巨量資料與統計分析

政治大學統計系余清祥

2024年10月29日

第七週：樣本代表性

http://csyue.nccu.edu.tw

抽樣與資料品質

- 除資料品質外，**選取適合資料（包括抽樣）**也要謹慎的考慮。

→統計是由觀察值（或現象）反推出發生原因，如何選取樣本非常重要。

→為了避免「瞎子摸象」及「以偏概全」的問題，檢查***樣本代表性***是資料分析時必須考慮的步驟。(i.e., Representative Sample)

有代表性的樣本 (Representative Sample)

- 抽血、檢體、切片等醫學檢驗，可視為能夠反映身體狀況的代表性樣本。

→蒐集能夠反映母體的樣本，未必是件容易的工作。（問題：**如何抽樣**？）

- 問卷調查常用於「顧客滿意度」、「品質管制」，但有些議題執行時較為棘手。

→例如：空氣（或土地）污染、品牌忠誠度、疾病盛行率。

樣本代表性的實例

- 美國調查業巨人《文摘》(Literary Digest)，1936年發出一千萬問卷，預言共和黨候選人Landon將大獲全勝，擊敗民主黨羅斯福。

→調查對象為擁有汽車與電話的家庭。

- 蓋洛普1948年錯誤預測共和黨杜依會獲勝。

→使用配額抽樣法，訪問五百位家庭主婦、二十位農夫與三百位老人。

註：資料來源【真實的謊言】

樣本代表性

- 「樣本代表性」意指抽出的樣本，其特性與母體類似，足以由樣本代表整個母體。

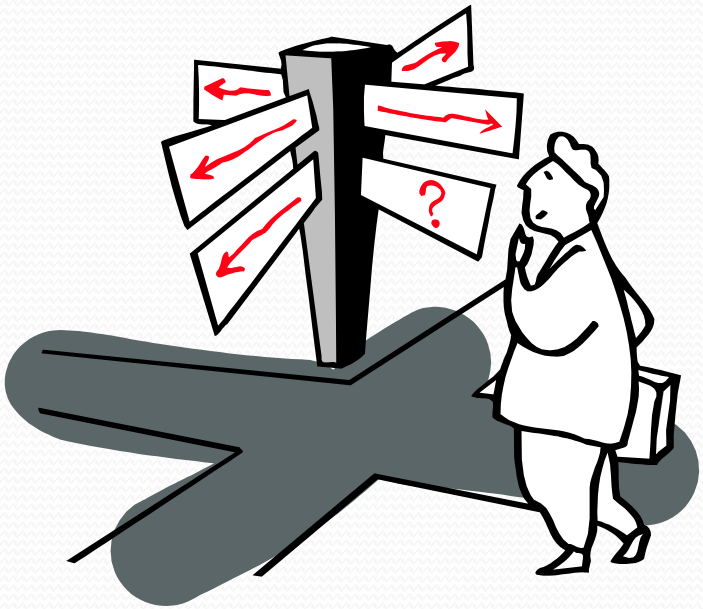
→檢查樣本代表性是資料分析的首要步驟，若樣本與母體差異過大，以樣本推測母體會有疑慮。

https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwiU0ML7iZzlAhXMG6YKHc51DH8QjRx6BAgBEAQ&url=https%3A%2F%2Farticle1000.com%2Ftag%2Fnon-probability-sampling%2F%3Fak_action%3Dreject_mobile&psig=AOvVaw2zFOLIYy03mFK3RMCHqxwp&ust=1571153585084780

