巨量資料與統計分析

政治大學統計系余清祥

2024年10月01日

第三週：蒐集資料

http://csyue.nccu.edu.tw

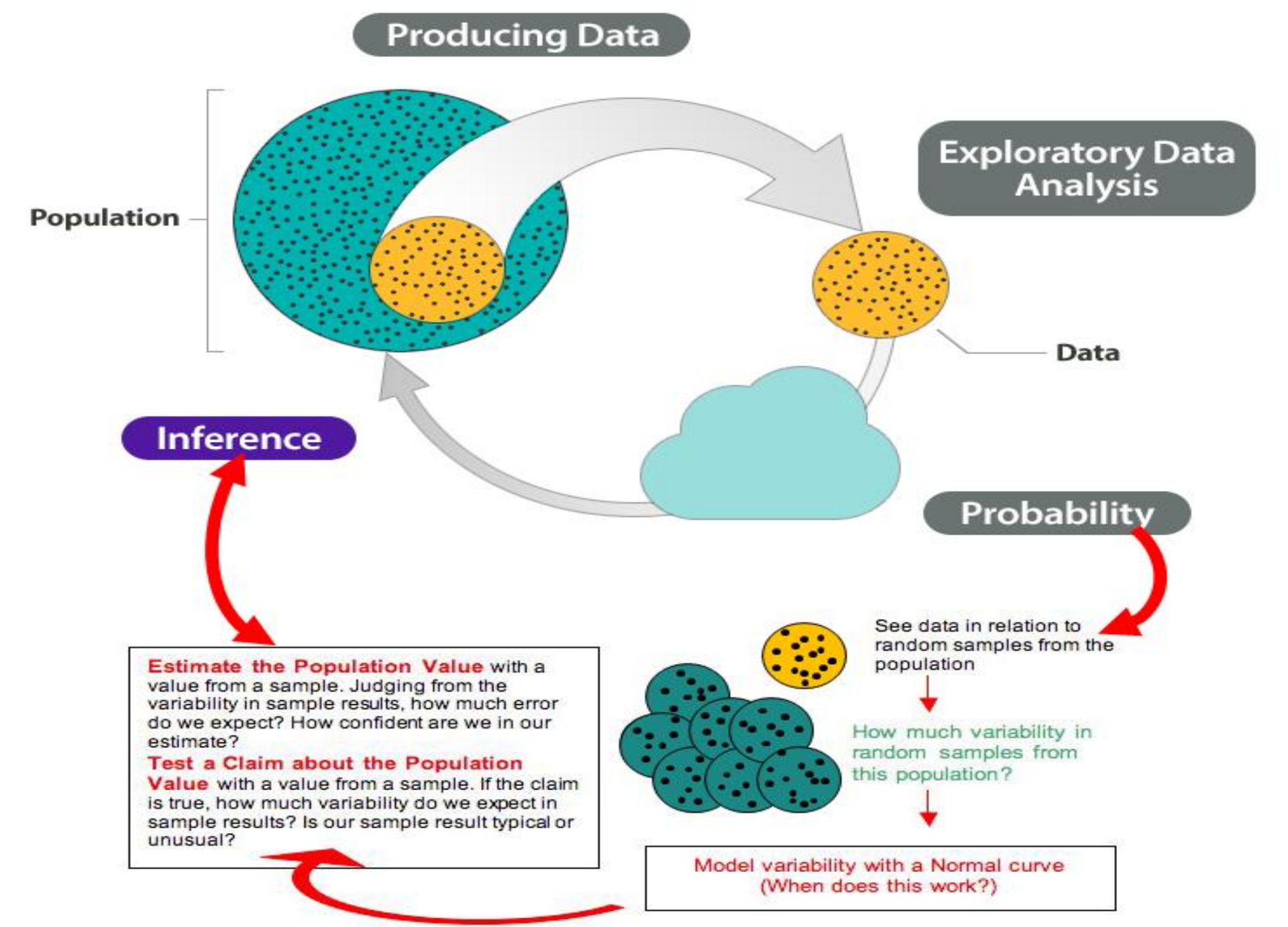
統計分析的流程

定義問題

蒐集資料

分析資料

詮釋結果



https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

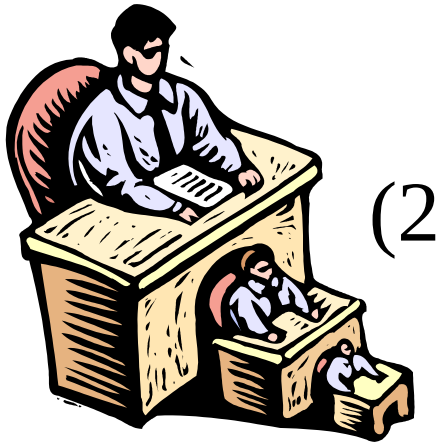
母體(Population)與樣本(Sample)

- 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(?)

