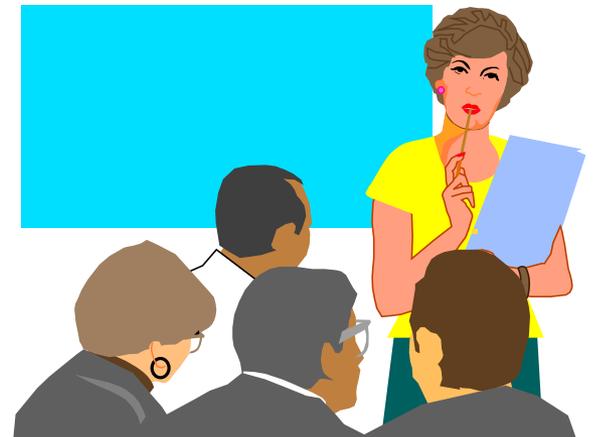


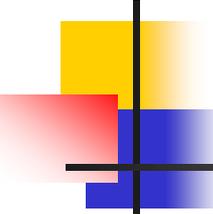
# 卡方檢定補充講義

授課教師：統計系余清祥

主題：卡方檢定

<http://csyue.nccu.edu.tw>





# 卡方檢定(Chi-Square Test)

---

- 卡方檢定可用於處理類別資料，常用於以下三種檢定：

## (1) Goodness of Fit Test (適合度檢定)

→ 通常用來檢查資料是否為某一特定分配，  
例如：樣本是否與母體類似。

## (2) Tests of Independence (獨立性檢定)

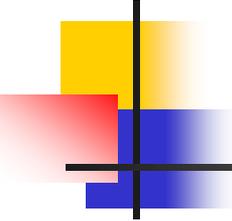
## (3) Tests of Homogeneity(齊一性檢定)



## 卡方檢定(續)

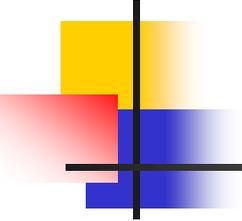
---

- 卡方檢定是處理問卷資料的利器，可用來檢查樣本資料和母體是否類似，以避免瞎子摸象的窘境。
  - 比較母體與樣本的結構差異。
- 卡方檢定也可用於檢查變數間是否相關，像是不同特質的受訪者，是否對某個問題的想法有不同反應。



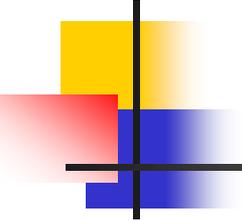
e.g. (a)調查台北對今年治安的看法，如下表：

看法	人數
比去年好	245
差不多	126
比去年差	713



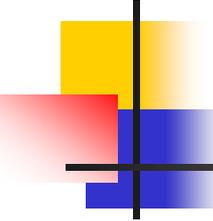
e.g. (b) 調查大學生中性別與選系的關係，得出

學院		性別			
		商	工	藝術	公衛
性別	男	21	16	145	8
	女	14	4	175	17



e.g. (c)調查年紀與買車習性的關係

年齡	大型車	中型車	小型車
20~39	90	18	92
40~59	40	60	100

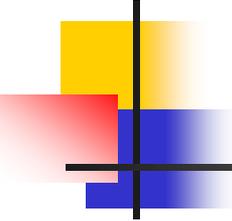


## 適合度檢定：

通常用來檢定母體是否為某一特定分配。

e.g. (d) 投擲一硬幣500次，得出230個正面，  
欲檢定此硬幣為公平硬幣

	正面	反面
觀察值	230	270
理論值	250	250

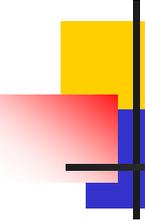


---

(i) Z 檢定:  $P(|Z| \geq \frac{|230 - 250|}{\sqrt{500 \times \frac{1}{2} \times \frac{1}{2}}} = 1.789) \cong 0.0736$

