# 第十三章卡方檢定(Chi-Square Test)

## 13.1適合度檢定(Goodness of Fit Tests)

- 檢定骰子是否公正
- 檢定一組樣本資料是否來自某一分配
- 檢定樣本資料是否隨機

分析如下：

1. 某一實驗結果可分為k個互斥的事件，分別為 $A_1, A_2, \ldots A_k$

2. 重複該實驗n次, 令 $P_i = P(A_i)$, $i=1, 2, \ldots, k$, 表示事件 $A_i$ 出現的機率

3. 令 $O_i$ 表示事件 $A_i$ 在實驗中出現的次數, $i=1, 2, \ldots, k$, 其中大寫 $O_i$ 表示隨機變數，小寫 $o_i$ 表示中觀察次數

4. $(O_1, O_2, \ldots, O_{k-1})$ 構成多項式分配具有參數 $P_1, P_2, \ldots, P_{k-1}$ ，其中 $\sum_{i=1}^{k} O_i = n$ ，$\sum_{i=1}^{k} P_i = 1$

當k=2

$O_1 \sim B(n, P_1)$

且 $E(O_1) = nP_1$ , $V(O_1) = nP_1(1-P_1)$

如果n≥30 由中央極限定理

$$\frac{O_1 - E(O_1)}{\sqrt{V(O_1)}} = \frac{O_1 - nP_1}{\sqrt{nP_1(1-P_1)}} \overset{\text{近似}}{\sim} N(0,1)$$

令 $$x^2 = \left(\frac{O_1 - nP_1}{\sqrt{nP_1(1-P_1)}}\right)^2 = \frac{(O_1 - nP_1)^2}{nP_1(1-P_1)}$$

$$= \frac{(O_1 - nP_1)^2}{nP_1} + \frac{(O_1 - nP_1)^2}{n(1-P_1)}$$

$$= \frac{(O_1 - nP_1)^2}{nP_1} + \frac{(O_2 - nP_2)^2}{nP_2} \overset{\text{近似}}{\sim} x^2(1)$$

# 推廣至一般式

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - nP_i)^2}{nP_i} \overset{\text{近似}}{\sim} x^2(k-1)$$

探討如下適合度檢定

Case1：以多項實驗所產生的資料來檢定母體機率是否為虛無假設所設定的特定值

$H_0$：$P_i = P_{io}$，i=1,2,...,k

$H_1$：至少一等號不成立

在$H_0$為真之下

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - nP_i)^2}{nP_i} = \sum_{i=1}^{k} \frac{(O_i - nP_{io})^2}{nP_{io}} \overset{\text{近似}}{\sim} x^2(k-1)$$

決策法則

拒絕$H_0$ if $x^2 = \sum_{i=1}^{k} \frac{(o_i - nP_{io})^2}{nP_{io}} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} > x_\alpha{}^2(k-1)$

其中 $o_i$ 表示事件 $A_i$ 觀察次數

其中$e_i = nP_{io}$ 表示事件 $A_i$ 期望次數

Ex:擲一骰子240次，1點至6點觀察次數分別為30,50,44,42,46,28，試檢定該骰子是否公正?

Sol：

$H_0$：骰子是公正

$H_1$：骰子不公正

令 $P_i$ 表是出現i點的機率，則上述假設檢定可改寫為

$H_0: P_i = \frac{1}{6}$ , i=1,2,...,6(k=6)

$H_1$ :至少一等號不成立

決策法則

拒絕 $H_0$ if $\sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} > x_\alpha^2(k-1)$

| 點數 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $O_i$ | 30 | 50 | 44 | 42 | 46 | 28 |
| $e_i$ | 40 | 40 | 40 | 40 | 40 | 40 |
| $O_i - e_i$ | -10 | 10 | 4 | 2 | 6 | -12 |
| $(O_i - e_{ii})^2$ | 100 | 100 | 16 | 4 | 36 | 144 |