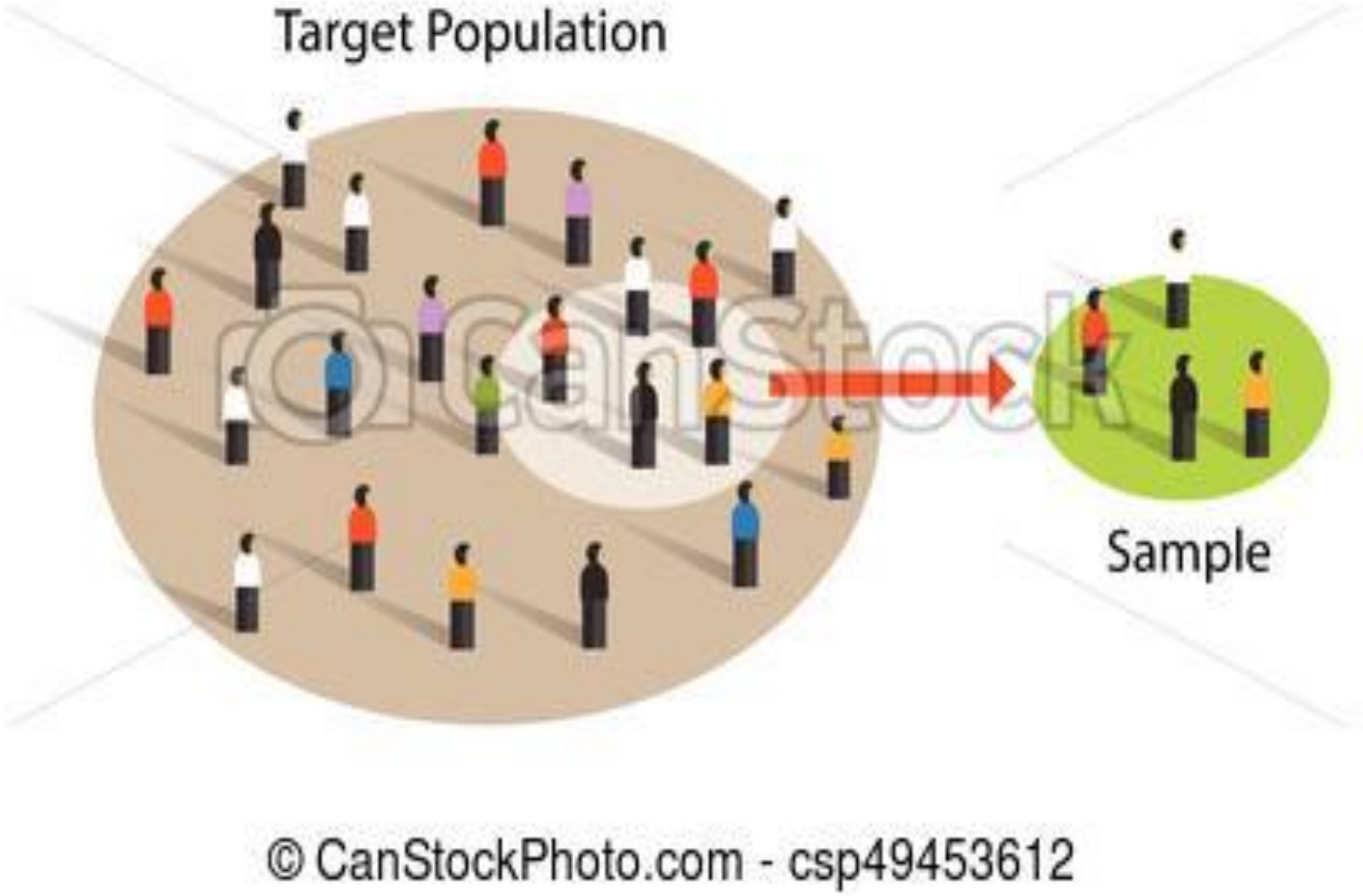
母體參數與估計

- 母體特性通常以參數（Parameter；母數）稱呼，像是成年人的身高、平均壽命、考試通過率。

→一般無法取得母體的資料，只能透過部分成員（亦即樣本）反推全體，統計因此也稱為反向推論(Inverse Inference)。

→但樣本是否能反映全體、分析方法是否適當都是關鍵！

統計分析：樣本→母體



https://comps.canstockphoto.com/sample-from-population-statistics-vector-clip-art_csp49453612.jpg

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：

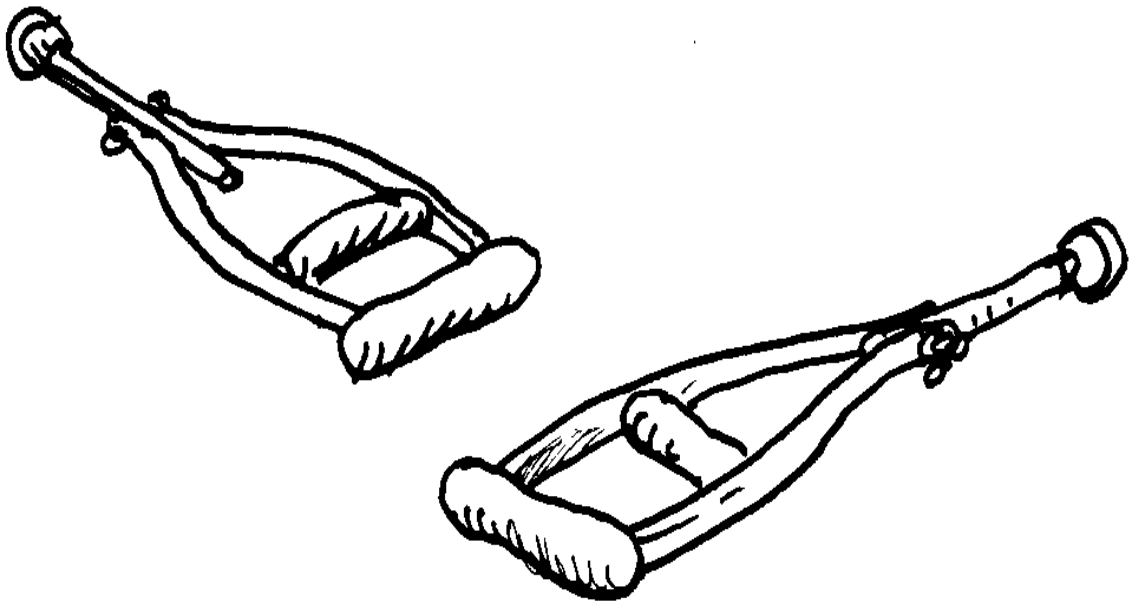
  1. 實驗設計(Experimental Design)
  2. 觀察研究(Observational Study)

→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：實驗設計較為費時費力，而且需要更為縝密的規劃設計，但投資報酬、附加價值通常也大許多。

規模最大的醫學實驗(沙克疫苗)

A MORE POSITIVE EXAMPLE IS THE *SALK POLIO VACCINE*. IN 1954, VACCINE TRIALS WERE PERFORMED ON SOME 400,000 CHILDREN, WITH STRICT CONTROLS TO ELIMINATE BIASED RESULTS. GOOD STATISTICAL ANALYSIS OF THE RESULTS FIRMLY ESTABLISHED THE VACCINE'S EFFECTIVENESS, AND TODAY POLIO IS ALMOST UNKNOWN.