常見的樣本代表性檢查項目

- 通常用於檢查樣本代表性的問項：

1.性別比例
2.年齡結構
3.居住地區
4.教育程度
5.職業別
6.婚姻狀態
7.其他因素



母體參考資料

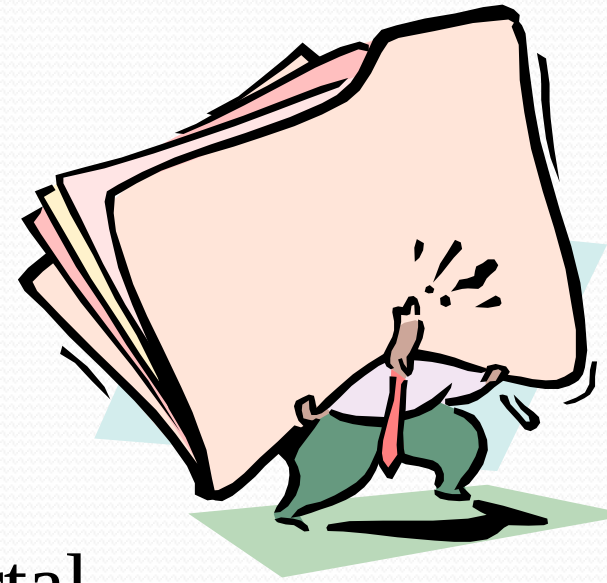
- 台灣地區的人口統計資料可到內政部統計處查詢，網址 https://statis.moi.gov.tw/micst/webMain.aspx?sys=100&funid=defjsp

→內政統計年報

→性別統計資料

→重要參考指標

- 內政部戶政司也有類似資料 https://www.ris.gov.tw/app/portal



樣本代表性的檢定方法

- 檢查樣本代表性的方法：

→卡方檢定(適度性；Goodness-of-fit)

→ 計算公式相似，但解釋方法不同。

$$\text{卡方檢定：}\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

o_i：觀察個數；e_i：理論個數

卡方檢定(Chi-Square Test)

- 卡方檢定用於處理類別資料，功能包括：

→ Goodness of Fit Test (適合度檢定)

檢查資料是否為某一特定分配，例如：樣本是否與母體類似。

→ Tests of Independence (獨立性檢定)

→ Tests of Homogeneity(齊一性檢定)

卡方檢定(續)

■範例一、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。

→由以下計算可知不拒絕男女嬰出生機會相同。

$$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = \frac{64}{100} + \frac{36}{100} + \frac{100}{100} + \frac{16}{100} = 2.16 < \chi^2_{0.05}(3) = 7.815$$

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

卡方檢定(續)

- 範例二、統計歷年台灣核能電廠的出事率，以確定其發生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）。

→不拒絕核能電廠每年出事次數服從Poisson(3)。

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

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

- 樣本代表性中的卡方檢定，理論個數 O_i 等於樣本數乘以母體中的比例值，例如：上例中的理論個數＝400×1/4＝100。
- 下例中的母體比例值分別是：

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

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

$$\vdots \qquad \vdots \qquad \vdots$$

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

獨立性檢定：

→範例三、檢定母體各項特性間是否互相影響。

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，$i=1,..,r,\ j=1,..,c$,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

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

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：

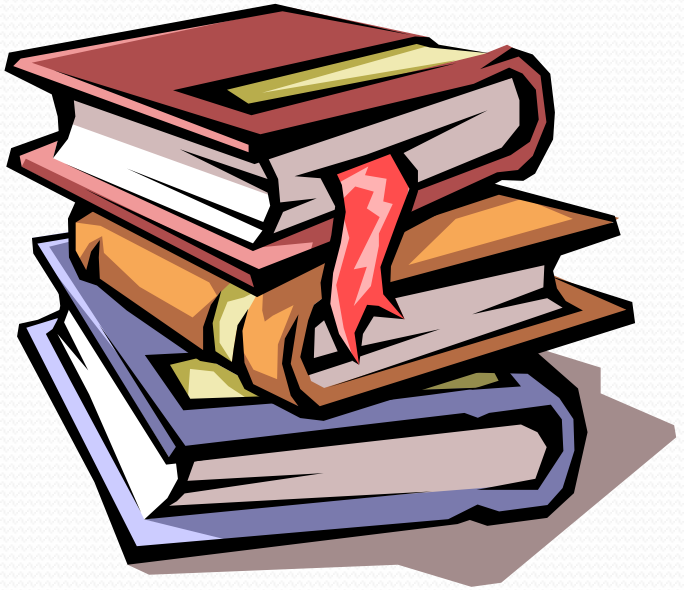
	商	工	藝術	公衛	列
男	0.0416 (16.625)	0.02375 (9.5)	0.38 (152)	0.0297 (11.875)	0.475
女	0.0459 (18.375)	0.02625 (10.5)	0.42 (168)	0.0328 (13.125)	0.525
行	0.0875	0.05	0.8	0.0625	1

