統計實務

Spring 2024

授課教師：統計系余清祥

日期：2024年3月20日

第五週：蒐集資料

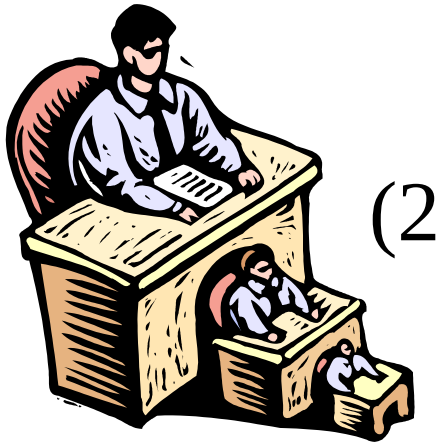
母體(Population)與樣本(Sample)

- 母體是具有共同特質的個體所組成的群體；樣本是自母體抽出的個體集合(母體的一部份)。

範例：(1)國中學生的智商(智力測驗)

母體→全台灣的國中學生

樣本→作智力測驗的國中學生



(2)台北市長候選人得票率(電話訪問)

母體→全台北市的合格選民

樣本→被訪談的台北市民(?)

資料蒐集的方式

- 一般將資料蒐集分類成：

1. 實驗設計(Experimental Design)

→ 包括臨床試驗 (Clinical Trials)，需要較精密計畫，一般分成實驗、對照組，較適合用於推論因果關係的研究。

2. 抽樣調查(Sampling Survey)

→ 設計問卷，藉由調查取得資訊。

- 目標：藉由蒐集的資料推得訊息。

- 另一種常見的資料來源分類，是依據資料產生分成：

  1. 實驗設計(Experimental Design)
  2. 觀察研究(Observational Study)

→兩者的差異在於資料蒐集者的參與，蒐集資料並不影響觀察研究，像是研究股市、利率、房地產價格，與實驗設計控制變因獲得觀察值不同。

註：參考Wikipedia及臨床實驗講義。

- 若以時間來區分，資料可分成：

  1. 縱向資料(Longitudinal Data)
  2. 橫向資料(Cross-sectional Data)

→縱向資料又稱為長期追蹤(Panel)資料，研究對象為固定個體，研究的特色在於可觀察相同個體因時間而有的變動，也稱為世代(Cohort)資料。橫向資料固定一個時間點，對當時的母體蒐集資料，在不同時間點獲得的資料不見得可互相比較。

註：國內外較知名的縱向資料包括「華人家庭動態研究」與PSID(Panel Study of Income Dynamics)。

為什麼要抽樣？

- 為什麼只看一部份的母體？

→普查(Census)：逐一檢查母體的所有個體。例如：戶口普查、工商業普查。

→普查需要較長的時間、較多的經費與人力，往往只有政府負擔得起。(政府也是每十年普查一次，其他時間輔以問卷調查、公務統計等等彌補資料的不足。)

→有時抽樣是唯一可行的方法。

抽樣的實例

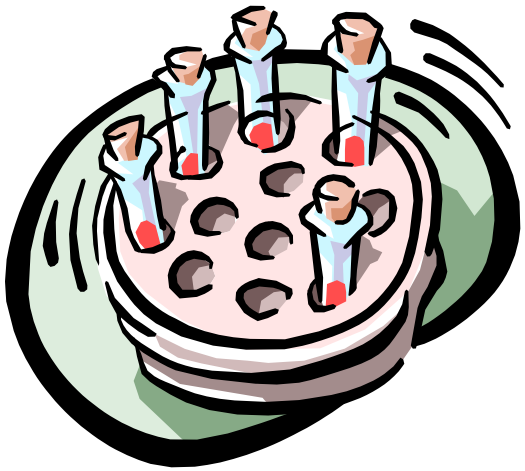
- 品質管制(Quality Control)

為確保品質，產品出廠時須經過檢查。但逐一檢查耗費過多的時間及金錢，通常每一批抽一個(或幾個)檢查。

→毀滅性抽樣(如鞭炮、罐頭等等產品)