(ii)  $\chi^2$  檢定:  $\chi^2 = \sum_{i=1}^2 \frac{(o_i - e_i)^2}{e_i} = \frac{400}{250} + \frac{400}{250} = 3.2$

其中  $o_i, e_i$  為第  $i$  個值的觀察值與理論值。查表  
得出  $\chi_{0.10}^2(1) = 2.706, \chi_{0.05}^2(1) = 3.841$ 。



e.g.(e) 調查400個家中有兩個子女的家庭以研究男女嬰出生的機會是否相同。

	男男	男女	女男	女女
觀察值	92	94	110	104
理論值	100	100	100	100

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \frac{64}{100} + \frac{36}{100} + \frac{100}{100} + \frac{16}{100} = 2.16 < \chi_{0.05}^2(3) = 7.815$$

⇒  $\alpha=0.05$  下，我們認為男女嬰出生機會相同。

e.g.(f)統計歷年台灣核能電廠的出事率，以確定其發出次數是否為平均每年三次的布阿松分配（假設資料，共30年資料）

每年次數	0~1次	2次	3次	4次	5次以上
觀察值	3	5	10	8	4
理論值	5.97	6.72	6.72	5.04	5.55

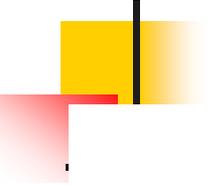
- 本例中的母體比例值分別是：

$$P(0) + P(1) = e^{-3} + 3e^{-3} = 0.1991 \Rightarrow \text{理論數} = 30 \times 0.1991 = 5.97$$

$$P(2) = \frac{3^2}{2} e^{-3} = 0.2240 \Rightarrow \text{理論數} = 30 \times 0.2240 = 6.72$$

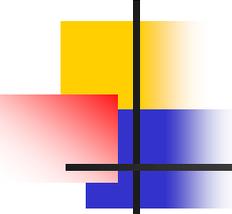
∴ ∴ ∴

( 因為  $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$  。 )


$$\chi^2 = \sum_{i=1}^5 \frac{(o_i - e_i)^2}{e_i} \cong 5.69 < \chi_{0.05}^2(4) = 9.488$$

因此我們不拒絕核能電廠每年出事次數為 Poisson(3)的假設。

[註]： $\chi$ 檢定要求每個理論值至少為5，以避免分母太小產生誤判。

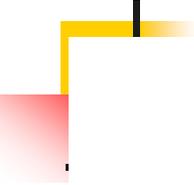


## 獨立性檢定：

用來檢定母體中的各項特性間是否互相影響。

e.g. (b)

學院 性別	商	工	藝術	公衛	列合計
男	21	16	145	8	190
女	14	4	175	17	210
行合計	35	20	320	25	400


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

其中  $o_{ij}$  與為第  $i$  列第  $j$  行觀察值與理論值，  
 $i = 1, \dots, r$ ,  $j = 1, \dots, c$ , 而自由度為  $(r-1)(c-1)$ 。

假設檢定  $H_0$ : 行與列的類別互相獨立

$$\longrightarrow H_0: p_{ij} = p_{i.} p_{.j}, i = 1, \dots, r, j = 1, \dots, c$$

其中  $p_{i.}$  為第  $i$  列平均， $p_{.j}$  為第  $j$  行平均，行列互

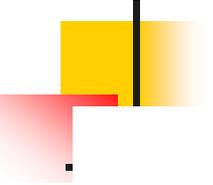
相獨立的理論值  $p_{ij} = p_{i.} \times p_{.j}$

理論值：

	商	工	藝術	公衛	列
男	0.0416 (16.625)	0.02375 (9.5)	0.38 (152)	0.0297 (11.875)	0.475
女	0.0459 (18.375)	0.02625 (10.5)	0.42 (168)	0.0328 (13.125)	0.525
行	0.0875	0.05	0.8	0.0625	1