檢定值＝13.675 ＞ $\chi^2(3)=7.815$

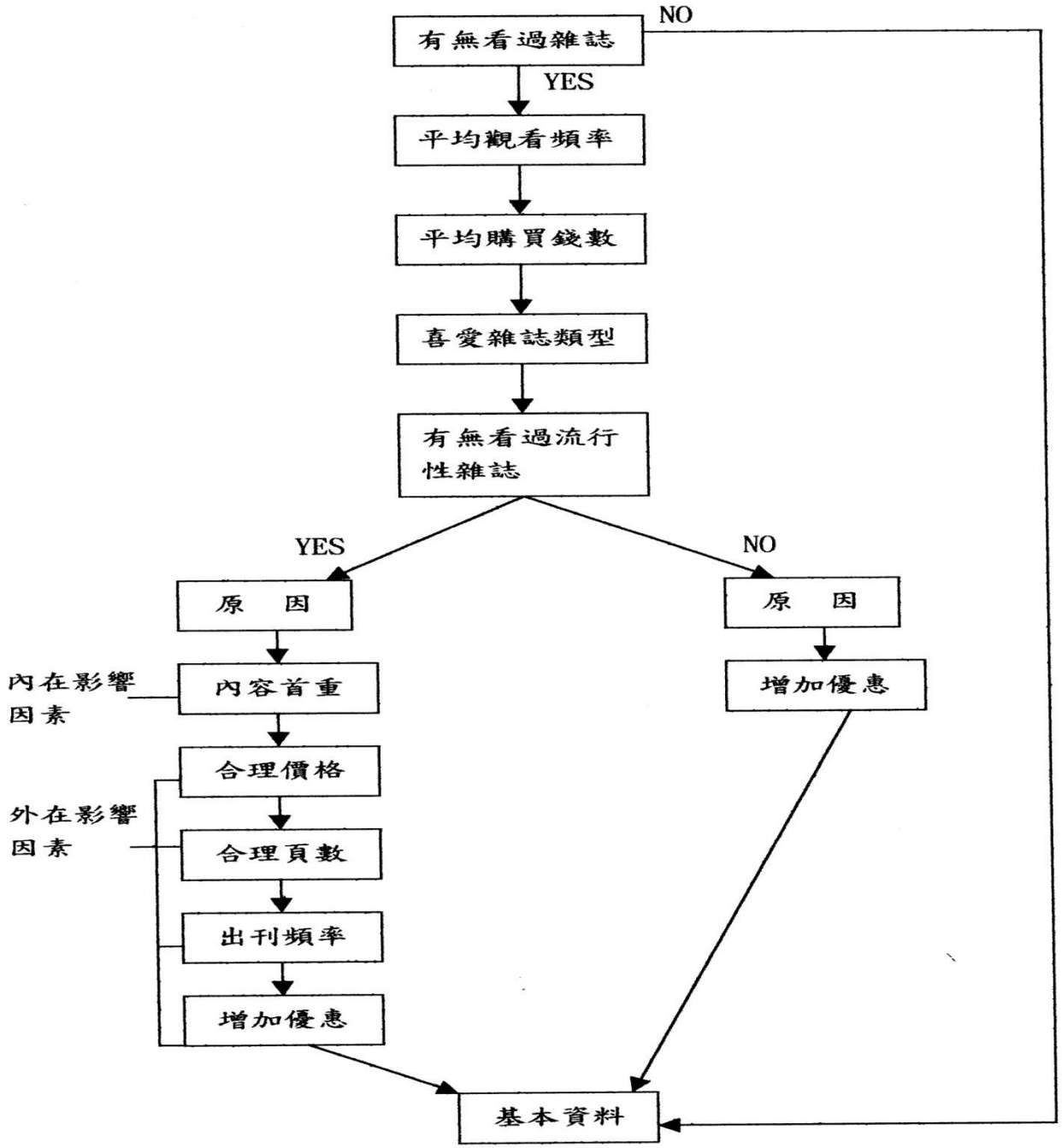
因此我們認為性別與選系有關聯。

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。



※問卷架構


有無看過雜誌
NO
YES
平均觀看頻率
平均購買錢數
喜愛雜誌類型
有無看過流行性雜誌
YES
NO
原　因
原　因
內在影響因素
內容首重
增加優惠
合理價格
外在影響因素
合理頁數
出刊頻率
增加優惠
基本資料

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