- 健康檢查

抽血、切片或抹片檢查

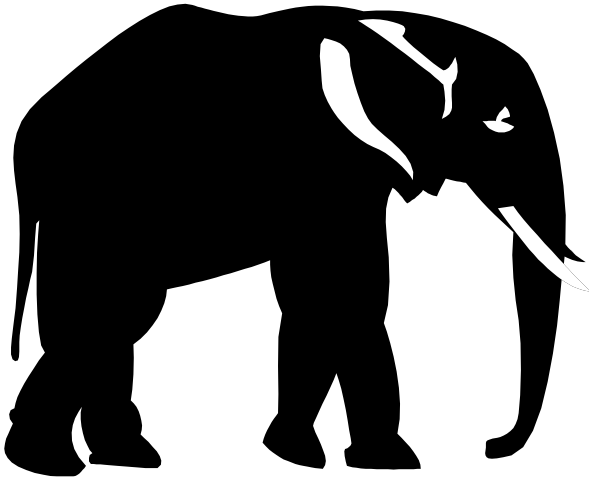


對樣本的要求

- 因為我們將從樣本推測出母體的原貌，抽出的部分必須能反映全體的特性，也就是說樣本需能代表母體。

  → ***樣本代表性！！！***

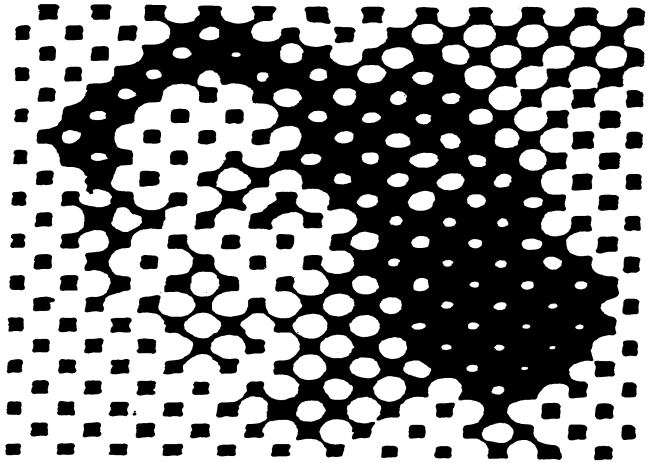
  → 最忌諱「瞎子摸象」



500,000

2,000

1,000



250

樣本對母體之代表性

目標樣本與實際樣本

- 無論是實驗設計或是觀察研究，抽取樣本需要謹慎規劃，確保目標與實際兩者一致。

→例如：藉由民意調查，獲取台北市民對市長的施政滿意度，先確定受訪者為台北市民，可先詢問受訪者是否為「居住」在台北市的市民。

註：「戶籍人口」vs.「常住人口」

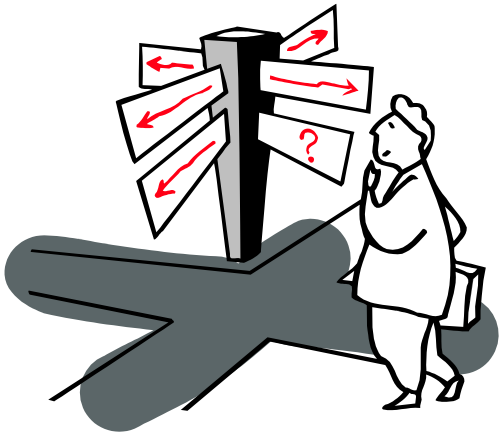
抽樣方法的分類

- 抽樣的方法可分為隨機抽樣(Random Sampling或稱為機率性抽樣)及非隨機抽樣，前者不加人為意志，僅以隨機抽取樣本；而後者則按人為意志選取具有典型代表性樣本。

