Unit4HW

May 10, 2020

1 Unit4 作业

[1]:

```
import scipy.stats as stats
import numpy as np
import statsmodels.stats.proportion as proportion
from IPython.core.interactiveshell import InteractiveShell
import matplotlib.pyplot as plt
import seaborn as sns
InteractiveShell.ast_node_interactivity = 'all'
```

HW-U4-1：请说明不同类型类型的卡方检验条件，并分别给出例子 （1分）

1. 单样本卡方检验 条件：频数不能过低。 例子：检验一个骰子是否均匀
2. 2×2联立表。条件：每个频数均不得小于5。例子：检验一种药物是否作用在治疗前和治疗后的影响。
3. $R \times C$联立表。条件：80%以上的频数不得小于5。例子：检验一道选择题一个班级的答案与标准答案的符合程度。
4. Yates卡方检验 条件：某个频数出现过少，需要进行不连续校正。

Drug	cured	uncured
A	510	100
B	110	50
total	620	150

HW-U4-2: 对两组病人分别进行两组（A,B)药物实验, 结果是治愈和未治愈，数据如下：

(1)计算两种药物治愈率的99%置信区间； （0.5分）

[3]:

```
def proportion_ci_wilson(p,alpha,n):
    za=stats.norm.isf(q=alpha/2,loc=0,scale=1)
    dn=2*(n+za**2)
    nl=2*(n*p+za**2)-(za*np.sqrt(za**2-1/n+4*n*p*(1-p)+4*p-2)+1)
    nu=2*(n*p+za**2)+(za*np.sqrt(za**2-1/n+4*n*p*(1-p)-4*p-2)+1)
    wl=max(0,nl/dn)
    wu=min(1,nu/dn)
    return wl,wu
def proportion_ci_asym(p,alpha,n):
    z_alpha005=stats.norm.isf(q=alpha/2,loc=0,scale=1)
    sigmap=np.sqrt(p*(1-p)/n)
    return p-z_alpha005*sigmap,p+z_alpha005*sigmap

proportion_ci_asym(510/610,0.01,610)
proportion_ci_wilson(510/610,0.01,610)

proportion_ci_asym(110/160,0.01,160)
proportion_ci_wilson(110/160,0.01,160)
```

```
[3]: (0.7974548969040598, 0.8746762506369239)

[3]: (0.7983705221403479, 0.8769091055402819)

[3]: (0.5931116378126502, 0.7818883621873498)

[3]: (0.6039121674361304, 0.7941905124430988)
```

(2) 用卡方检验分析两种药物疗效是否有差异；（1分）

```
[6]: stats.chisquare([510,100],[110,50],ddof=0)
```

```
[6]: Power_divergenceResult(statistic=1504.5454545454545, pvalue=0.0)
```

p-value<0.05，说明两种药物疗效存在差异。

(3)用z-检验比较两种药物的治愈率差异， 并与置信区间方法比较。（1分）

```
[9]: proportion.proportions_ztest([100,50],[510,110])
```

```
[9]: (-5.741004111057862, 9.411680590755424e-09)
```

HW-U4-3:用卡方检验分析一个4列3行的RC联立表（contingency table)，其卡方检验的自由度是多少？（0.5分） $dof = (4-1)\times(3-1) = 6$

HW-U4-4:卡方检验中，如果自由度为2，卡方统计量为8.1的时候，对应的p值是多少？（1分）

```
[10]: stats.chi2.sf(8,df=2)
```

```
[10]: 0.018315638888734182
```

对应的p值是0.018

HW-U4-5: 如果要研究健康教育是否会让人公众提高防疫意识，从而更加注重勤洗手/戴口罩，采用什么统计方法合适？（0.5分） 可以调查若干不同教育水平的勤洗手/戴口罩频率得到RC联立表，并采用卡方检验分析一个RC联立表的方法。

HW-U4-6: 分析全国34个省、自治区、直辖市、特别行政区的新冠病人确诊数是否符合正态分布，采用卡方检验的话，对应的自由度是多少，说出理由？（0.5分） 对应的自由度是$34-3=31$。