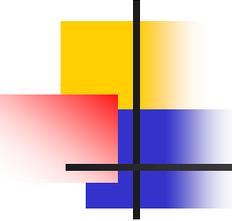
檢定值 =  $13.675 > \chi^2(3) = 7.815$

因此我們認為性別與選系有關聯。



e.g.(g) 隨機抽樣調查200位大學生，獲得其就讀年級與吸煙習慣的資料如下：

年級 \ 習慣	一年級	二年級	三年級	四年級
吸煙	21	33	25	20
不吸煙	47	26	19	9



---

在 $\alpha=0.01$ 下檢定大學生之就讀年級與吸煙習慣是否獨立，可得

$$\text{檢定值} = 15.7438 > \chi^2(3) = 11.34$$

因此我們認為抽煙與就讀的年齡有關聯。

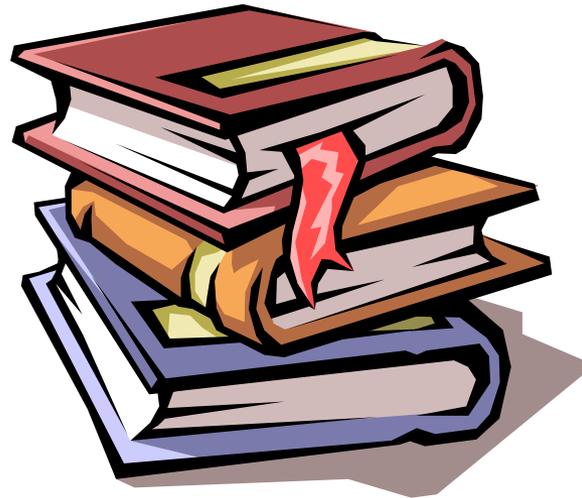
# 樣本代表性



- 樣本代表性意指抽出的樣本，其特性與母體非常類似，足以由樣本來代表整個母體。
- 檢查樣本代表性是分析問卷資料的首要步驟，若樣本與母體差異過大，則以樣本的資訊推測母體的特性，將顯得不合適。

# 分析案例

- 政治大學統計系大學部三年級學生在民國90年，調查政治大學學生對流行性雜誌的偏好，包括雜誌價格、內容、出版期數等問項。



1. 研究對象：國立政治大學全體學生—因為我們身處在政大校園裡面，對於同校同學們，他們對於雜誌的傾向為和是我們所關心的話題。而且目前坊間得雜誌更是千奇百樣，尤其以流行性雜誌一本接一本的發行，更為獨特。所以我們選擇以政大全體學生來做為我們研究的對象，來了解他們對於各類型雜誌及流行性雜誌的偏好情形。

2. 抽樣方法：先分層抽樣在行群集抽樣

本次抽樣調查因我們認為年級以及學院在雜誌的閱讀習慣上會有差別，而同一學院內的系別不會有差異，所以把學院及年級同時做為分層標準，再針對每一學院的比例抽出等比例的樣本數，進行調查。母體即為全政大之學生。其中，抽出各學院的人數比為：

學院	人數	有效份數
文學院	60	52
社會科學院	120	95
商學院	150	126
傳播學院	40	34
法學院	30	21
理學院	30	24
國際事務學院	15	15
外國語文學院	55	45
總計	500	412

$$\text{有效問卷率} = 500 - (\text{無效問卷} + \text{未回收問卷}) / 500 = 0.824$$

\* 無效問卷：答題不完整、單題複選皆列入無效問卷（71份）

\* 未回收問卷：17份

## ※樣本代表性

為了要檢定此份問卷是否可代表母體（全部政大學生），所以我們必須逐項檢查基本問項以確定此份問卷的樣本代表性是否足夠。母體資料來源為「國立政治大學九十學年度第一學期學生註冊人數統計表」，學生總人數共9,883人