  →隨機抽樣法因樣本以隨機抽出，較具代表性，但需要較完備的規劃，通常衍生的費用也較高。

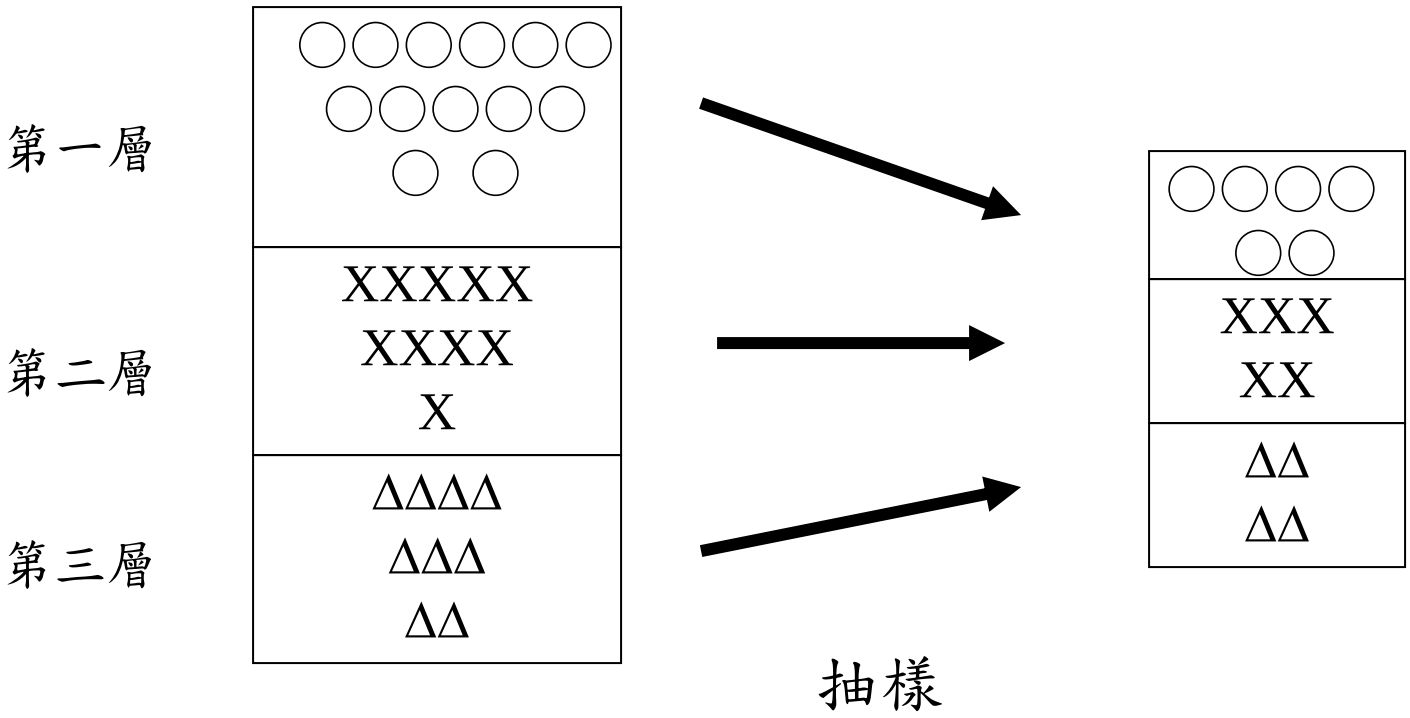
較常見的隨機抽樣法

- 簡單隨機抽樣(Simple Random Sampling)
- 分層隨機抽樣
- 集體隨機抽樣
- 系統抽樣
