Climent

June 27, 2020

[1]:
```
import scipy.stats as scs
import requests
import numpy as np
import pandas as pd
```

0.1 Cancer recurrency rates comparison (chemo VS non-chemo)

Authors on Climent et al. assessed recurrency after chemotherapy in breast cancer patients with negative lymph nodes. The difference in the rate of recurrency after chemotherapy was not found to be significant. In this notebook we are going to replicate that result.

0.1.1 Data

185 patients with lymph node–negative breast cancer. Biopsies were selected randomly from a pool of cryopreserved tumors from 1979 to 2000 at the University of Valencia if they complied with the following: a) invasive breast carcinoma of any size; b) mastectomy or surgery with or without radiotherapy; c) negative lymph-node d) complete clinical data e) 50% or more tumor cells in sample. Data is public and available at http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE6448

0.1.2 Reference:

Climent J et al. (2007) Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. Cancer Research 67: 818-826.

0.1.3 Retrieving the data

Download the compressed file to unpack it (Need to run only once)

[95]:
```
url = 'https://ftp.ncbi.nlm.nih.gov/geo/series/GSE6nnn/GSE6448/miniml/
 ↪GSE6448_family.xml.tgz'
r = requests.get(url, allow_redirects=True)
open('GSE6448_family.xml.tgz', 'wb').write(r.content)
```

[95]:
```
2999132
```

Unpack only the file of interest

[104]:
```
!tar -zxvf GSE6448_family.xml.tgz GSE6448-tbl-1.txt
```

```
x GSE6448-tbl-1.txt
```

[3]:
```
columns = ['Id', 'TumorNo', 'Age', 'HormStatus', 'TNM', 'Stage', 'Gender',␣
↪'Recurrence', 'Treatment', 'DFSmonths', 'ERpos', 'PRpos']
```

[32]:
```
clim_table = 'GSE6448-tbl-1.txt'
clim = pd.read_csv(clim_table,sep='\t',header=None, names = columns,␣
↪usecols=list(range(1,12)))
```

[31]:
```
clim.head()
```

[31]:

	TumorNo	Age	HormStatus	TNM	Stage	Gender	Recurrence	\
0	19	35	PREMENOPAUSIC	T1N0M0	I	FEMALE	1	
1	49	49	PREMENOPAUSIC	T1N0M0	I	FEMALE	1	
2	139	71	POSTMENOPAUSIC	T2N0M0	II	FEMALE	0	
3	154	42	PREMENOPAUSIC	T1N0M0	I	FEMALE	0	
4	203	29	PREMENOPAUSIC	T2N0M0	II	FEMALE	1	

	Treatment	DFSmonths	ERpos	PRpos
0	ADCHEM: Anthracycline	166.03	NEGATIVE	POSITIVE
1	ADCHEM: Anthracycline	67.20	POSITIVE	POSITIVE
2	ADCHEM: Anthracycline	170.90	POSITIVE	POSITIVE
3	ADCHEM: Anthracycline	173.60	NEGATIVE	POSITIVE
4	ADCHEM: Anthracycline	153.37	NEGATIVE	NEGATIVE

0.1.4 Data

From the 185 women, 90 received anthracycline-based chemotherapy (CHEMO group) and 95 did not. The mayority of those with positive ER or PR tumor also received tamoxifen (Chemo or not). Some patients did not receive any treatment

[33]:
```
pd.crosstab([clim.ERpos, clim.PRpos],clim.Treatment)
```

[33]:

Treatment		ADCHEM: Anthracycline	ADH:Tamoxifen	No Treatment
ERpos	PRpos			
.	.	1	1	0
NEGATIVE	NEGATIVE	24	4	14
	POSITIVE	9	6	3
POSITIVE	NEGATIVE	8	11	4
	POSITIVE	34	29	15

Table above shows two samples with an incorrect value of "." for ERpos and PRpos that should be recoded as missing values. Let's take care of that.

