統計，讓數字說話！

Statistics

concepts and controversies

Chapter 1

樣本

2009/07/31

不管有錢沒錢、黑人還是白人、民主黨或共和黨，
被選入的機會都是一樣。運氣不好時，
我們可能選出包含過多有錢的黑人共和黨員，
但是選取樣本的方法是不偏的。

抽樣詞彙

- **母體（population）：**
  我們求取資訊的對象全體，可能是人、動物或事物。

- **個體（unit）：**
  母體中的一份子。若母體包含的是人，我們則常稱這些人為受對象（subject）。

- **樣本（sample）：**
  母體的一份子，我們蒐集其資訊以便對整個母體做某些結論。

抽樣詞彙 (Cont.)

- 抽樣底冊（sampling frame）：
  個體的清單，我們從抽樣底冊中抽取樣本。

- 變數（variable）：
  個體的某種特質，被選入樣本的個體就會被度量這種特質。

Ex. 民意調查

- **Ex.**

蓋洛普及許多新聞機構常舉辦民意調查，探詢人民對某些議題的意見。此處的「變數」，就是人們對有關公共政策的問題的回答。這類民意調查整年都持續進行，但到了選舉前才特別受注意。一個典型的民意調查，其母體及樣本可能是以下狀況：

母體：

18歲以上的美國居民，包括非公民、甚至非法移民。

樣本：

從母體中選出、經由電話訪談的人。其人數在1000至1500之間。

Ex. 市場調查

- **Ex.**

  為了了解消費者的喜好及產品使用情形。市場調查的一個著名例子-尼爾森媒體研究-電視收視率調查服務。為了決定廣告商主花多少錢買某節目的廣告，也決定節目播不播得下去。對應於尼爾森全國電視收視率：

母體：

所有九千五百萬戶有電視機的美國住戶。

樣本：

約5000個主戶，住戶同意使用個人收視記錄器來記錄該戶中每個人收視的節目。

為什麼要抽樣？

- 當母體很大時，普查（census）既費錢又費時。
- 破壞性檢驗。
- 普查可提供每個很小區域的詳細料。
- 美國普查的主要作用，就是提小區域的地方資料。

怎樣取得爛樣本

- 有偏抽樣法：

  如果統計問題的設計使得結果總是往某個方向偏，此設計為有偏的。

- 方便抽樣法（convenience sampling）：

  從母體抽樣時，如果選最容易取得的。

- 自發性回應樣本（voluntary response sample）：

  經由對某一訴求的回應而自然形成的。

  ex. 意見調查。

★★方便樣本及自發性回應樣本常常是有偏的。

Ex. 購物中訪談

- Ex.

  製造業者和廣告代理商常利用在購物中心的訪談，以蒐集消費者習性及廣告效用等資訊。在購物中心裡取樣本既快速又省錢。但在購物中心裡訪談到的人並不能充分代表整個美國人口。例，這些人比較有錢，而且有很多青少年或退休人士。此外，訪問者傾向於從顧客群中選擇外表整潔、看起來不具威脅的人。購物中心樣本是有偏的，某些族群的比重太重，而有些族群的比重太輕。這樣一個方便樣本的意見，可能和全美大眾的意見有很大的出入。

簡單隨機抽樣

簡單隨機抽樣（simple random sample）

- 大小為*n*的簡單隨機樣本是有*n*個個體的樣本，其選取的方法是：使得抽樣底冊中任一組*n*個個體，被選中的機率都相同。（*n*代表樣本中的個體個數）

- 優點：公平、不偏（unbiased）

隨機數字

隨機數字表列出0,1,2,3,4,5,6,7,8及9這些數字，且滿足下述兩個性質：

- 表中任一個位置的數字，其為0,1,2,3,4,5,6,7,8或9中任何一個的機率相同 。
- 不同位置的數字之間是獨立的（independent），意指，一個位置的數字之數值，完全不會影響到其他位置的數字之數值。