$$\sum_{i=1}^{6}\frac{(O_i-e_i)^2}{e_i}=\frac{100}{40}+\frac{100}{40}+\frac{16}{40}+\frac{4}{40}+\frac{36}{40}+\frac{144}{40}$$

$$=\frac{400}{40}=10$$

當 $\alpha=0.05$ ⇨ $x^2_{0.05}(5)=11.07$

從 $\sum_{i=1}^{6}\frac{(O_i-e_i)^2}{e_i}<x^2_{0.05}(5)$

$10$ $11.07$

結論在 $\alpha=0.05$ 之下，不拒絕 $H_0$

即由實驗結果顯示無法推翻骰子是公正

當 $\alpha = 0.1$ ⇨ $x^2_{0.1}(5) = 9.236$

從 $$\sum_{i=1}^{6} \frac{(O_i - e_i)^2}{e_i} > x^2_{0.1}(5)$$

$$\| \qquad\qquad \|$$

$$10 \qquad\qquad 9.236$$

結論在 $\alpha = 0.1$ 之下，拒絕 $H_0$

即由實驗結果顯示骰子不公正

Case2:檢定樣本資料是否來自某一特定分配

a.特定分配為離散型分配

步驟1.如果欲檢定分配的母體參數未知，須用樣本資料估計之，假如估計r個參數，則自由度再減r

2. 由於欲檢定的分配為離散型分配，必須列出每個可能值的機率 $P_i$，及計算期望值 $e_i = nP_i$，如果 $e_i < 5$，則與鄰組合併，直到 $e_i \geq 5$

3. 計算每個可能值出現的次數 $O_i$

4. 計算 $x^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$

5. 如果 $x^2 > x_\alpha^2(k-1-r)$，則拒絕 $H_0$，否則不拒絕 $H_0$

# Ex. 13.2

記錄過去90週，每週下雨的天數，資料如下

| 每週下雨天數 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 週數 | 6 | 12 | 27 | 18 | 21 | 6 | 0 | 0 |

X表示一週內下雨的天數,在顯著水準 $\alpha$=0.05 之下，試檢定X是否服從二項分配?

Sol： $H_0$ : X~B(7,P)

$H_1$ : X ~ B(7,P)

由於P未知，以 $\hat{P}$ 估計之

$$\hat{P} = \frac{0\times 6 + 1\times 12 + \cdots + 7\times 0}{90\times 7} = 0.37$$

$$\therefore P(X=x) = C_x^7 0.37^x (1-0.37)^{7-x}$$

$$= C_x^7 0.37^x 0.63^{7-x}, x = 0,1,2,\ldots,7$$

| x | $O_i$ | $P_i = P(X = i)$ | $e_i = nP_i$ |
|---|---|---|---|
| 0 | 6 | 0.0394 | 3.54 |
| 1 | 12 | 0.1619 | 14.58 |
| 2 | 27 | 0.2853 | 25.68 |
| 3 | 18 | 0.2793 | 25.14 |
| 4 | 21 | 0.1640 | 14.76 |
| 5 | 6 | 0.0578 | 5.19 |
| 6 | 0 | 0.0113 | 1.02 |
| 7 | 0 | 0.001 | 0.09 |

| x | $O_i$ | $P_i$ | $e_i$ |
|---|---|---|---|
| 0~1 | 18 | 0.2013 | 18.21 |
| 2 | 27 | 0.2853 | 25.68 |
| 3 | 18 | 0.2793 | 25.41 |
| 4 | 21 | 0.1640 | 14.76 |
| 5~7 | 6 | 0.0701 | 6.3 |

決策法則

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} > x_\alpha{}^2(k-1-r)$$

$$x^2 = \frac{(18-18.21)^2}{18.21} + \frac{(27-25.68)^2}{25.68} + \cdots + \frac{(6-6.3)^2}{6.3} = 4.75$$

$$x_\alpha{}^2(k-1-r) = x^2_{0.05}(5-1-1) = x^2_{0.05}(3) = 7.815$$

從 $x^2 < x_\alpha{}^2(k-1-r)$

$4.75 \quad 7.815$

結論：在$\alpha$=0.05之下不拒絕$H_0$，即無法推翻一週內下雨的天數來自二項分配

b.特定分配為連續型分配

常用方法
- (1)等機率分組法(固定$e_i$)
- (2)等組距分組法(固定$o_i$)

