=df['Gr2'].var()
df
```

Out[23]:

	Gr1	Gr2	Mean_Gr1	Mean_Gr2	Std_Dev_Gr1	Std_Dev_Gr2	Var_Gr1	Var_Gr2
0	10	5	30.0	15.0	15.811388	7.905694	250.0	62.5
1	20	10	30.0	15.0	15.811388	7.905694	250.0	62.5
2	30	15	30.0	15.0	15.811388	7.905694	250.0	62.5
3	40	20	30.0	15.0	15.811388	7.905694	250.0	62.5
4	50	25	30.0	15.0	15.811388	7.905694	250.0	62.5

In [26]:

```
# Calculate the P Values
# Hypothesis Test
print('Null Hypothesis Group1 = Group2')

α =0.05  # significance level, confidence level 95%
print('\nSignificance level:\t', round(α,4))

# F test
# F-Test Formula:\t (Varience of Group 1)/(Varience of Group 1)
F_Val=df['Gr1'].var()/df['Gr2'].var()
print('F Test Results:\t\t',F_Val)

p_Val = stats.f.cdf(F_Val, len(df['Gr1'])-1,len(df['Gr1'])-1)

print('p Values is:\t\t',p_Val)

print('\nHere:\t p Value:\t', round(p_Val,4), ' >', round(α,4) , '(Significance level)' )
print('\t So, Reject Null Hypothesis: \t Group1=Group2' )
```

```
Null Hypothesis Group1 = Group2

Significance level:      0.05
F Test Results:          4.0
p Values is:             0.896

Here:    p Value:        0.896  > 0.05 (Significance level)
         So, Reject Null Hypothesis:     Group1=Group2
```