選取簡單隨機抽樣

- **Step1.**
  編代碼：對抽樣底冊中每個個體指定一個數字代碼。

- **Step2.**
  用表：利用隨機數字來隨機選取代碼。

從樣本看母體

- 要想從樣本中得出什麼結論，要先知道樣本代表的母體是什麼。
- 取樣本是對母體做結論，不是樣本本身做結論。
- 測量樣本過程，可得到個體特質的描述（變數）-樣本統計量。
- 樣本統計量可用來作為描述母體特質-母體參數。
- 好的抽樣樣本可以得到具代表整個母體資訊。

參數及統計量

- **參數（parameter）：**

  描述母體的數字。參數是一個固定數字，但我們實際上無法知道參數的值。

- **統計量（statistic）：**

  描述樣本的數字。一旦取了樣本，統計量的值就知，但是換個不同的樣本，統計量的值就會改變。我們常用統計量來估計未知的參數。

抽樣變異及抽樣分佈

- **抽樣變異：**

  如果我們不斷從同一個母體抽取樣本，樣本統計量的值會隨樣本而變。

- **抽樣分佈（sampling distribution）：**

  若我們從同一個母體抽許多個樣本，其對應之樣本統計量會具有某種可預測的抽樣變異型態。

偏差及欠精確

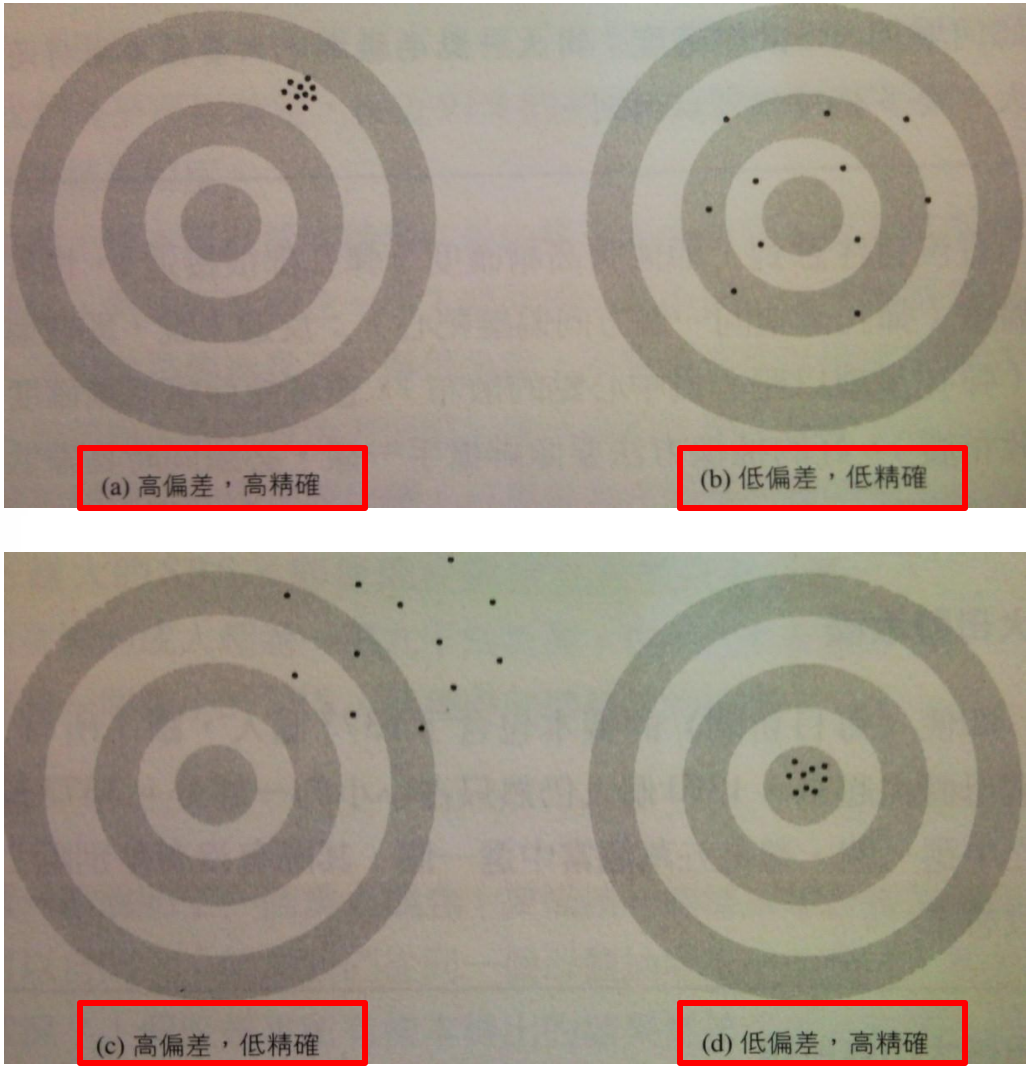
- **偏差：**

  樣本統計量朝同一個方向偏離母體參數值。

- **欠精確：**

  若不斷抽取樣本，在不同樣本之下，同一個統計量計算出來的值差異很大、很分散。無法指望每次抽樣的結果都差不多。

偏差及欠精確 (Cont.)



減低偏差

若要減低偏差：

- 利用隨機抽樣即可。若將整個母體列在抽樣底冊，再從中抽取簡單隨機樣本，就會得到不偏估計值（unbiased estimate），即，以簡單隨機抽樣得到的統計量估計母體參數，既不會高估，也不會低估。

增加精確度

如何增加簡單隨機抽樣的精確度：

- 用大一點的樣本。只要樣本取得足夠大，要多精確都可以做到。

母體大小無所謂