- 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，研究相同個體在各時間的變動趨勢，也稱為世代(Cohort)資料。橫向資料則研究某個時間點的母體，但不同時間點的資料未必可互相比較。

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。

→普查需要較長時間、較多經費與人力，往往只有政府負擔得起。（政府也是每十年普查一次，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

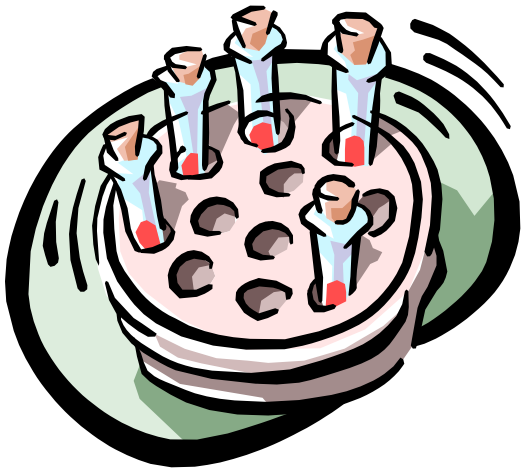
- 品質管制(Quality Control)

→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。

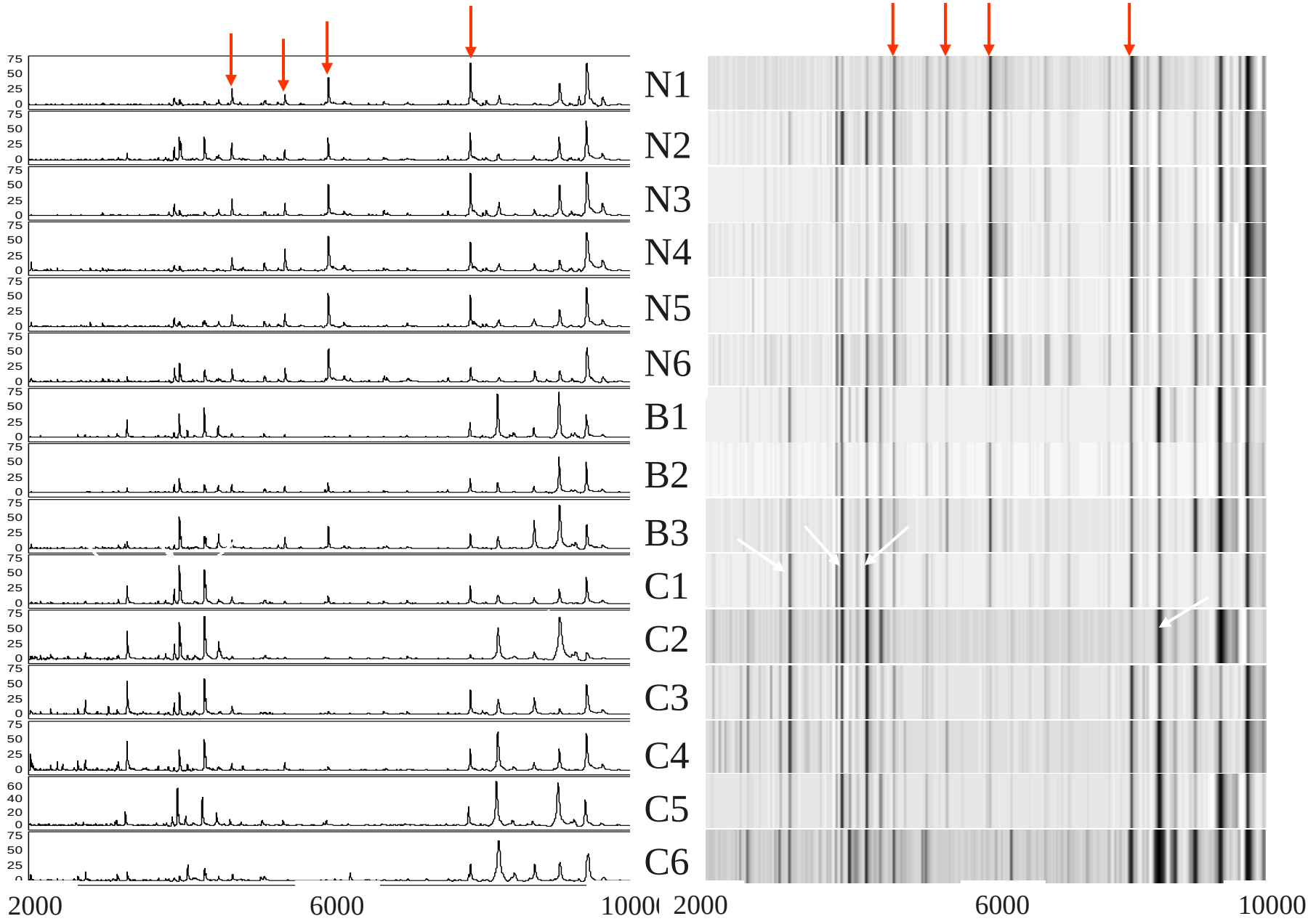
→毀滅性抽樣(如鞭炮、罐頭等等產品)