考慮等機率分組法

步驟1.如果欲檢定分配的母體參數未知，須用樣本資料估計之，假如估計r個參數，則自由度再減r

2.將欲檢定的分配k等分(k=4~10)，每等分之機率為 $\frac{1}{k}$

3.利用等機率求出每組的端點

4.計算每組的觀察次數 $O_i$

5.計算 $x^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$

6.當 $x^2 > x_\alpha^2(k-1-r)$ ，則拒絕 $H_0$ ，否則不拒絕 $H_0$

# Ex. 13.4

假設隨機變數X為由40頭母牛每個所生產出來的牛奶總量(公升)，資料如下

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 18.79 | 14.62 | 13.98 | 15.79 | 12.39 | 13.20 | 16.08 | 13.79 | 16.16 |
| 16.12 | 17.81 | 18.74 | 15.79 | 13.32 | 13.63 | 16.40 | 13.76 | 16.58 |
| 15.25 | 18.97 | 18.36 | 15.04 | 18.79 | 18.08 | 17.32 | 16.32 | 17.54 |
| 18.08 | 14.20 | 18.04 | 13.00 | 13.25 | 12.43 | 16.56 | 14.12 | 20.55 |
| 16.75 | 13.29 | 18.23 | 16.93 | | | | | |

試問在顯著水準為 $\alpha$=0.01之下，X是否服從常態分配?

