Data Science Masters :Assignment 18

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:

	HS	BE	MS	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?

Solution:

Step 1) Define Hypothesis

H0 => Gender and Education are independent
HA => Gender and Education are dependent

We will be using Chi-Square χ2 test to check the dependency of two given catagorial variable (Education and Gender)

Step 2) Here's the table of expected counts:

	High School	Bachelors	Masters	Ph.d.	Total
Female	(201*100)/395	(201*98)/395	(201*99)/395	(201*98)/395	201
Male	(194*100)/395	(194*98)/395	(194*99)/395	(194*98)/395	194
Total	100	98	99	98	395

	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
Total	100	98	99	98	395

Step 3) Degrees of freedom => (4-1)(2-1) = 3

Step 4) Chi Square χ2 computation
χ2 = Summation[(Oi-Ei)^2/Ei]
So, working this out,
χ2=(60–50.886)^2/50.886 + (54–49.868)^2/49.868 + (46–50.377)^2/50.377 + (41–49.868)^2/49.868 + (40–49.114)^2/49.114 + (44–48.132)^2/48.132 + (53–48.635)^2/48.635 + (57–48.132)^2/48.132

= 8.006

Step 5) Conclude result
The critical value of χ2 with 3 degree of freedom is 7.815 (from table).
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.

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:

Step 1)Calculate all the means
Sample means for the groups: = 48.2, 35.4, 69.8

Step 2)Define Hypothesis

H0: Mean(G1) = Mean(G2) = Mean(G3)
Ha: Mean(G1) ≠ Mean(G2) ≠ Mean(G3)

We also specify the as well as the rejection criteria.
alpha = 0.05
Rejection criteria: K0.05 < F
This means that if the critical value of F from tables is less than the calculated value of F, we reject the null hypothesis

Step 3)Calculate the Sum of Squares

Intermediate steps in calculating the group variances:

Group 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

Group 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

Group 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:

G1 = 612.8 G2=515.2 G3=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 + 183.23 = 155.07

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+69.83=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

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

dfgroups=3−1=2

SSgroup=(1511.45)(3−1)=3022.9

Step 4)Calculate the Test statistic and critical value

F=1511.45155.07=9.75

Fcritical(2,12)=3.89

Since the calculated(absolute value) of F is greater than the tabulated value, we
reject the null hypothesis and conclude that at least two of the means are
significantly different from each other.

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 = 3022.9 / 4883.7 = 0.62

APA writeup ->

F(2, 12)=9.75, p <0.05, Effect size=0.62.
```

```
Problem Statement 3:
Calculate F Test for given 10, 20, 30, 40, 50 and 5,10,15, 20, 25.

# Solution:

Step 1) Calculate Variance of first set (10, 20, 30, 40, 50)

Mean =>
     = (x1+x1+x2...xn)/N i.e. N=5
Mean = 150/5
Mean = 30

Std.Dev =>
        =  sqrt(1/(5-1)((10-30)^2+(20-30)^2+(30-30)^2+(40-30)2+(50-30)2))
        =  sqrt(250)
        =  15.8114

Variance = Std.Dev^2
Variance = 15.8114^2
Variance(var1) = 250.00037 = 250


Step 2) Calculate Variance of second set (5, 10,15,20,25)
```

```
Mean =>
     = 75/5
Mean = 15

Std.Dev =>
   = sqrt(1/(5-1)((5-15)2+(10-15)2+(15-15)2+(20-15)2+(25-15)2))
   = sqrt(62.5)
   = 7.9057

Variance = Std.Dev^2
Variance = 7.9057^2
Variance(var2) = 62.5

Step 3) Calculate F Test using the below formuala
F Test  = var1/var2
        = 250/62.5
        = 4.00000592
        = 4

The F Test value is 4
```