有效問卷率 = 500-（無效問卷+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

$$\sum \frac{(Oi-Ei)^2}{Ei} \sim\sim X^2_{(K-1)} \qquad k=1\ldots i$$

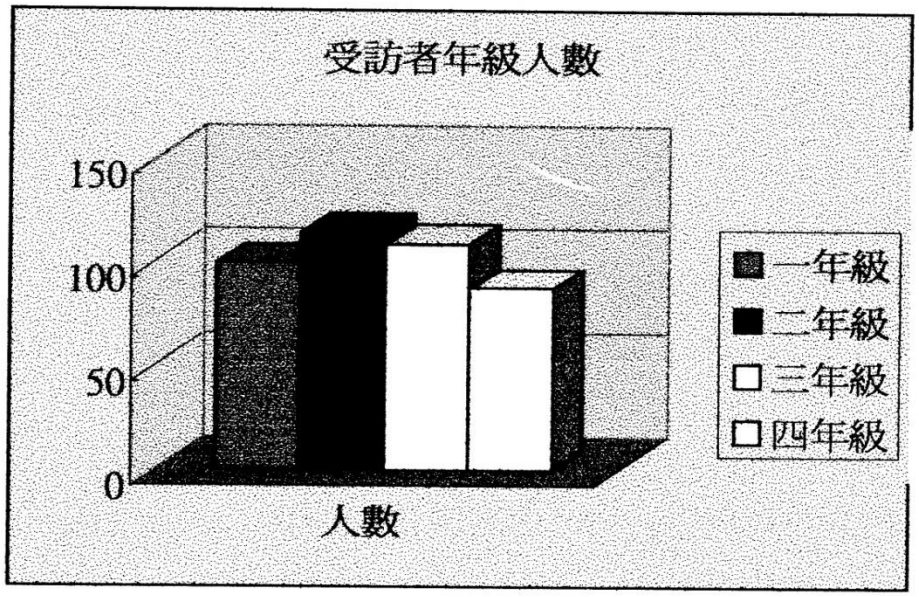
Oi：樣本觀察次數（實際出現次數）

Ei：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$X^2 < X^2_{(K-1)}$，即表示樣本代表性足以代表母體



