Anexo-Chi-2 test example

July 17, 2017

In [ ]:
```
#Test chi-2: Es un test que busca probar la independencia entre variables.
#Referencias:
#http://hamelg.blogspot.cl/2015/11/python-for-data-analysis-part-25-chi.html
#https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html
#https://www.slideshare.net/armando310388/prueba-chicuadrado
```

In [1]:
```
import pandas as pd
from pandas import read_csv
import numpy as np
import scipy as sp
import scipy.stats as stats
```

In [2]:
```
#El fragmento siguiente carga el conjunto de datos de inicio de diabetes de los indios
#Link a los datos https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
url = "https://goo.gl/vhm1eU"
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
df = read_csv(url, names=names)
df.head()
```

Out[2]:

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

In [3]:
```
#Primero calculamos los valores observados que son los que asignamos en la tabla de co
contingencyTable = pd.crosstab(index=df['class'],columns=df['preg'],margins=True)
#Ahora calculamos los valores esperados
expected =  np.outer(contingencyTable.iloc[0:2]['All'],
                     contingencyTable.loc["All"][0:2]/ 768)
expected = pd.DataFrame(expected)
expected.columns = contingencyTable.columns[0:2]
expected.index = contingencyTable.index[0:2]
expected
```

Out[3]:
```
preg           0          1
class
```

```
0      72.265625  87.890625
1      38.734375  47.109375
```

In [4]:
```
contingencyTable = pd.crosstab(index=df['class'],columns=df['preg'])
chi_squared_stat = (((contingencyTable-expected)**2)/expected).sum().sum()
print("el x2 calculado es de: "+str(chi_squared_stat))
```

```
el x2 calculado es de: 10.714170122361168
```

In [5]:
```
#Note: The degrees of freedom for a test of independence equals the product of the num
#each variable minus 1. In this case we have a 2x2 table so df = 1x1 = 1.
#Encontramos el valor crítico a un 95% de confianza*
crit = sp.stats.chi2.ppf(q = 0.99,df = 1)
print("El valor crítico es de:")
print(crit)
#Encontramos el p-valor
p_value = 1 - sp.stats.chi2.cdf(x=chi_squared_stat,df=1)
print("el p-valor es de:")
print(p_value)
```

```
El valor crítico es de:
6.63489660102
el p-valor es de:
0.00106318133612
```

In [6]:
```
chi_squared_stat<=crit
```

Out[6]: False

In [7]:
```
#Recordemos que el test X2 tiene como hipótesis nula-->H0:La variable class es indepen
#El output entrega el valor del estadístico chi-2, el valor p, los grados de libertad
#esperados.Como podemos ver, en este caso no se acepta H0,ya que el p-valor es menor q
#x2 calculado es mayor que el estadístico x2. En consecuencia existe una relación entr
contingencyTable = pd.crosstab(index=df['class'],columns=df['preg'])
sp.stats.chi2_contingency(observed=contingencyTable,correction=False)
#Al agregar el parámetro correction como falso le estamos pidiendo que no aplique la c
```

Out[7]:
```
(64.594808687230056,
 8.648349123362548e-08,
 16,
 array([[ 72.265625  ,  87.890625  ,  67.05729167,  48.828125  ,
          44.27083333,  37.109375  ,  32.55208333,  29.296875  ,
          24.73958333,  18.22916667,  15.625     ,   7.16145833,
           5.859375  ,   6.51041667,   1.30208333,   0.65104167,
           0.65104167],
        [ 38.734375  ,  47.109375  ,  35.94270833,  26.171875  ,
          23.72916667,  19.890625  ,  17.44791667,  15.703125  ,
          13.26041667,   9.77083333,   8.375     ,   3.83854167,
           3.140625  ,   3.48958333,   0.69791667,   0.34895833,
           0.34895833]]))
```