- 兩段抽樣

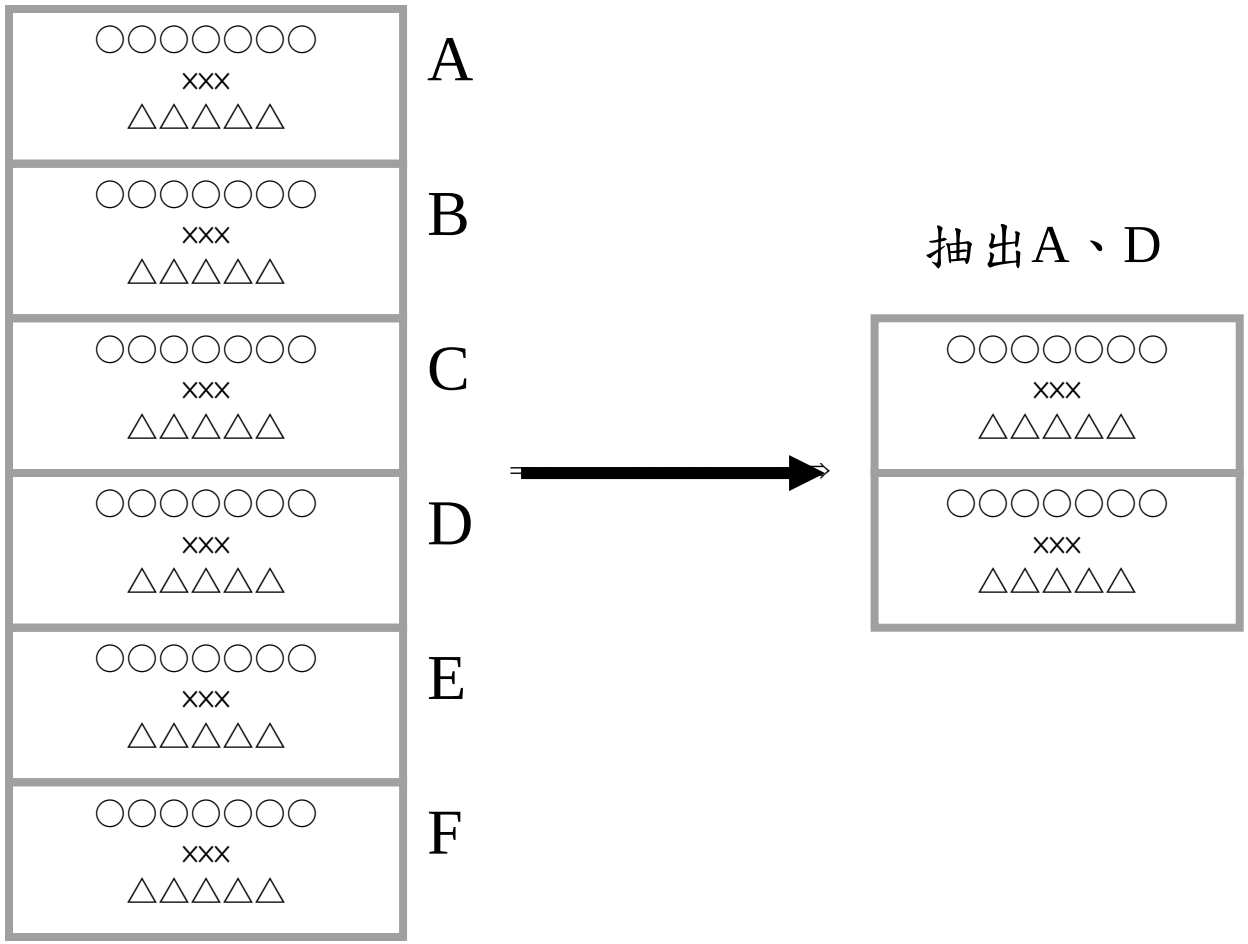


→簡單隨機抽樣如同摸彩，將所有的個體逐一編號再抽出。

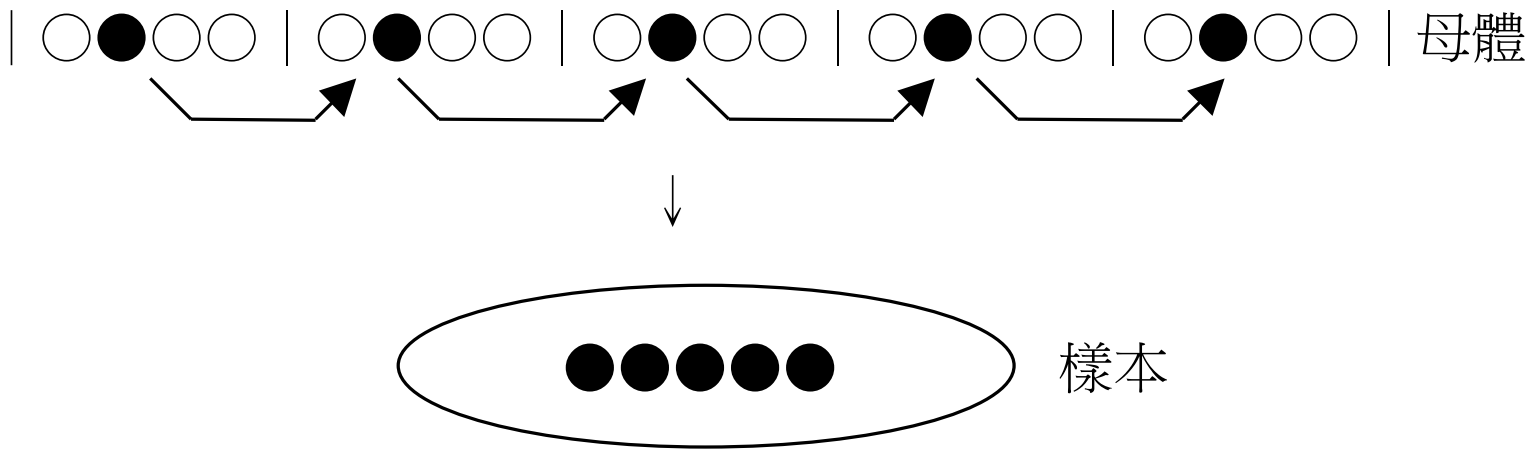
分層隨機抽樣(Stratified Random Sampling)



