112-2 生物統計學一

列連表分析

2024/05/14

助教: 廖振博

列連表(Contingency table)

- 適用情境：兩個類別變項的推論，以下方法可用來進行列連表分析：
  - ✓ 獨立樣本：卡方檢定，費雪精確檢定
  - ✓ 配對樣本：McNemar’s test
- 列連表：以r列×c欄的表格，將兩個類別變數分組呈現

性別

		男	女
體重	體重過輕	10	20
	正常體重	30	25
	體重過重	20	10

卡方檢定(Chi–squared test)

- 適用情境：檢定兩個類別變項是否獨立(沒有關聯性)
- 假說　H_0: A與B彼此獨立/A與B無關　vs. H_1: A與B彼此不獨立/A與B有關
- 假設A為性別, B為體重：$P(A=a,B=b)\overset{H_0}{=}P(A=a)\times P(B=b)$

	男	女	Total
體重過輕	10	20	30
正常體重	30	25	55
體重過重	$\boxed{20}$	10	30
Total	60	55	115

觀察值 = 期望值？

$$\frac{20}{115}=\frac{60}{115}\times\frac{30}{115}$$

- 在虛無假設下，性別與體重彼此獨立，例如：

$$P(A=\text{男},B=\text{體重過重})=P(A=\text{男})\times P(B=\text{體重過重})$$

卡方檢定(Chi–squared test)

- 假說　H_0: A與B彼此獨立/A與B無關　vs.　H_1: A與B彼此不獨立/A與B有關
- 若虛無假說成立，每一格的觀察值應該要等於期望值$O_{ab} \approx E_{ab}$
- $E_{ab} = n \times P(A = a, B = b) = n \times P(A = a) \times P(B = b)$
- 檢定統計量：$\chi^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{df}, df = (a-1)(b-1)$

  (觀察值與期望值相差過大時應拒絕虛無假設)

- 前提假設：
  - ✓ 隨機樣本(且為類別變數)
  - ✓ 每組樣本之間彼此獨立
  - ✓ 虛無假設下每格的期望次數必須滿足$E_{ab} \geq 5$ (rule of five)

卡方檢定(Chi–squared test)

```
> chisq.test(table(allpass_new, studata1$Gender), correct = F)

	Pearson's Chi-squared test

data:  table(allpass_new, studata1$Gender)
X-squared = 1.1102, df = 1, p-value = 0.292
```

```
> table(allpass_new, studata1$Gender)

allpass_new  0  1
          0  5  7
          1 23 16
```

- 若自由度為1時，可以做葉式校正(Yate’s correction)得到較保守的結果

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E}$$

```
> chisq.test(table(allpass_new, studata1$Gender), correct = T)

	Pearson's Chi-squared test with Yates' continuity correction

data:  table(allpass_new, studata1$Gender)
X-squared = 0.52122, df = 1, p-value = 0.4703
```

費雪精確檢定(Fisher exact test)

- 適用情境：檢定兩個類別變項是否獨立(沒有關聯性)

(在有期望次數小於5的情形下，使用fisher exact test可得檢定力較高的結果)

猜測先放的飲料

實際先放的飲料

	牛奶	茶	Total
牛奶	a	b	$a+b$
茶	c	d	$c+d$
Total	$a+c$	$b+d$	n

- 在虛無假設下，利用超幾何分佈(hypergeometric distribution)計算p–value

$$P(X=a)=\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \Rightarrow p-value = P(X \geq a)$$

費雪精確檢定(Fisher exact test)

```
> table(allpass_new, Gender)
           Gender
allpass_new  0  1
          0  5  7
          1 23 16

> fisher.test(table(allpass_new, Gender))

	Fisher's Exact Test for Count Data

data:  table(allpass_new, Gender)
p-value = 0.3361
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1054431 2.2274338
sample estimates:
odds ratio 
 0.5038965 
```

練習

請先匯入Studata.csv，並命名為studata1，並進行以下檢定：

(假設本學期生統一修課同學是一組由台大學生抽出的隨機樣本)

- 計算每個人的BMI，並命名為bmi新增到studata1中
- 若以不同體態滿意度 (satisfy_body) 分組，請問各組同學的平均 bmi 是否相同
- 承上，如果有不同，請問是哪幾組不同

(補充) McNemar's test

- 適用情境：檢定兩個類別變項是否獨立(沒有關聯性)

	Test1	Test2
A	Pass	No
B	Pass	Pass
C	No	Pass
D	No	Pass
E	No	No

	Test1	Test2	Total
Pass	2	3	5
No	3	2	5
Total	5	5	10

只有5個人的資料，列連表卻有10人
(違反每個資料點須為獨立的假設)

Test2

Test1		Pass	No	Total
	Pass	1 (B)	1 (A)	2
	No	2 (C, D)	1 (E)	3
	Total	3	2	5

- 轉換成此列連表後的資料點互相獨立
- 利用配對情形來檢定不同考試的通過率是否相同
- 一致配對(concordant pair)
- 不一致配對(discordant pair)

(補充) McNemar's test

- 適用情境: 檢定兩個類別變項是否獨立(沒有關聯性)

- 假說:

  H_0: 通過與否跟哪一次考試無關 (兩次考試的通過率相同)

  H_1: 通過與否跟哪一次考試有關 (兩次考試的通過率不同)

Test2

Test1		Pass	No	Total
	Pass		O_1	2
	No	O_2		3
	Total	3	2	5

- 兩次考試結果不一致的格子的期望值:

$$E_1 = E_2 = \frac{O_1 + O_2}{2}$$

- 檢定統計量:

$$\chi^2 = \frac{(|O_1 - O_2| - 1)^2}{O_1 + O_2} \overset{H_0}{\sim} \chi^2_{df=1}$$

- 前提假設:

✓ 隨機樣本(且為類別變數)

✓ 每組樣本之間彼此不獨立

✓ 虛無假設下每格的期望次數必須滿足$E_{ab} \geq 5$ (rule of five)