# 13.3獨立性檢定

- 檢定一個母體的兩個分類屬性是否獨立

例如：薪資的高低和學歷的程度是否獨立？

婚姻狀況與宗教信仰是否有關？

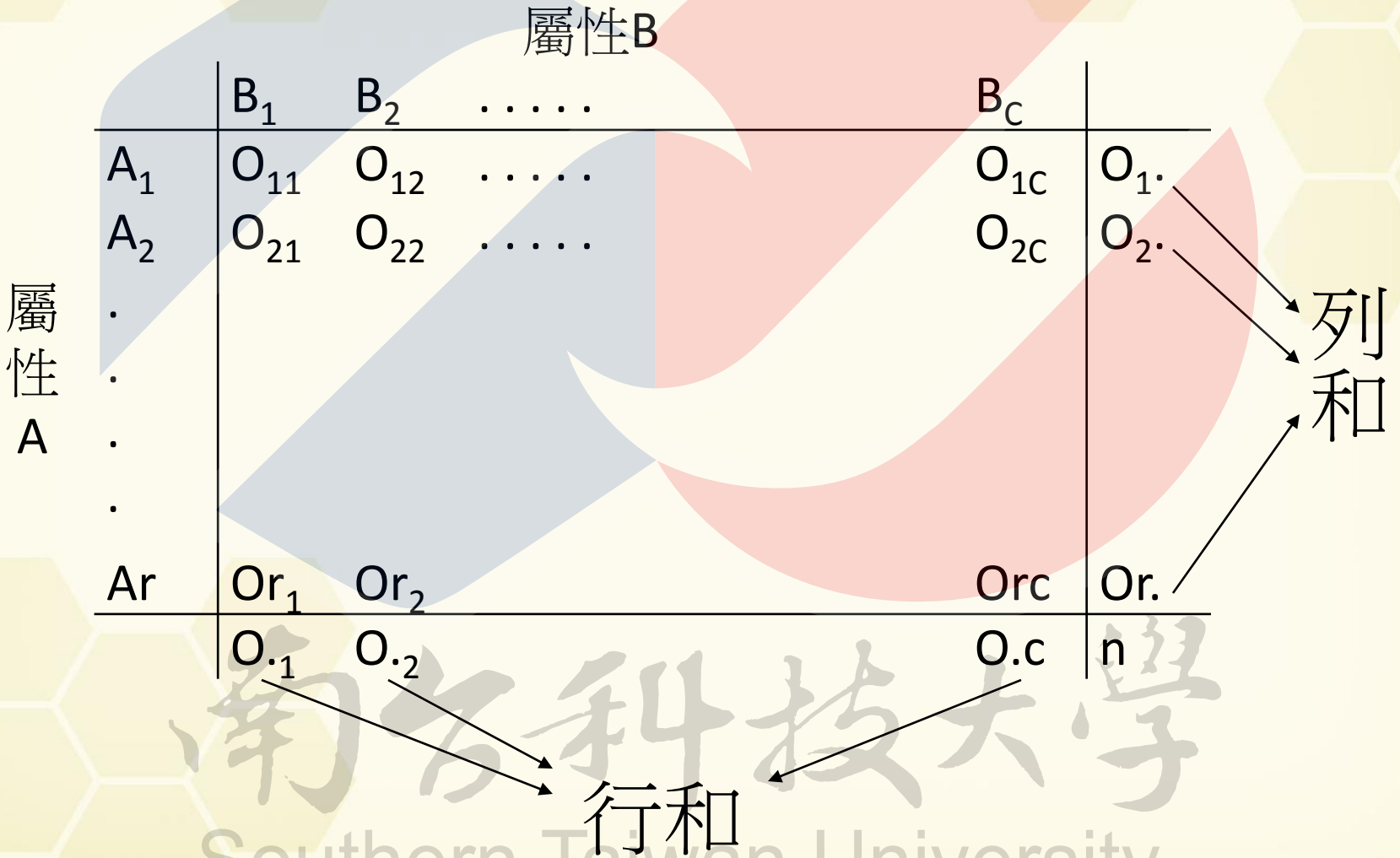
- 分析如下：

自一研究母體抽取n筆資料，按屬性A、B加以分類，

其中屬性A分成r類互斥事件

屬性B分成c類互斥事件

- 則列聯表表示如下：




上述稱為rxc列聯表

令

- Oij表示同時具有屬性A第i類及屬性B第j類的個數
- Pij表示同時具有屬性A第i類及屬性B第j類的機率
- Oi表示屬性A第i類的個數
- Oj表示屬性B第j類的個數
- Pi表示屬性A第i類的機率
- Pj表示屬性B第j類的機率
- $Pij = P(Ai \cap Bj), Pi = P(Ai), Pj = P(Bj)$

- 給定顯著水準α下，試檢定兩屬性A 、B是否獨立性

  $H_0$ ：兩屬性獨立

  $H_1$ ：兩屬性相關

- Sol ：上述的假設檢定可改寫為

  $H_0$ ：$P(A_i \cap B_j) = P(A_i)P(B_j)$，

  $i = 1,2,\ldots,r$，$j = 1,2,\ldots,c$

  $H_1$ ：至少一等號不成立

- 即 $H_0$ ：$P_{ij} = P_i P_j$，$i = 1,2,\ldots,r$，$j = 1,2,\ldots,c$

  $H_1$ ：至少一等號不成立

case1： Pi ,Pj皆已知i＝1,2,…,r，j＝ 1,2,…,c

在 $H_0$為真之下

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-nP_{ij})^2}{nP_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-nP_iP_j)^2}{nP_iP_j} \text{近似} \; X^2(rc-1)$$

拒絕$H_0$ if $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-nP_iP_j)^2}{nP_iP_j} > X^2_{\alpha}(rc-1)$

case2： Pi ,Pj皆未知 i＝1,2,…,r，j＝ 1,2,…,c

由於Pi ,Pj未知,須估計之

令 $\hat{P}_i = \frac{O_i}{n} \xrightarrow{\text{估計}} P_i$ , i＝1,2,…,r

$\hat{P}_j = \frac{O_j}{n} \xrightarrow{\text{估計}} P_j$ , j＝ 1,2,…,c

在 $H_0$為真之下

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - nP_{ij})^2}{nP_{ij}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n\hat{P}_i\hat{P}_j)^2}{n\hat{P}_i\hat{P}_j} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n\frac{O_i}{n}\frac{O_j}{n})^2}{n\frac{O_i}{n}\frac{O_j}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \frac{O_iO_j}{n})^2}{\frac{O_iO_j}{n}}$$

$$\sim X^2((r-1)(c-1))$$

- 拒絕$H_0$ if

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - \frac{o_i o_j}{n})^2}{\frac{o_i o_j}{n}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} > X_{\alpha}^2((r-1)(c-1))$$

