# 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 [7]:

```
import scipy.stats as sts
from scipy.stats import norm
import math
import numpy as np
import pandas as pd

lfemale = [60,54,46,41]
lmale = [40,44,53,57]
s = [40,60]
b = [44,54]
m = [53,46]
p = [57,41]
marks = lfemale + lmale

sex =  ['Male','Male','Male','Male','Female','Female','Female','Female'
]
```

```
edu = ['High School', 'Bachelors', 'Masters', 'Ph.d.','High School', 'B
achelors', 'Masters', 'Ph.d.']
df_edu = pd.DataFrame({"Sex":sex,"Edu":edu,"Marks":marks})

print(df_edu)
```

```
[60, 54, 46, 41, 40, 44, 53, 57]
           Edu  Marks     Sex
0  High School     60    Male
1    Bachelors     54    Male
2      Masters     46    Male
3        Ph.d.     41    Male
4  High School     40  Female
5    Bachelors     44  Female
6      Masters     53  Female
7        Ph.d.     57  Female
```

In [8]:

```
df2 = pd.crosstab(df_edu.Sex, df_edu.Edu,df_edu.Marks, aggfunc="sum",ma
rgins=True)

df2.columns = ["Bachelors","High School","Masters","Ph.d.","row_totals"
]

df2.index = ["Female","Male","col_totals"]

df2
```

Out[8]:

| | Bachelors | High School | Masters | Ph.d. | row_totals |
|---|---|---|---|---|---|
| **Female** | 44 | 40 | 53 | 57 | 194 |
| **Male** | 54 | 60 | 46 | 41 | 201 |
| **col_totals** | 98 | 100 | 99 | 98 | 395 |

In [9]:

```
observed = df2.iloc[0:2,0:4]   # Get table without totals for later use
observed
```

Out[9]:

| | Bachelors | High School | Masters | Ph.d. |
|---|---|---|---|---|
| **Female** | 44 | 40 | 53 | 57 |
| **Male** | 54 | 60 | 46 | 41 |

In [13]:
```
expected =  np.outer(df2["row_totals"][0:2],
                     df2.loc["col_totals"][0:4]) / 395.0
expected = pd.DataFrame(expected)
expected.columns = ["Bachelors","High School","Masters","Ph.d."]
expected.index = ["Female","Male"]
expected
```

```
Female    194
Male      201
Name: row_totals, dtype: int64
[[19012 19400 19206 19012]
 [19698 20100 19899 19698]]
```

Out[13]:

| | Bachelors | High School | Masters | Ph.d. |
|---|---|---|---|---|
| **Female** | 48.131646 | 49.113924 | 48.622785 | 48.131646 |
| **Male** | 49.868354 | 50.886076 | 50.377215 | 49.868354 |

In [14]:
```
# We call .sum() twice: once to get the column sums and a second time t
o add the column sums together, returning the sum of the entire 2D tabl
e

chi_squared_stat = (((observed-expected)**2)/expected).sum().sum()

print(chi_squared_stat)
```

```
8.006066246262538
```

In [29]:
```
#The degrees of freedom for a test of independence equals the product o
f the number of categories in each variable minus 1.
#In this case we have a 2x4 table so df = 1x3 = 3.
```

```
critical = sts.chi2.ppf(q = 0.95, # Find the critical value for 95% con
fidence*
                      df = 3)   # *

print("Critical value", critical)

p_value = 1 - sts.chi2.cdf(x=chi_squared_stat, df=3)  # Find the p-valu
e

print("P value", p_value)
```

```
Critical value 7.8147279032511765
P value 0.04588650089174717
```

In [16]:
```
sts.chi2_contingency(observed= observed)
```

Out[16]:
```
(8.006066246262538,
 0.045886500891747214,
 3,
 array([[48.13164557, 49.11392405, 48.62278481, 48.13164557],
        [49.86835443, 50.88607595, 50.37721519, 49.86835443]]))
```

The output shows the chi-square statistic = 8, the p-value as 0.045 and the degrees of freedom as 3 followed by the expected counts. The critical value 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.

# 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 [28]:
```
Group1 = [51, 45, 33, 45, 67]
```

```
Group2 = [23, 43, 23, 43, 45]
Group3 = [56, 76, 74, 87, 56]

# ANOVA Test
statistic, pvalue = sts.f_oneway(Group1,Group2,Group3)

print("F Statistic value {} , p-value {}".format(statistic,pvalue))

if pvalue < 0.05:
    print('\nTrue')
else:
    print('\nFalse')

print("\nThe test result suggests the groups don't have the same sample
means in this case, since the p-value is significant at a 99% confidenc
e 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 don't have the same sample means in
this case, since the p-value is significant at a 99% confidence level.
Here the p-value returned is 0.00305 which is < 0.05
```

# 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 [27]:

```
grp1 = [10, 20, 30, 40, 50]
grp2 = [5,10,15, 20, 25]

mean_1 = np.mean(grp1)
mean_2 = np.mean(grp2)

sum_grp1 = 0
sum_grp2 = 0
```

```
for items in grp1:
    sum_grp1 += (items - mean_1)**2

for items in Group2:
    sum_grp2 += (items - mean_2)**2

var1 = sum_grp1/(len(grp1)-1)
var2 = sum_grp2/(len(grp2)-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
```