母體大小無所謂：

- 只要母體比樣本大得多，隨機樣本的統計量之精確性就和母體大小沒關係。

抽樣總結

- 要描述一個樣本是否值得信任

  →如果我們從同一個母體抽取很多樣本，會發生什麼狀況？

- 假設幾乎所有樣本得出的結果都接近真正的值，那麼即使並不知道樣本是否接近真正的值，還是可以對這個樣本有信心。

- 用大的簡單隨機抽樣可以保證幾乎所有的樣本都會得出精確的結果。

信賴敘述

信賴敘述：誤差界限&信賴水準

- 誤差界限：

  誤差界限告訴我們樣本統計量離母體參數多遠。

- 信賴水準（level of confidence）：

  信賴水準告訴我們所有可能樣本中有多少百分比滿足這樣的誤差界限。

信賴敘述的提示

- 信賴敘述的結論永遠是針對母體而不是針對樣本。
- 我們對母體所做的結論永遠不會是完全確定的。
- 如果我們要求99%的信賴水準，則必須接受比95%信心時大的誤差界限。
- 報告誤差界限時，用95%的信賴水準是很普遍的。
- 想在同樣的信賴水準下要求較小的誤差界限嗎？
  →取個大點的樣本就成了。

Ex. 全國調查

- Ex. 蓋洛普調查索取樣本之誤差界限之部分值

表1-1　蓋洛普調查1972年之前所使用抽樣步驟之精確性

母體比例	樣本大小						
	100	200	400	600	750	1000	1500
接近10	7	5	4	3	3	2	2
接近20	9	7	5	4	4	3	2
接近30	10	8	6	4	4	4	3
接近40	11	8	6	5	4	4	3
接近50	11	8	6	5	4	4	3
接近60	11	8	6	5	4	4	3
接近70	10	8	6	4	4	4	3
接近80	9	7	5	4	4	3	2
接近90	7	5	4	3	3	2	2

來源：George Gallup, *The Sophisticated Poll Watcher's Guide*（Princeton Opinion Press, 1972）, p. 228

Ex. 全國調查 (Cont.)