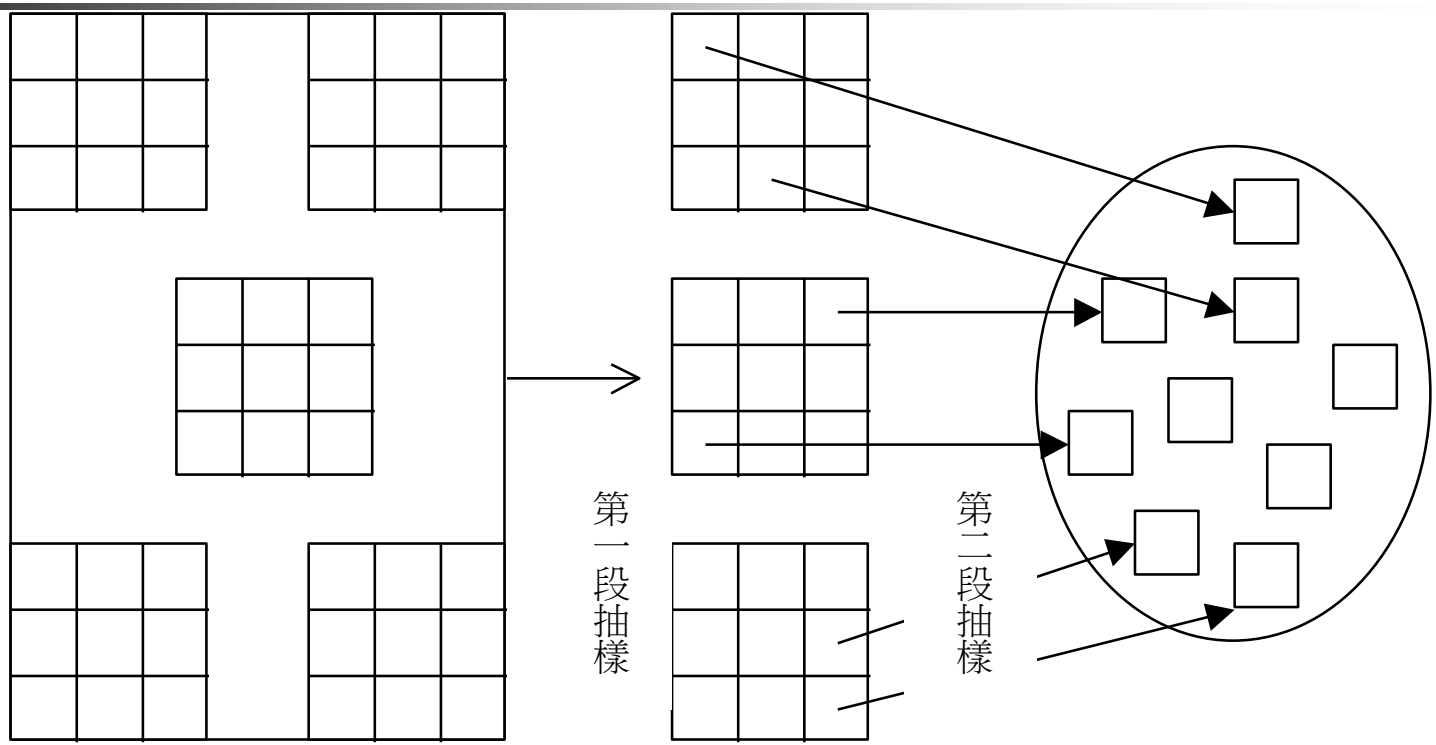
集體隨機抽樣(Cluster Random Sampling)



系統抽樣(Systematic Sampling)



兩段抽樣(Two Stage Sampling)



母體
設正方格共有 100 個為第一段抽樣單位

副次母體
設抽出五個大方格,每個大正方格內各含九個小正方格為第二抽樣單倍

樣本
每個大正方格中各抽出二個小正方格,共十個合成樣本

較常見的非隨機抽樣法

- 立意抽樣：不依隨機原則抽取樣本，而由母體中選取部份具有典型代表樣本。(e.g. 專家意見)
- 便利抽樣：事先不預定樣本，碰到即問或樣本自動回答。(e.g. 街頭調查)
- 滾球抽樣：利用樣本尋找樣本，對於特定族群樣本取得不易時採用。(e.g. 愛滋病的罹病人數)
- 配額抽樣：規定具有某種特性的樣本比例，類似分層隨機抽樣。

樣本好壞的判斷

- 母體的特性→參數(Parameter)
- 樣本中用以推測母體特性的估計值
  → 統計量(Statistic)

- 對統計量的要求：

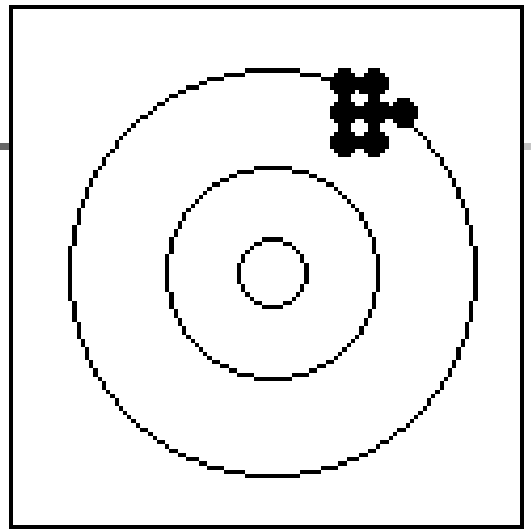
(1) 不偏(Unbiased): E(統計量) = 參數

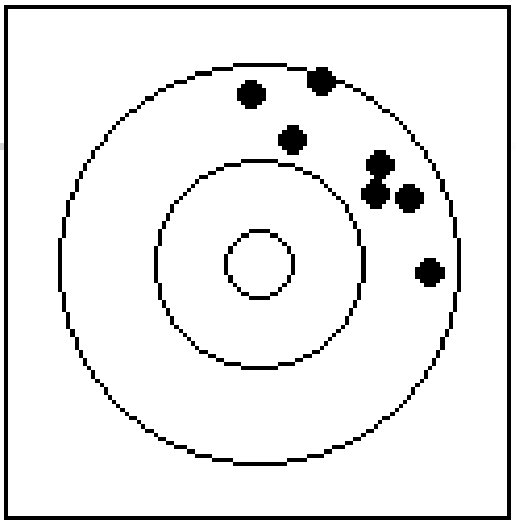
(2) 變異數(Variance)愈小愈好
  → 變異數與風險(Risk)有相似的涵意