- 其中

$$e_{ij} = \frac{o_i o_j}{n} = \frac{\text{第 i 列和} \times \text{第 j 列和}}{\text{樣本數}}$$

- Yates'修正法

在2 ×2列聯表中檢定統計量 $x^2$ 有時會做出修正

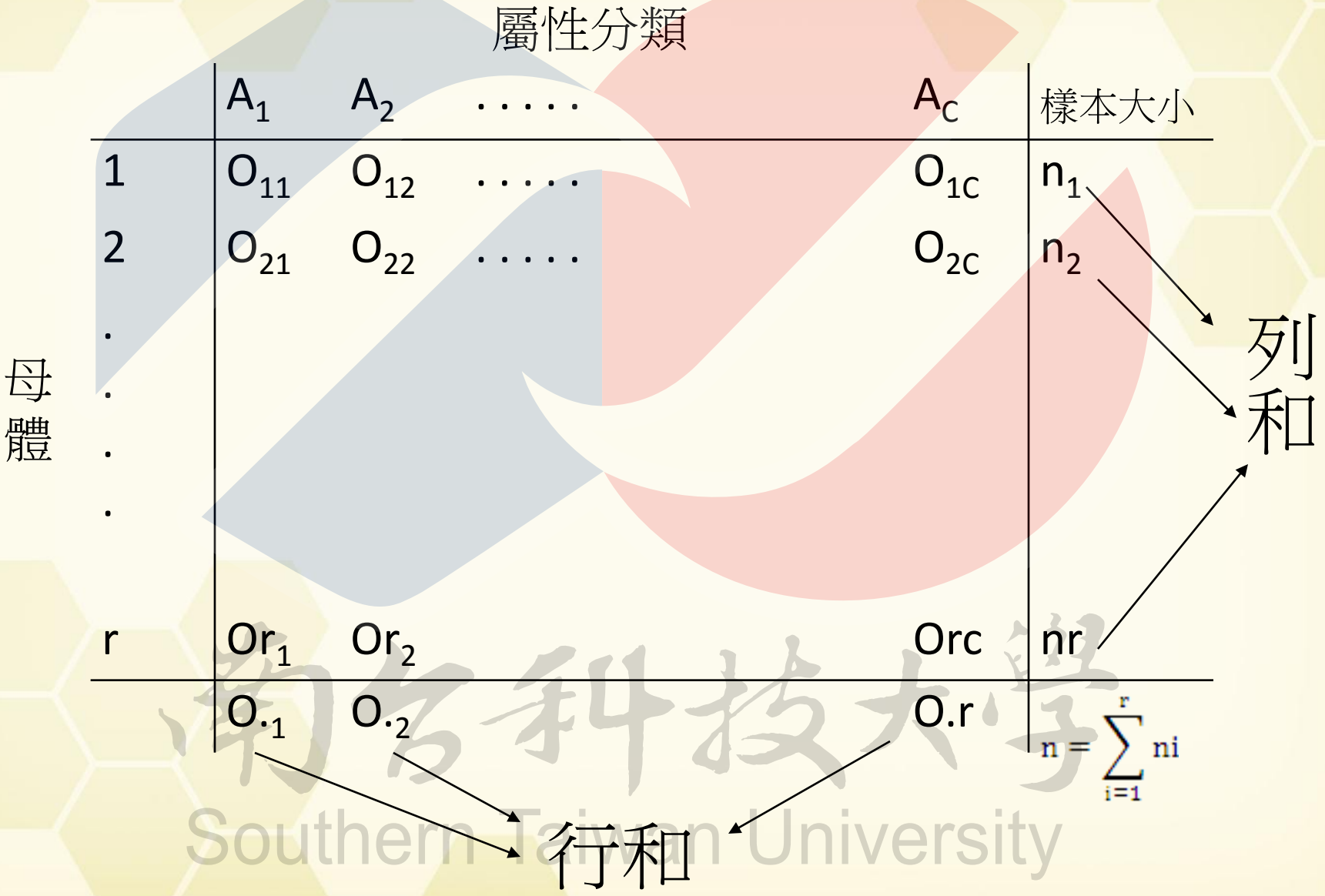
$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|O_{ij} - e_{ij}| - 0.5)^2}{e_{ij}} \text{近似} \; x^2(1)$$

