**Assignment 18.1**

# 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 |
|  | 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?

Answer :

Below is the table of expected counts based on below formula:

E=(row total × column total)/sample size

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| E | High School | Bachelors | Masters | Ph.d. | Total |
| Female | 50.886 | 49.868 | 50.377 | 49.868 | 201 |
| Male | 49.114 | 48.132 | 48.623 | 48.132 | 194 |
|  | 100 | 98 | 99 | 98 | 395 |

Chi-Square Test Statistic

*χ*2=∑(*O*−*E*)2/*E*

So, working this out, *χ*2 =(60−50.886)2/50.886+…+(57−48.132)2/48.132=**8.006**

The critical value of *χ*2 with 3 degree of freedom is 7.815.

Since 8.006 > 7.815, therefore we reject the null hypothesis and conclude that the education level depends on gender at a 5% level of significance.

**Using Python code:**

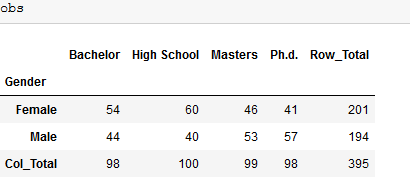
obs= pd.DataFrame({'Gender':['Female','Male','Col\_Total'],'High School':[60,40,100],'Bachelor':[54,44,98],\

'Masters':[46,53,99],'Ph.d.':[41,57,98]})

obs["Row\_Total"]=[201,194,395]

obs.set\_index('Gender',inplace=True)

obs



expected = pd.DataFrame()

for i in obs["Row\_Total"][0:2].values:

expected = expected.append(pd.Series(rows\_list),ignore\_index=True)

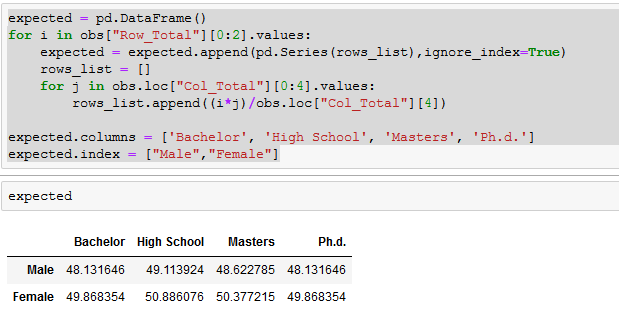
rows\_list = []

for j in obs.loc["Col\_Total"][0:4].values:

rows\_list.append((i\*j)/obs.loc["Col\_Total"][4])

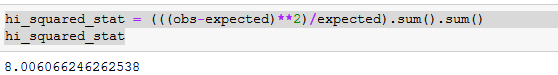
expected.columns = ['Bachelor', 'High School', 'Masters', 'Ph.d.']

expected.index = ["Male","Female"]



hi\_squared\_stat = (((obs-expected)\*\*2)/expected).sum().sum()

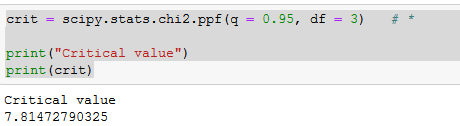
hi\_squared\_stat



crit = scipy.stats.chi2.ppf(q = 0.95, df = 3) # \*

print("Critical value")

print(crit)



Hence chi-square test value 8.006 is greater than CV value 7.814 , therefore null hypothesis is rejected.