檢定方法：Goodness-of-fit test

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)} \quad k=1 \dots i$$

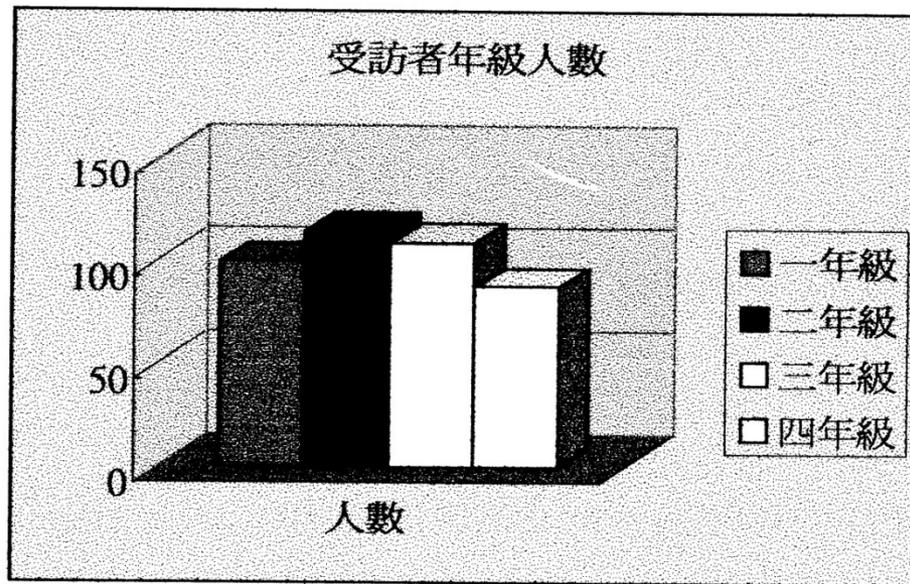
$O_i$ ：樣本觀察次數（實際出現次數）

$E_i$ ：理論次數（期望次數）

$H_0$ ：母體比例與樣本比例相同

$H_1$ ：母體比例與樣本比例不相同

當所得  $\chi^2 < \chi^2_{(k-1)}$ ，即表示樣本代表性足以代表母體



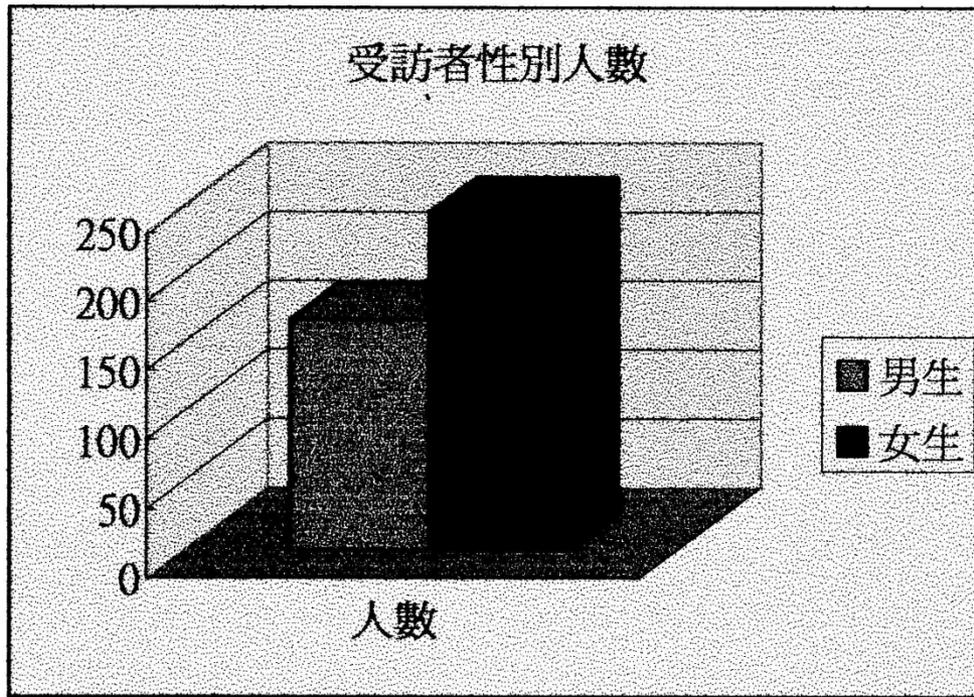
各年級中，以二年級受訪人數所佔比例最高，約有 27.9%，其次為三年級 26.5%。最少則為四年級，只有 21.4%。