[34]:
```
clim.ERpos = clim.ERpos.replace('.',np.nan)
clim.PRpos = clim.PRpos.replace('.',np.nan)
```

[35]:
```
pd.crosstab([clim.ERpos, clim.PRpos],clim.Treatment)
```

[35]:

Treatment		ADCHEM: Anthracycline	ADH:Tamoxifen	No Treatment
ERpos	PRpos			
NEGATIVE	NEGATIVE	24	4	14
	POSITIVE	9	6	3
POSITIVE	NEGATIVE	8	11	4
	POSITIVE	34	29	15

Let's create the Chemo group

[36]:
```
clim['Chemo']=clim.Treatment
dicothomic = {'ADCHEM: Anthracycline': 'Chemo', 'ADH:Tamoxifen': 'NoChemo', 'No␣
↪Treatment': 'NoChemo'}
clim['Chemo']=clim.Chemo.replace(dicothomic)
```

[37]:
```
pd.crosstab([clim.ERpos, clim.PRpos],clim.Chemo)
```

[37]:

Chemo		Chemo	NoChemo
ERpos	PRpos		
NEGATIVE	NEGATIVE	24	18
	POSITIVE	9	9
POSITIVE	NEGATIVE	8	15
	POSITIVE	34	44

0.1.5 Missing data

There are 24 samples missing both ER and PR status

[38]:
```
clim[['ERpos', 'PRpos','Chemo', 'Recurrence']].isna().sum()
```

[38]:
```
ERpos         24
PRpos         24
Chemo          0
Recurrence     0
dtype: int64
```

Whenever ERpos is missing, it is also missing for PRpos.

[39]:
```
len(clim[clim.ERpos.isna() & clim.PRpos.isna()])
```

[39]:
```
24
```

0.2 Is recurrency rate related to chemotherapy?

Below are the relative and absolute frequencies in a contingency table for recurrence and chemotherapy

[40]:
```
pd.crosstab(clim.Chemo,clim.Recurrence, normalize='index')
```

[40]:
```
Recurrence         0         1
Chemo
Chemo       0.722222  0.277778
NoChemo     0.747368  0.252632
```

[41]:
```
t_rec = pd.crosstab(clim.Chemo,clim.Recurrence)
t_rec
```

[41]:
```
Recurrence   0   1
Chemo
Chemo       65  25
NoChemo     71  24
```

Recurrence rate for those undertaken Chemo is actually higher than the rate for those that did not have chemotherapy. We can still test if the rate is the same for both groups using a Chi-square test. That is H0: p1 = p2 where p1 is the recurrence proportion for those who had chemo and p2 the proportion for those who didn't have chemo.

[43]:
```
stat, p, dof, expected = scs.chi2_contingency(t_rec, correction=False)
print('p-value=%.4f' % (p))
print('Expected values:\n',expected)
```

```
p-value=0.6985
Expected values:
 [[66.16216216 23.83783784]
 [69.83783784 25.16216216]]
```

The alternative hypothesis is not rejected (p-value=0.6985). Therefore, there is not statistical evidence to think that the proportions differ.

0.3 Does treatment have an effect on recurrence?

The focus of the study is the effect of chemotherapy. However Tamoxifen was considered as an additional treatment; And a group of patients with no treatment is available too leaving the experiment with three groups to compare from. Tamoxifen is actually the group with the lowest recurrency rate. Is the difference statistically significant?

First we will check absolute and relative frequencies with contingency tables:

[44]:
```
t_allT = pd.crosstab(clim.Treatment,clim.Recurrence)
t_allT
```

[44]:
```
Recurrence              0   1
Treatment
ADCHEM: Anthracycline  65  25
ADH:Tamoxifen          47  12
No Treatment           24  12
```

[45]:
```
pd.crosstab(clim.Treatment,clim.Recurrence,normalize='index')
```

[45]:
```
Recurrence                   0         1
Treatment
ADCHEM: Anthracycline  0.722222  0.277778
ADH:Tamoxifen          0.796610  0.203390
No Treatment           0.666667  0.333333
```

Now we can test for difference in proportions among the three groups

[46]:
```
stat, p, dof, expected = scs.chi2_contingency(t_allT)
print('p-value=%.4f' % (p))
print('Expected values:\n',expected)
```

```
p-value=0.3519
Expected values:
 [[66.16216216 23.83783784]
 [43.37297297 15.62702703]
 [26.46486486  9.53513514]]
```

The null hypothesis for equallity of proportions (H0: p1=p2=p3) is not rejected (p-value = 0.3519). There is no statiscal evidence to think that there is a difference of proportions among the three groups

0.4 Conclusions

- Recurrency rate is higher in the sample for those that went under chemo than for those who didn't. Difference in recurrence rate for these two groups is not statistically significant (p-value=0.6985)
- There is not evidence to indicate that recurrency rate is associated to treatment options (p-value=0.3519: No-treatment, Tamoxifen, Chemo)