拒絕$H_0$ if

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}} > X^2_{\alpha}(1)$$

# 13.4齊一性檢定

- 探討r個( $r \geq 2$)母體的某一屬性(教學方法、施政滿意度、分配....)是否一致？
- 分析如下：
- 將所要探討的屬性分成C類互斥事件，依次為A1,A2,....,$A_C$ ，自第 i 個母體抽取$n_i$個樣本，i ＝1,2,...,r ，則列聯表如下：



- 令 $P_{ij} = P_i(A_j)$ 表示第i個母體的屬性 $A_j$ 事件發生的機率 i＝1,2,…,r，j＝ 1,2,…,c
- 給定顯著水準 α ，齊一性假設檢定如下

$H_0$ ：

$$P_{11} = P_{21} = \cdots = P_{r1} = P_1$$
$$P_{12} = P_{22} = \cdots = P_{r2} = P_2 \Rightarrow \text{r個母體比例相同}$$
$$\vdots$$
$$P_{1c} = P_{2c} = \cdots = P_{rc} = P_c$$

$H_1$ ：至少有一等號不成立

case1： Pj皆已知 j＝ 1,2,…,c

在 $H_0$ 為真之下

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-nP_{ij})^2}{nP_{ij}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-n_iP_j)^2}{n_iP_j} \text{近似} X^2(r(c-1))$$

拒絕$H_0$ if $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-n_iP_j)^2}{n_iP_j} > X_\alpha^2(r(c-1))$

case2： Pj皆未知 ，j＝ 1,2,…,c

令 $\hat{P}_j = \frac{O_j}{n} \xrightarrow{\text{估計}} P_j$ , j＝ 1,2,…,c

在 $H_0$為真之下

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n_i P_{ij})^2}{n_i P_{ij}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n_i \hat{P}_j)^2}{n_i \hat{P}_j}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(O_{ij} - n_i \frac{O_j}{n}\right)^2}{n_i \frac{O_j}{n}} \text{近似} \chi^2((r-1)(c-1))$$

拒絕$H_0$ if

$$X^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij}-n_i\frac{o_j}{n}\right)^2}{n_i\frac{o_j}{n}}=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij}-e_{ij}\right)^2}{e_{ij}}>X_{\alpha}^{2}((r-1)(c-1))$$

其中 $e_{ij}=\dfrac{\text{第 i 個母體的樣本}\times\text{第 j 個屬性和}}{\text{總樣本數}}$

- 獨立性檢定與齊一性檢定比較

| 獨立性檢定 | 齊一性檢定 |
| --- | --- |
| 列聯表檢定問題 | 列聯表檢定問題 |
| 針對單一母體進行抽樣 | 針對多個母體進行抽樣 |
| 列總和及行總合是隨機變數 | 各母體的樣本數是事先決定好的，屬性和是隨機變數 |
| 檢定統計量 $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ | 檢定統計量 $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ |
| $e_{ij} = \frac{\text{第 i 列和} \times \text{第 j 行和}}{\text{樣本數總和}}$ | $e_{ij} = \frac{\text{第 i 列和} \times \text{第 j 行和}}{\text{樣本數總和}}$ |