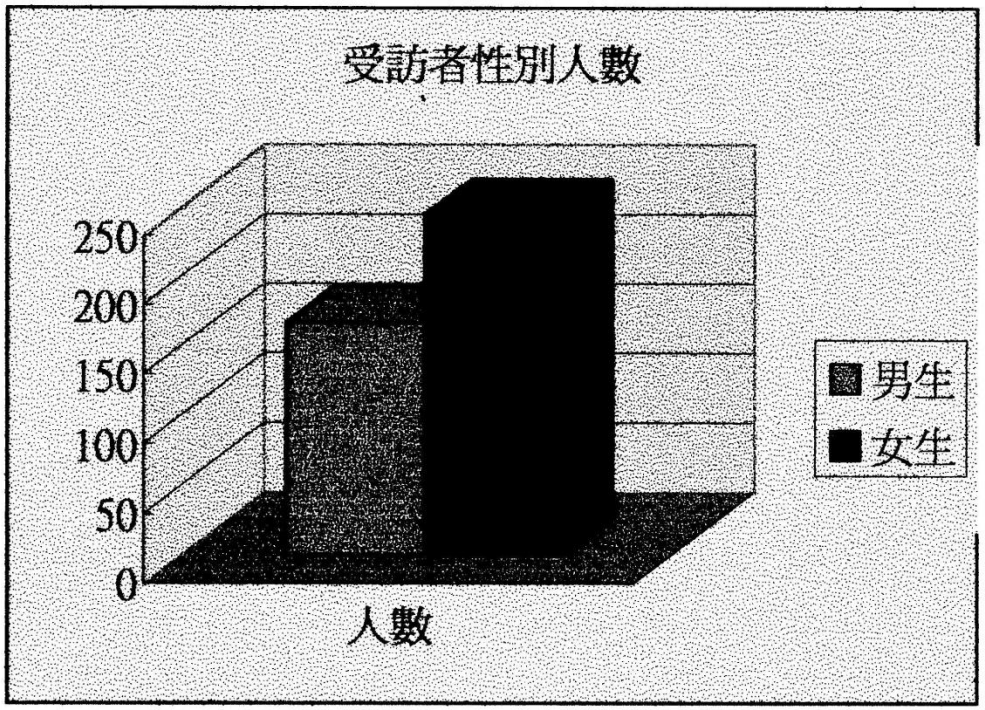
各年級中，以二年級受訪人數所佔比例最高，約有 27.9%，其次為三年級 26.5%。最少則為四年級，只有 21.4%。

	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。



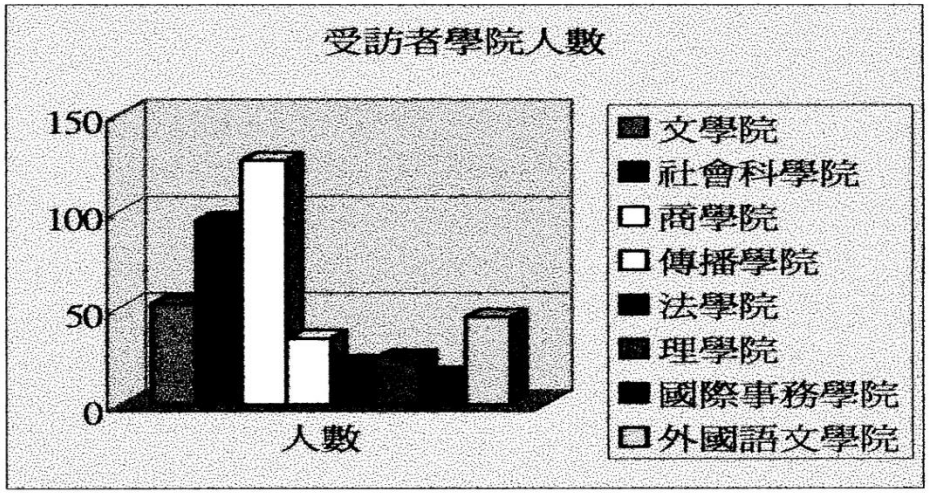
在受訪者中，男生佔了 41%，而女生的比例則為 59%。性別比，男：女約為 2：3。

	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。



受訪者學院已商學院居多，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

2. 獨立性檢定

根據我們的目的及興趣，利用基本資料和問卷內容中的項目做獨立性檢定，以及利用問卷內容中題目對題目做獨立性檢定，以探討兩個分類間的相關性程度如何。其中基本資料有：

＊一個月可使用的金錢：$0-$4000、$4001-$8000、$8000-$12000、$12001以上
＊ 學院：文學院、社會科學院、商學院、傳播學院、法學院、理學院、國際事務學院、外國語文學院、
＊年級：一年級、二年級、三年級、四年級
＊性別：男生、女生