Precise

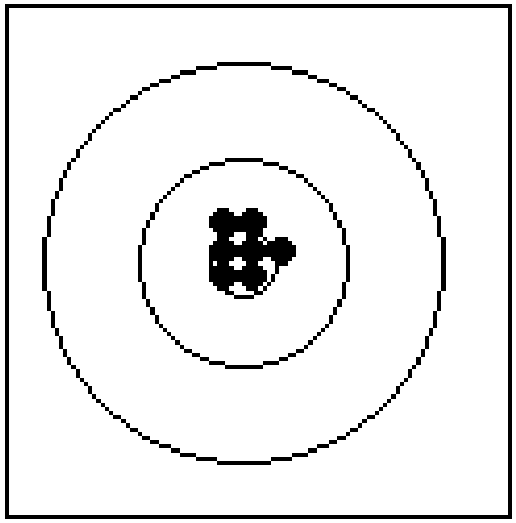
Imprecise

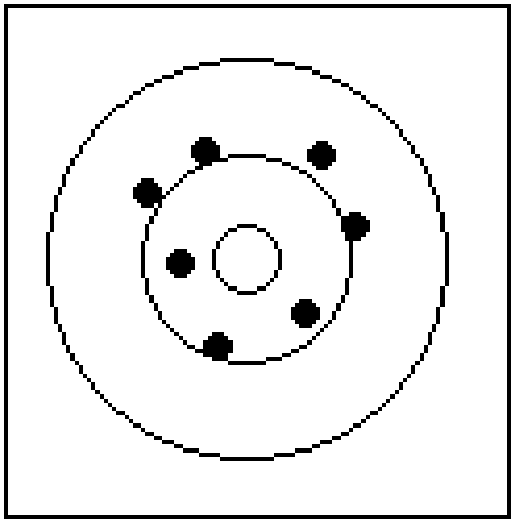
Biased





Unbiased





調查方法

- 調查方法通常可因獲取資料方法之異，通常分為：

(1)人員調查法(Personal Survey)

(2)電話調查法(Telephone Survey)

(3)郵件調查法(Mail Survey)

(4)網路調查法(Internet Survey)

問卷種類

- 問卷調查的題目通常分成三類：

1. 開放式：不列出可能答案，由被調查者自由作答。
2. 封閉式：(1)是否式(2)選擇式(3)排列式(4)填入式(5)尺度式
3. 半封閉式：封閉式為主體；若選項皆非填答者的選擇，則自由作答。

問卷題目範例

(1) 請問您本次購買的機車是

什麼廠牌_____ 汽缸大小_____c.c.

(2) 請問上一部機車行駛多少公里？

__15,000公里以下 __15,001~30,000公里

__30,001~45,000公里 __45,001~60,000公里

__60,001公里以上

（3）請問您打算幾年後換購新機車？

__1年以下 __1-2年 __3-4年

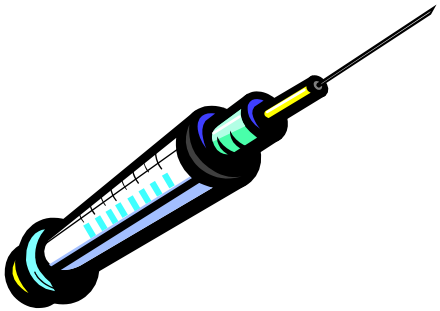
__5年以上 __其他(請說明)