	一年級	二年級	三年級	四年級
人數	100	115	109	88

※檢查年級的分布比例

	男	女	總計	百分比	樣本百分比
一年級	904	1470	2374	24.0%	24.3%
二年級	922	1587	2509	25.4%	27.9%
三年級	963	1376	2339	23.7%	26.5%
四年級	1108	1553	2661	26.9%	21.4%

利用 Goodness-of-fit test，得到  $X^2=7.059 > X^2_{3,0.9}=6.25$ ，我們發現在此一部分的檢查，樣本跟母體的比例是有差異的，樣本代表性可能不太足夠。但是由於在母體中的四年級包含了延畢生，所以在人數方面自然會比真正四年級人數高，因此如果在抽樣時能加入這項影響因素的話，整份問卷會更有其代表性。



在受訪者中，男生佔了41%，而女生的比例則為59%。性別比，男：女約為2：3。

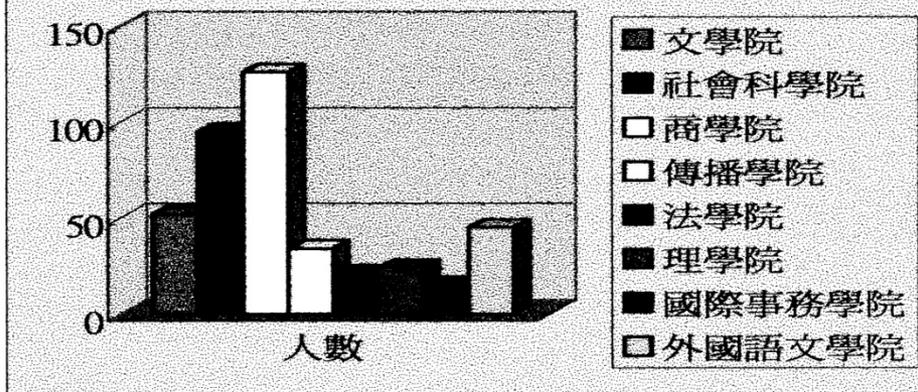
	男生	女生
人數	167	245

### ※檢查男女比例

全政大為母體下，男女比例為40.5：59.5 (3897：5986)；而樣本比例為41：59 (167：245)

由Goodness-of-fit test 檢定的結果， $X^2=0.21024 < X^2_{1,0.9}=2.71$ ，可知男女的採樣方面的小誤差是可接受的，即在男女方面，樣本代表性為足夠的。

受訪者學院人數



受訪者學院以商學院居多，佔 30.8%，其次為社會科學院佔 23.1%，最少的為國際事務學院，佔 3.6%

	文學院	社會科學院	商學院	傳播學院
人數	52	95	126	34
	法學院	理學院	國際事務學院	外國語文學院
人數	21	24	15	45

※檢查學院的分布比例

	男	女	總計	百分比	樣本百分比
文學院	330	856	1186	12.00%	12.4%
社會科學院	975	1369	2344	23.72%	23.1%
商學院	1250	1704	2954	29.89%	30.8%
傳播學院	269	597	866	8.76%	8.3%
法學院	304	305	609	6.16%	5.1%
理學院	422	126	548	5.55%	5.8%
國際事務學院	94	174	268	2.71%	3.6%
外國語文學院	253	855	1108	11.21%	10.9%

利用 Goodness-of-fit test，得到  $X^2=2.5264 < X^2_{7,0.9}=12.02$ ，所以能推之我們的假設「母體比例與樣本比例」是相同的。在學院這一方面可以說樣本對母體的代表性是足夠的。