分類一	分類二	chi-square值	P-value
一個月可使用金錢	多久看一次雜誌	15.396	0.081
一個月可使用金錢	每月平均花多少錢買雜誌	23.982	0.020
一個月可使用金錢	流行性雜誌的合理價格	26.252	0.010
一個月可使用金錢	流行性雜誌的合理頁數	28.921	0.004
學院	多久看一次雜誌	39.381	0.009
學院	最常閱讀的雜誌類型	63.440	0.080
學院	流行性雜誌內容重點	62.538	0.021
學院	流行性雜誌的出刊頻率	39.397	0.075
年級	平常有無看雜誌	9.285	0.026
年級	最常閱讀的雜誌類型	1.367	0.071
年級	不閱讀流行性雜誌的原因	8.950	0.030
年級	閱讀流行性雜誌的原因	18.707	0.096
年級	流行性雜誌的合理頁數	22.187	0.035
性別	最常閱讀的雜誌類型	61.759	0.000
性別	有無看過流行性雜誌	27.503	0.000
性別	閱讀流行性雜誌的原因	15.001	0.005
性別	流行性雜誌內容重點	14.103	0.029
性別	流行性雜誌的合理價格	11.122	0.025
性別	流行性雜誌的出刊頻率	10.890	0.028
多久看一次雜誌	流行性雜誌出刊的頻率	9.243	0.682
每月平均花多少錢買雜誌	流行性雜誌的合理價格	114.091	0.000
有無看過流行性雜誌	是否會因增加優惠而購買	14.858	0.000
不閱讀流行性雜誌的原因	是否會因增加優惠而購買	2.159	0.142

上述的獨立性檢定過程可見於後面的【分項說明】或【附錄】由於我們所使用的抽樣方法為先分層在群集抽樣，所以原先設定若以400份有效問卷做為樣本數時，樣本錯誤率α=0.05應比實際為小，所以我們把α擴大為0.1，若其p-value<α，則表示兩種分類間有相關性存在。

樣本代表性與加權調整

- 如果無法通過樣本代表性的檢定，通常採取加權調整加以補救，常見方法有二：

→事後分層加權(Post-stratification)：將樣本某些重要特質，經過加權轉換，使得該特質與母體一致。(Joint Probability Distribution!)

→反覆多重加權(Raking)：在臺灣較為常用，一次只調整一個特質（變數），通過檢定後再考慮另一變數。

參考資料：「樣本代表性檢定與最小差異加權：以2001年台灣選舉與民主化調查為例」選舉研究(2003)

表1　訪問成功樣本之代表性檢定：性別（加權前）

	樣　本		母群	檢　定　結　果
	人　數	百分比	百分比	
男	1012	50.0495%	50.8202%	卡方值=0.456
女	1010	49.9505%	49.1798%	p >0.05
合　計	2022	100.0%	100.0%	樣本與母群一致

*「母群」依據2000年戶口普查資料。

表2　訪問成功樣本之代表性檢定：年齡（加權前）

	樣　本		母群	檢　定　結　果
	人　數	百分比	百分比	
20—29歲	444	21.9585%	24.6171%	
30—39歲	470	23.2443%	24.5794%	卡方值＝16.643
40—49歲	450	22.2552%	21.9066%	$p < 0.05$
50—59歲	273	13.5015%	12.0141%	樣本與母群不一致
60歲以上	385	19.0406%	16.8827%	
合　　計	2022	100.0001%	99.9999%	