多少樣本才足夠？

- 抽樣時常見的迷思：

→樣本數必須至少達到母體的一定比例？

範例(1)一般抽血不多於 10 c.c.，不論大人或小孩。



範例(2)台灣與美國人口數差了10倍以上，但民意調查多半只抽1,000份左右。

抽取1,000份樣本的原因

- 民意、市場調查的多為封閉問卷，有興趣的多為某個問項佔的比例，例如：某位候選人的支持程度→二項分配。
- 在信心水準為95%及最大誤差不大於 3% 的要求下：

$$1.96\times\sqrt{\frac{p(1-p)}{n}}\le 0.03$$

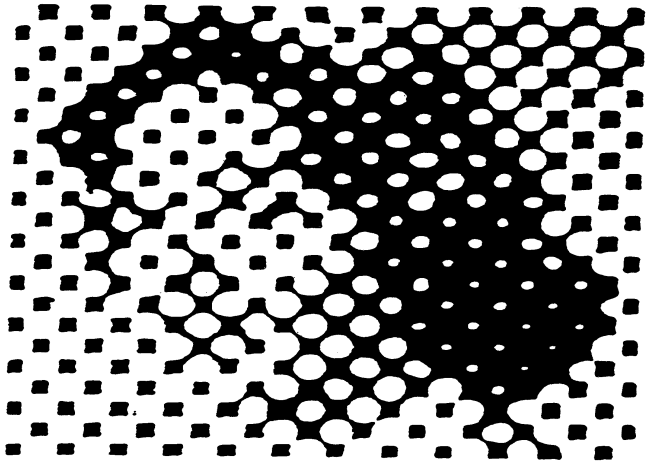
$$\Leftrightarrow \sqrt{n}\ge \frac{1.96\sqrt{p(1-p)}}{0.03}\cong \frac{1.96\times 1/2}{0.03}$$

$$\Leftrightarrow \quad n\ge 1{,}067$$

500,000

2,000

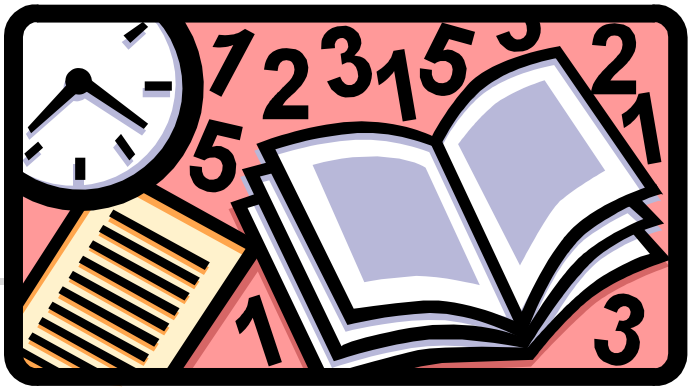
1,000



250

樣本對母體之代表性

問卷調查的步驟



- 定義問題、確定抽樣方法
- 問卷設計(Questionnaire Design)
- 問卷預試(Pretest)、訪員訓練
- 修訂問卷
- 正式訪問(發出問卷)
- 收回問卷、資料偵錯、資料輸入與整理

實驗設計

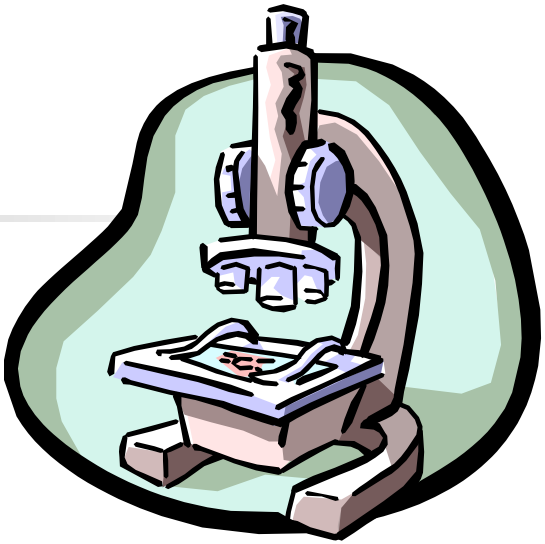
- 問卷調查蒐集的資料絕大多數屬於觀察研究(Observational Study)，經常無法確定觀察出的結果之成因。

→ 例如：研究發現國小學童中腳較大者，拼字能力也較強。(腳的大小影響拼字？)

- 實驗設計控制外在環境，只容許有興趣的部分(稱為「處理」；Treatment)變動，藉以分離出影響結果的原因。

- 由實驗設計應可推論出較精確的結果，但實驗設計的人力、金錢、時間的需求較高，且需更為精密的事前規劃。然而，實驗設計也無法使用於所有情形，有時問卷調查是唯一可能獲得資料的方法。