- 健康檢查

→抽血、切片或抹片檢查

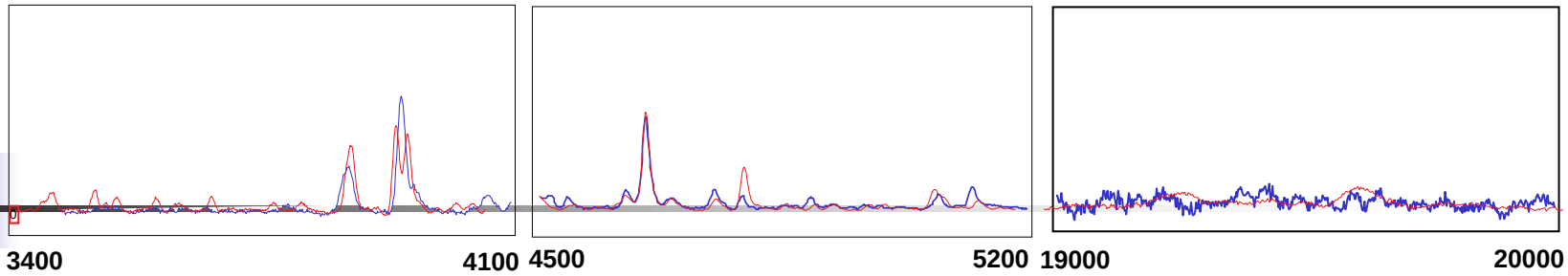


SELDI Serum Protein Profile Analysis-Prostate Cancer

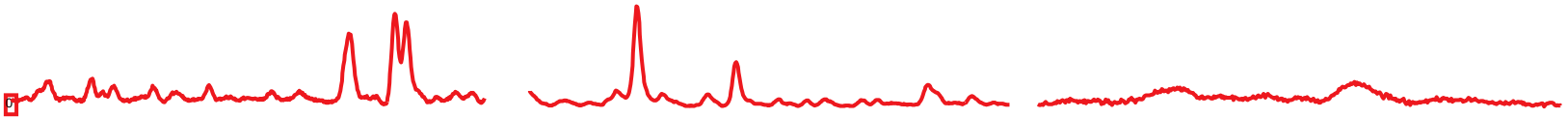


Proteomic Pattern of Sera from Patient

