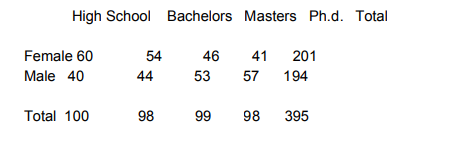
Problem Statement 1: Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:



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?

Solution:

# import libraries

import numpy as np

import pandas as pd

import scipy.stats as stats

from ipykernel import kernelapp as app

import matplotlib.pyplot as plt

import math

# Create DataFrame from the given Data

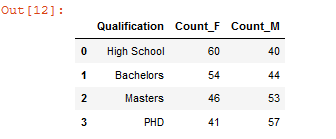
lst\_qualification = ['High School','Bachelors','Masters','PHD']

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



**##USING Z SCORE AND p VALUE**

# 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()

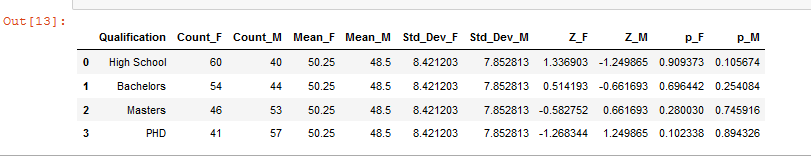
df['Z\_F']=stats.zscore(df['Count\_F'])

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.head()



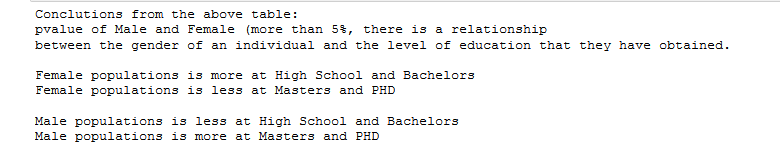
print('Conclutions from the above table: \npvalue of Male and Female (more than 5%, there is a relationship \n''between the gender of an individual and the level of education that they have obtained.\n')

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')



**##Using Chi-square test**

# redefine the dataset

df=df[['Qualification','Count\_F','Count\_M']]

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

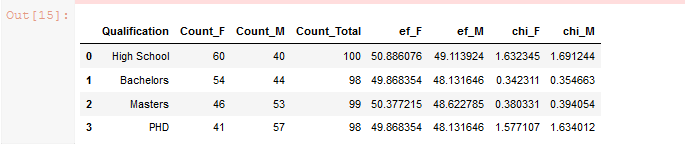
df['ef\_M']=df.Count\_Total-df.ef\_F

# 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



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%

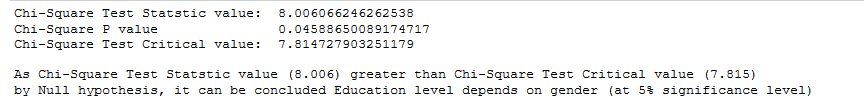
#Calculate chi-square crtical value

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)')



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]

Solution:

# import libraries

import numpy as np

import pandas as pd

import scipy.stats as stats

from ipykernel import kernelapp as app

import matplotlib.pyplot as plt

import math

# Create DataFrame from the given Data

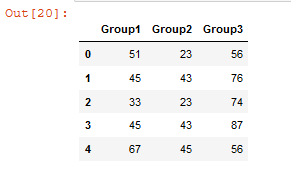
lst\_group1 = [51, 45, 33, 45, 67]

lst\_group2 = [23, 43, 23, 43, 45]

lst\_group3 = [56, 76, 74, 87, 56]

df=pd.DataFrame({'Group1':lst\_group1,'Group2': lst\_group2 ,'Group3': lst\_group3})

df



p\_Val=stats.f\_oneway(df['Group1'],df['Group2'],df['Group3']).pvalue

F\_Val=stats.f\_oneway(df['Group1'],df['Group2'],df['Group3']).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. ')

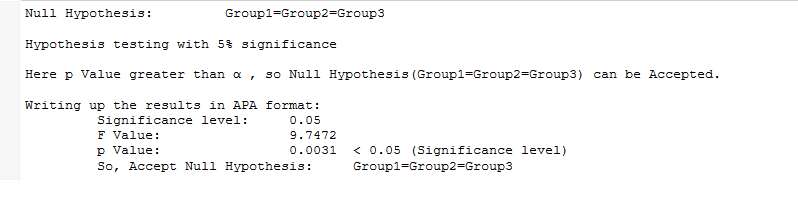
print('\nWriting up the results in APA format:')

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' )



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:

Solution:

# import libraries

import numpy as np

import pandas as pd

import scipy.stats as stats

from ipykernel import kernelapp as app

import matplotlib.pyplot as plt

import math

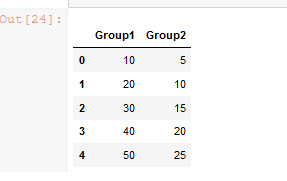
# Create DataFrame from the given Data

lst\_group1 = [10,20,30,40,50]

lst\_group2 = [5,10,15, 20, 25]

df=pd.DataFrame({'Group1':lst\_group1,'Group2': lst\_group2})

df



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

df['Mean\_Group1']=df['Group1'].mean()

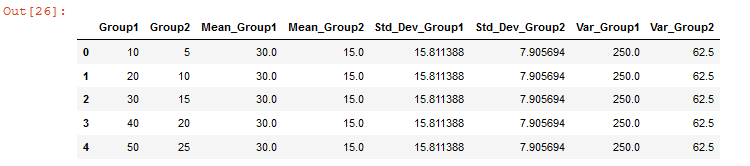
df['Mean\_Group2']=df['Group2'].mean()

df['Std\_Dev\_Group1']=df['Group1'].std()

df['Std\_Dev\_Group2']=df['Group2'].std()

df['Var\_Group1']=df['Group1'].var()df['Var\_Group2']=df['Group2'].var()

df



# 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['Group1'].var()/df['Group2'].var()

print('F Test Results:\t\t',F\_Val)

p\_Val = stats.f.cdf(F\_Val, len(df['Group1'])-1,len(df['Group1'])-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' )