1.1 附加题

1. 请用bootstrapping方法计算单个比例样本(p0,n0)的置信区间，并与demo中asymptotic，wilson score 方法比较，然后改变p0,n0,给出观察结论（0.5分）。

[18]:
```
def proportion_ci_bootstrap(p,alpha,n,n_boot=200):
    bootstrap_means=[]
    num=round(n*p)
    data=[1 for i in range(num)]+[0 for i in range(n-num)]
    for i in range(n_boot):
        random_sample=np.random.choice(data,len(data),replace=True)
        bootstrap_means.append(np.array(random_sample).sum()/len(data)-p)
    ci_l,ci_h=p+ np.percentile(bootstrap_means, [(1-alpha)/2*100,(1+alpha)/
 ↪2*100])
    return ci_l,ci_h

proportion_ci_asym(510/610,0.05,610)
proportion_ci_wilson(510/610,0.05,610)
proportion_ci_bootstrap(510/610,0.05,610)
```

[18]:
```
0.8360655737704918
```

[18]:
```
(0.8066864778914842, 0.8654446696494995)
```

[18]:
```
(0.8068562899634564, 0.8670355254537873)
```

[18]:
```
(0.8360655737704918, 0.8360655737704918)
```

[18]:
```
0.01639344262295082
```

[18]:
```
(0.006316495870918962, 0.02647038937498268)
```

[18]:
```
(0.011480874567673587, 0.03360064248304972)
```

[18]:
```
(0.014754098360655738, 0.01639344262295082)
```

[20]:
```
#  p0
proportion_ci_bootstrap(310/610,0.05,610)
proportion_ci_bootstrap(110/610,0.05,610)
proportion_ci_bootstrap(10/610,0.05,610)
```

[20]:
```
(0.5114754098360655, 0.5138934426229508)
```

[20]:
```
(0.1819672131147541, 0.18360655737704917)
```

```
[20]: 0.01639344262295082
```

```
[20]: (0.014754098360655738, 0.01639344262295082)
```

```
[21]: #  n0
      proportion_ci_bootstrap(10/610,0.05,610)
      proportion_ci_bootstrap(10/610,0.05,310)
      proportion_ci_bootstrap(10/610,0.05,110)
```

```
[21]: (0.014754098360655738, 0.015532786885245916)
```

```
[21]: (0.016129032258064516, 0.016129032258064516)
```

```
[21]: (0.01818181818181818, 0.01818181818181818)
```

结论:

1.可以看到bootstrap方法的置信区间范围要明显小于asym和wilson方法。

2.当p0增大时，bootstrap的置信区间范围减小。

3.由于重采样的次数n_boots相同，n0越小，置信区间范围越小。

2. 用bootstrapping方法计算两个样本（p1=r1/n1，p2=r2/n2)比例（p1,p2)差异的置信区间（0.5分）；并与z-分布方法进行比较（0.5分）

```
[26]: def proportion_ci_bootstrap2(p1,n1,p2,n2,alpha,n_boot=200):
          bootstrap_means=[]
          num1=round(n1*p1)
          num2=round(n2*p2)
          data1=[1 for i in range(num1)]+[0 for i in range(n1-num1)]
          data2=[1 for i in range(num2)]+[0 for i in range(n2-num2)]
          length=n1+n2
          for i in range(n_boot):
              random_sample1=np.random.choice(data1,length,replace=True)
              random_sample2=np.random.choice(data2,length,replace=True)
              random_sample=random_sample2-random_sample1
              bootstrap_means.append(np.array(random_sample).sum()/length-(p2-p1))
          ci_l,ci_h=p2-p1+ np.percentile(bootstrap_means, [(1-alpha)/2*100,(1+alpha)/
       ↪2*100])
          return ci_l,ci_h

      proportion_ci_bootstrap2(510/610,610,110/160,160,0.05)
      proportion.proportions_ztest([100,50],[510,110])
```

```
[26]: (-0.14805194805194805, -0.14415584415584415)
```

```
[26]: (-5.741004111057862, 9.411680590755424e-09)
```

同样，bootstrap方法的置信区间宽度要显著低于z-test方法。