- 原則上而言，實驗可對因果關係提供好的證據。

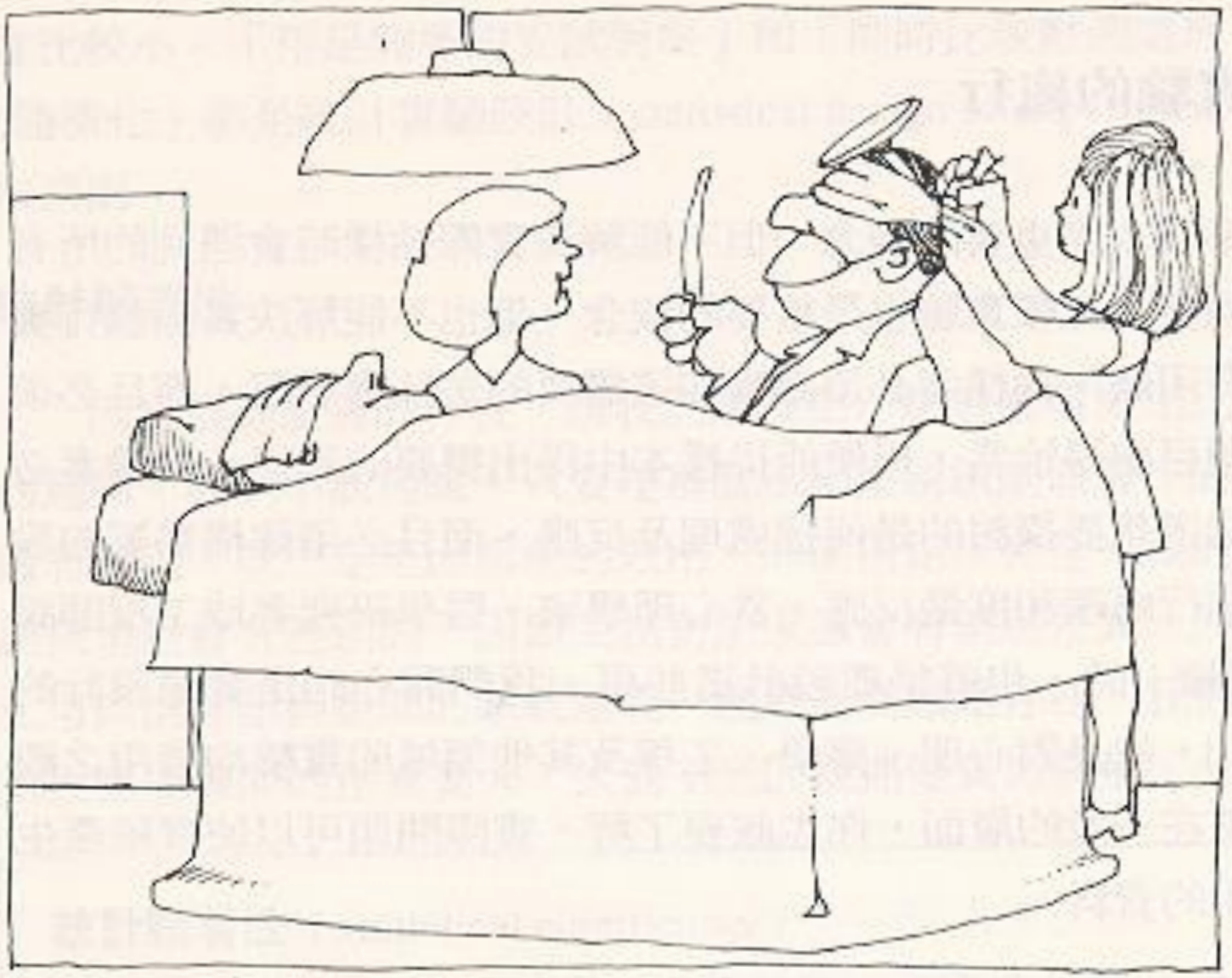


- 實驗組→處理；處方
  對照組→安慰劑(Placebo)

- 單盲與雙盲實驗：

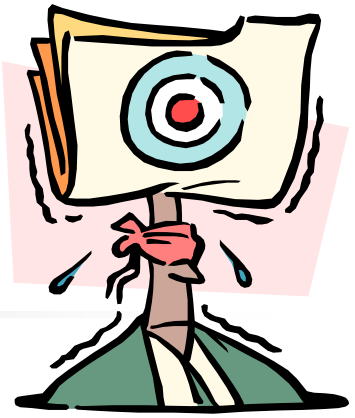
→ 單盲：只有受試者不知道自己的處方

→ 雙盲：醫生與受試者都不知道處方的分配方式



「伯恩斯醫師，您確定統計學家說的雙盲實驗是這個意思嗎？」

設計實驗時成本以外的考量



- 道德因素(Ethical factor)

  →讓重病病患使用可能較差的處方(或服用安慰劑)，雖然可證明實驗處方較佳，但也因此令病人縮短壽命(Patients' Right)。

- 公共政策的實驗

  →新的福利制度、健康保險等等公共政策的制訂，經常根據很多想像與很少資訊。對問題較小的政策且需比較的處理明確，通常較容易成功。

統合分析(Meta-analysis)

- 因為成本、時間及其他因素，研究可能需要合併不同研究的資料，這些資料可能有來不同來源、蒐集時間不一樣、或甚至有不同母體，如何因應問題需要結合資料，也是近年來另一種資料蒐集的方法。

→例如：各國蒐集該國罹患SARS、AIDS等疾病，希望找到共同的特性；選舉研究如何整合不同地區及時間得出的電訪結果，以獲得當前選舉的趨勢。