*「母群」依據2000年戶口普查資料。

表6　訪問成功樣本之代表性檢定：年齡（加權後）

	樣　本		母群	檢　定　結　果
	人　數	百分比	百分比	
20—29歲	470	23.2558%	24.6171%	卡方值＝0.133 p ＝0.998 樣本與母群一致
30—39歲	481	23.8001%	24.5794%	
40—49歲	452	22.3652%	21.9066%	
50—59歲	255	12.6175%	12.0141%	
60歲以上	363	17.9614%	16.8827%	
合　　計	2021	100.0%	99.9999%	

*「母群」依據2000年戶口普查資料。

表4　訪問成功樣本之代表性檢定：地理區域（加權前）

	樣　本		母群	檢　定　結　果
	人　數	百分比	百分比	
大台北都會	434	21.4639%	22.3455%	
大高雄都會	160	7.9130%	7.5282%	
北縣基隆	164	8.1078%	8.2584%	
桃竹苗	218	10.7814%	14.0094%	卡方值＝241.428
中彰投	319	15.7765%	17.9101%	p ＜0.05
雲嘉南	346	17.1118%	15.3915%	樣本與母群不一致
高屏澎	157	7.7646%	10.1363%	
宜花東	224	11.0781%	4.4205%	
合　　計	2022	100.0001%	99.9999%	

*「母群」依據2000年戶口普查資料。

表8　訪問成功樣本之代表性檢定：地理區域（加權後）

	樣　本		母群	檢　定　結　果
	人　數	百分比	百分比	
大台北都會	447	22.1068%	22.3456%	
大高雄都會	168	8.3086%	7.5282%	
北縣基隆	159	7.8635%	8.2584%	
桃竹苗	273	13.5015%	14.0094%	卡方值＝2.520
中彰投	382	18.8922%	17.9101%	p ＝.926
雲嘉南	318	15.7270%	15.3915%	樣本與母群一致
高屏澎	178	8.8032%	10.1363%	
宜花東	97	4.7972%	4.4205%	
合　　計	2022	100.0%	100.0%	

*「母群」依據2000年戶口普查資料。

民國 100 年 6 月

單位：人數；%

地區別	母體結構		加權前樣本分配		加權後樣本分配		卡方檢定
	人口數	百分比	樣本數	百分比	樣本數	百分比	
總計	8,670,326	100.00	6,455	100.00	6,455	100.00	加權前：
新北市	1,545,826	17.85	1,039	16.10	1,152	17.85	卡方值為 1155.72＞
臺北市	1,018,190	11.78	683	10.58	760	11.78	31.41（自由度為
臺中市	1,011,117	11.70	683	10.58	755	11.70	20，顯著度為 5%）
臺南市	694,559	8.04	466	7.22	519	8.04	在 5%顯著水準下，
高雄市	1,057,578	12.21	712	11.03	788	12.21	樣本與母體的地區
宜蘭縣	163,536	1.89	145	2.25	122	1.89	別結構有顯著差異。
桃園縣	760,260	8.72	511	7.92	563	8.72	加權後：
新竹縣	180,290	2.06	139	2.15	133	2.06	卡方值為 0.00<31.41
苗栗縣	194,080	2.23	138	2.14	144	2.23	在 5%顯著水準下，
彰化縣	458,188	5.28	310	4.80	341	5.28	樣本與母體的地區
南投縣	183,320	2.11	140	2.17	136	2.11	別結構無顯著差異。
雲林縣	237,018	2.72	158	2.45	175	2.72	
嘉義縣	181,132	2.07	139	2.15	134	2.07	
屏東縣	311,836	3.58	213	3.30	231	3.58	
臺東縣	78,645	0.91	139	2.15	58	0.91	
花蓮縣	120,532	1.39	142	2.20	90	1.39	
澎湖縣	33,965	0.39	138	2.14	25	0.39	
基隆市	144,391	1.66	142	2.20	107	1.66	
新竹市	154,278	1.78	139	2.15	115	1.78	
嘉義市	101,020	1.17	141	2.18	75	1.17	
金馬地區	40,564	0.47	138	2.14	31	0.47	

資料來源：100年婦女調查抽樣設計