# 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]

Answer :

Sample means (*x*¯) for the groups: = 48.2, 35.4, 69.8

**Intermediate steps in calculating the group variances:**

[[1]]

value mean deviations sq deviations

1 51 48.2 2.8 7.84

2 45 48.2 -3.2 10.24

3 33 48.2 -15.2 231.04

4 45 48.2 -3.2 10.24

5 67 48.2 18.8 353.44

[[2]]

value mean deviations sq deviations

1 23 35.4 -12.4 153.76

2 43 35.4 7.6 57.76

3 23 35.4 -12.4 153.76

4 43 35.4 7.6 57.76

5 45 35.4 9.6 92.16

[[3]]

value mean deviations sq deviations

1 56 69.8 -13.8 190.44

2 76 69.8 6.2 38.44

3 74 69.8 4.2 17.64

4 87 69.8 17.2 295.84

5 56 69.8 -13.8 190.44

Sum of squared deviations from the mean (SS) for the groups:

[1] 612.8 515.2 732.8

Var1=612.8/5−1=153.2

Var2=515.2/5−1=128.8

Var3=732.8/5−1=183.2

MSerror=(153.2+128.8+183.2)/3=155.07 - *Note: this is just the average within-group variance; it is not sensitive to group mean differences!*

Calculating the remaining error (or within) terms for the ANOVA table:

dferror=15−3=12

SSerror=(155.07)(15−3)=1860.8

Intermediate steps in calculating the variance of the sample means:

Grand mean (x¯grand) = (48.2+35.4+69.8)/3=51.13

group mean grand mean deviations sq deviations

48.2 51.13 -2.93 8.58

35.4 51.13 -15.73 247.43

69.8 51.13 18.67 348.57

Sum of squares (SSmeans)=604.58

Varmeans=604.58/(3−1)=302.29

MSbetween=(302.29)(5)=1511.45

Note: This method of estimating the variance IS sensitive to group mean differences!

Calculating the remaining between (or group) terms of the ANOVA table:

dfgroups=3−1=2

SSgroup=(1511.45)(3−1)=3022.9

Test statistic and critical value

F=1511.45/155.07=9.75

Fcritical(2,12)=3.89

 Decision: reject H0

ANOVA table

| source | SS | df | MS | F |
| --- | --- | --- | --- | --- |
| group | 3022.9 | 2 | 1511.45 | 9.75 |
| error | 1860.8 | 12 | 155.07 |  |
| total | 4883.7 |  |  |  |

Effect size

η2=3022.9/4883.7=0.62

APA writeup :

F(2, 12)=9.75, p <0.05, η2

=0.62.

# 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:

Answer :

**Calculate Variance of first set**   
  
Total Inputs (N) =(10,20,30,40,50)   
Total Inputs (N)=5   
Mean (xm)= (x1+x1+x2...xn)/N   
Mean (xm)= 150/5   
Means(xm)= 30   
SD=sqrt(1/(N-1)\*((x1-xm)2+(x2-xm)2+..+(xn-xm)2))   
=sqrt(1/(5-1)((10-30)2+(20-30)2+(30-30)2+(40-30)2+(50-30)2))   
=sqrt(1/4((-20)2+(-10)2+(0)2+(10)2+(20)2))   
=sqrt(1/4((400)+(100)+(0)+(100)+(400)))   
=sqrt(250)   
=15.8114   
Variance=SD2   
Variance=15.81142   
Variance=250

**Calculate Variance of second set**   
For 5, 10,15,20,25:   
Total Inputs(N) =(5,10,15,20,25)   
Total Inputs(N)=5   
Mean (xm)= (x1+x2+x3...xN)/N   
Mean (xm)= 75/5   
Means (xm)= 15   
SD=sqrt(1/(N-1)\*((x1-xm)2+(x2-xm)2+..+(xn-xm)2))   
=sqrt(1/(5-1)((5-15)2+(10-15)2+(15-15)2+(20-15)2+(25-15)2))   
=sqrt(1/4((-10)2+(-5)2+(0)2+(5)2+(10)2))   
=sqrt(1/4((100)+(25)+(0)+(25)+(100)))   
=sqrt(62.5)   
=7.9057   
Variance=SD2   
Variance=7.90572   
Variance=62.5

**To calculate F Test**   
F Test = (variance of 10, 20,30,40,50) / (variance of 5, 10, 15, 20, 25)   
= 250/62.5   
= 4.   
  
The F Test value is 4.