- Ex. 蓋洛普調查訪問了1514為成人，得知其中53%反對將學期延長。我們能對這結果做出怎樣的信賴敘述？

→我們有95%的信心說，介於50%~56%之間的成年美國人反對學期加長。

抽樣的實際面

- 抽樣誤差（error in sampling）：

  是抽樣這個動作造成的誤差。抽樣誤差使得樣本結果和普查結果不同。

- 隨機抽樣誤差（sampling error）：

  樣本統計量和母體參數之間的差距，是在選取樣本時因機遇造成的。信賴敘述中的誤差界限只包含隨機抽樣誤差。

- 非抽樣誤差（nonsampling error）：

  是和「從母體取樣本」這個動作無關的誤差。非抽樣誤差即使在普查中也可能出現。

抽樣誤差

Ex. 電話抽樣：

- 未登錄
- 同時擁有兩支電話
- 偏遠地區

非抽樣誤差

非抽樣誤差→連普查都可能逃不過的差錯

- 處理誤差→經由電腦普及化，已大幅減少此誤差
  - -計算錯誤
  - -Key In 錯誤
- 回應誤差
  - -記錯答案
  - -沒聽懂問題
  - -敏感議題
- 無回應→無法得到已經被選入樣本中個體的資料
  - -聯絡不上受訪對象
  - -拒絕回答

問題的措辭

- 不當或加料的問題
  → 足以左右抽樣調查之結果

- 對策：
  → 將問題寫的更清楚及更中立

相信調查結果之前該自問的問題

- 誰做的調查？
- 母體是什麼？
- 樣本怎樣選取的？
- 樣本多大？
- 應答率是多少？
- 用什麼方式聯絡受訪者？
- 調查是什麼時候做的？
- 問題確實是怎麼問的？

其他抽樣設計

- 多段抽樣設計

  縣(市)→區→路(街)

- 分層隨機抽樣

  Step1. 將母群組分成數個子群,稱每個子群為層

  Step2. 在每個子群中進行簡單隨機抽樣或系統抽樣

設計抽樣調查

- Step1. 決定母體
- Step2. 明確陳述要估量的變數
- Step3. 建立抽樣底冊
- Step4. 針對樣品做統計設計
- Step5. 注意細節

Ex. 民意調查與政治活動

反對選前民調 vs. 贊成選前民調

- 贊成者的論點：

  提供社社會大眾表達意見的機會。

- 反對者的論點：

  -抽樣結果的意見值代表什麼？

  -候選人做民調已是全世界通用的競選策略工具。

  -候選人運用民調進行反宣傳行銷策略。

Ex. 民意調查與政治活動 (Cont.)

用以判斷選情的選前民調永遠無法真正的了解以下兩個問題：

- 是哪些人真的去投票？
- 各候選人支持者的profiles到底如何？
  到底是誰支持誰？

Ex. 以隨機選取做為公共政策

1970 年美國首次舉辦徵兵抽籤

- 對象：
  19～25 歲合格役男
- 抽籤母體：
  編號 001～366 的等重塑膠球，001 代表 1 月 1 日出生者，031 代表 1 月 31 日...
- 366 個塑膠球放入滾筒中，混合，再一個個抽出。
- 抽出的號碼順序，代表徵兵順位。

Ex. 以隨機選取做為公共政策 (Cont.)

什麼時候應該用隨機選取決定公眾問題呢？

- 隨機選取的觀念就是抽籤，目的是給每一個人相同的機率。

- 如果政策要一視同仁，則隨機選取可以執行這個政策。

Ex. 資訊道德

基本資訊道德

- 施行研究的機構必須設立制度審查委員會(institutional review board)，負責事先審查所有的研究計畫，以保護受試對象，使受試對象免於受到可能的傷害。
- 在蒐集資料前，研究中的每一個受試對象，都必須在知情的狀況下同意受試。
- 任何個人資料都必須保密，只有整體的統計結果可以公開。

Thanks for your attention