Statistics Assignment 4
Submitted by: Amit Kumar
21/10/2020

In [2]:
```
# 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 [1]:
```
#Chi-Square test of independence
#H0 :Null Hypothesis: The two categorical variables are independent.
#H1:Alternative Hypothesis: The two categorical variables are dependent.

import numpy as np
import pandas as pd
import scipy.stats as stats

male = [40,44,53,57]
female = [60,54,46,41]
High_school=[60,40]
Bachelors = [54,44]
Masters = [46,53]
Phd = [41,57]

marks = male+female
print(marks)
sex=['M','M','M','M','F','F','F','F']
education =['High_school','Bachelors','Masters','Ph.d','High_school','Bachelors','Masters','Ph.d']
DF=pd.DataFrame({"Education":education,"Marks":marks,"Sex":sex})
DF
print(DF)
```

```
[40, 44, 53, 57, 60, 54, 46, 41]
     Education  Marks Sex
0  High_school     40   M
1    Bachelors     44   M
2      Masters     53   M
3         Ph.d     57   M
4  High_school     60   F
5    Bachelors     54   F
6      Masters     46   F
7         Ph.d     41   F
```

In [2]:
```python
cross_tab = pd.crosstab([DF.Sex,DF.Marks],DF.Education,margins=True)
cross_tab
```

Out[2]:

Sex	Marks	Bachelors	High_school	Masters	Ph.d	All
F	41	0	0	0	1	1
	46	0	0	1	0	1
	54	1	0	0	0	1
	60	0	1	0	0	1
M	40	0	1	0	0	1
	44	1	0	0	0	1
	53	0	0	1	0	1
	57	0	0	0	1	1
All		2	2	2	2	8

In [3]:
```python
DF1 = pd.crosstab(DF.Sex, DF.Education,DF.Marks, aggfunc="sum",margins=True)
DF1
```

Out[3]:

Education Sex	Bachelors	High_school	Masters	Ph.d	All
F	54	60	46	41	201
M	44	40	53	57	194
All	98	100	99	98	395

In [4]:
```python
DF1.columns = ["Bachelors","High_School","Masters","Ph.d.","Genderwise_total"]
DF1.index = ["Female","Male","Combined"]
DF1
```

Out[4]:

	Bachelors	High_School	Masters	Ph.d.	Genderwise_total
Female	54	60	46	41	201
Male	44	40	53	57	194
Combined	98	100	99	98	395

In [5]:
```python
# Creating a table exlcuding the total for later use
DF2 = DF1.iloc[0:2,0:4]
DF2
```

Out[5]:

	Bachelors	High_School	Masters	Ph.d.
Female	54	60	46	41
Male	44	40	53	57

In [6]:

```python
# For a test of independence, we use the same chi-squared formula that we used
for the goodness-of-fit test.
# The  main difference is we have to calculate the expected counts of each cell
in a 2-dimensional table instead  of
# a 1-dimensional table. To get the expected count for a cell, multiply the ro
w total for that cell by the column
# total for that cell and then divide  by  the  total number  of  observations.  We
 can  quickly  get  the  expected  counts
# for all cells in the table by taking the row totals and column totals of the
table, performing an outer product
# on them with the np.outer() function and dividing by the number of observati
ons:

DF3=np.outer(DF1["Genderwise_total"][0:2],DF1.loc["Combined"][0:4]) / 395.0
DF3 = pd.DataFrame(DF3)
DF3.columns = ["Bachelors","High_School","Masters","Ph.d."]
DF3.index = ["Female","Male"]
DF3
```

Out[6]:

	Bachelors	High_School	Masters	Ph.d.
Female	49.868354	50.886076	50.377215	49.868354
Male	48.131646	49.113924	48.622785	48.131646

In [7]:

```python
# Now we will calculate the chisquare statistic, critical value and p value.
# We  called the .sum() twice, once to get the column sum  and second time to ad
d  the  sum,  returning  the  sum of  entire
# 2D table

chi_squared_stat = (((DF3-DF2)**2)/DF3).sum().sum()
print(chi_squared_stat)
```

```
8.006066246262538
```

In [9]:

```python
#Find the critical value for 95% confidence and degree of freedom (df) is 3
cvalue = stats.chi2.ppf(q = 0.95,df= 3)
print("Critical  value")
print(cvalue)
```

```
Critical value
7.814727903251179
```

In [10]:

```python
# Find the p-value
p_value = 1 - stats.chi2.cdf(x=chi_squared_stat,df=3)
print("P value")
print(p_value)
```

```
P value
0.04588650089174717
```

In [11]:
```
# Use stats.chi2_contingency() function to conduct a test of independence auto
matically given a frequency table
# of observed counts:
result = stats.chi2_contingency(observed= DF2)
print(result)
print('-'*115)
print('The output shows the chi-square statistic = 8, the p-value as 0.045 and
the degrees of freedom as 3')
print('The critical value with 3 degree of freedom is 7.815. Since 8.006 > 7.8
15, therefore we reject the null hypothesis and conclude that the education le
vel depends on gender at a 5% level of significance.')
```

```
(8.006066246262538, 0.045886500891747214, 3, array([[49.86835443, 50.8860759
5, 50.37721519, 49.86835443],
       [48.13164557, 49.11392405, 48.62278481, 48.13164557]]))
-------------------------------------------------------------------------------
------------------------------------
The output shows the chi-square statistic = 8, the p-value as 0.045  and  the  d
egrees of freedom as 3
The critical value with 3 degree of freedom is 7.815. Since 8.006 > 7.815, th
erefore we reject the null hypothesis and conclude that the education level d
epends on gender at a 5% level of significance.
```

In [12]:
```
# 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, 7
4, 87, 56]
```

In [16]:

```
#The analysis of variance or ANOVA is a statistical inference test that lets y
ou compare multiple groups at the same
# time. The one-way ANOVA tests whether the mean of some numeric variable diff
ers across the levels of one categorical
# variable.It essentially answers the question: do any of the group means diff
er from one another?

#The scipy library has a function for carrying out one-way ANOVA tests called
 scipy.stats.f_oneway()
import scipy.stats as stats
Group1 = [51, 45, 33, 45, 67]
Group2 = [23, 43, 23, 43, 45]
Group3 = [56, 76, 74, 87, 56]
# Perform the ANOVA
statistic, pvalue = stats.f_oneway(Group1,Group2,Group3)
print("F Statistic value {} , p-value {}".format(statistic,pvalue))
if pvalue < 0.05:
    print('True')
else:
    print('False')
print("-"*115)
print("The test result suggests the groups have different same sample means in this example, since the p-value is significant at a 99% confidence level. Here the p-value returned is 0.00305 which is < 0.05")
```

```
F Statistic value 9.747205503009463 , p-value 0.0030597541434430556
True
----------------------------------------------------------------------------
---------------------------------------
The test result suggests the groups have different same sample means in this
example, since the p-value is significant at a 99% confidence level. Here the p-
value returned is 0.00305 which is < 0.05
```

In [ ]:

```
# Problem Statement 3:
# Calculate F Test for given 10, 20, 30, 40, 50 and 5,10,15, 20, 25. For 10, 2
0, 30, 40, 50:
```

In [14]:
```
stats.f_oneway([10, 20, 30, 40, 50],[5,10,15, 20, 25])

Group1 = [10, 20, 30, 40, 50]
Group2 = [5,10,15, 20, 25]
mean_1 = np.mean(Group1)
mean_2 = np.mean(Group2)
grp1_sub_mean1 = []
grp2_sub_mean2 = []
add1 = 0
add2 = 0
for items in Group1:
    add1 += (items - mean_1)**2
for items in Group2:
    add2 += (items - mean_2)**2
var1 = add1/(len(Group1)-1)
var2 = add2/(len(Group2)-1)

F_Test = var1/var2
print("F Test for given 10, 20, 30, 40, 50 and 5,10,15, 20, 25 is : ",F_Test)
```

```
F Test for given 10, 20, 30, 40, 50 and 5,10,15, 20, 25 is :  4.0
```