Stage 1 Profile

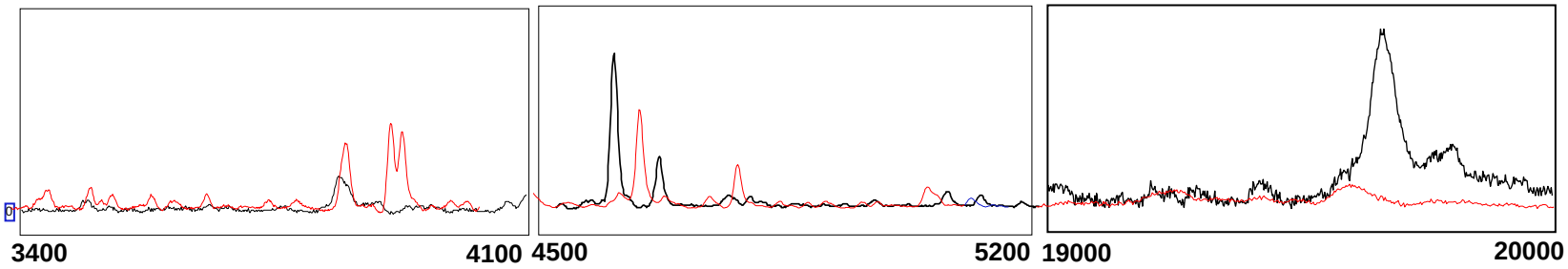


Sample



not a S2 pattern

Stage 2 Profile

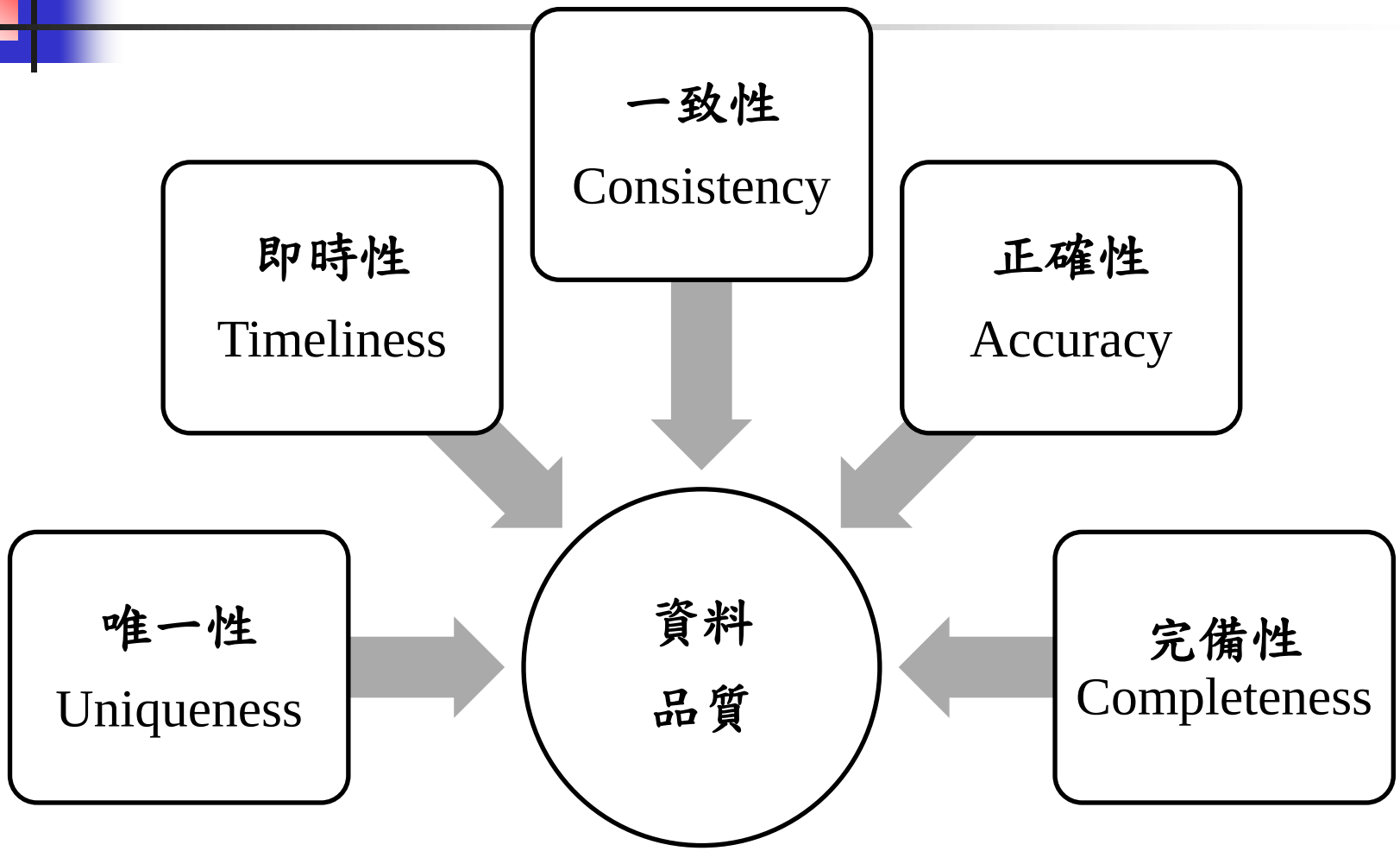


資料品質(Data Quality)

→Garbage in, garbage out!

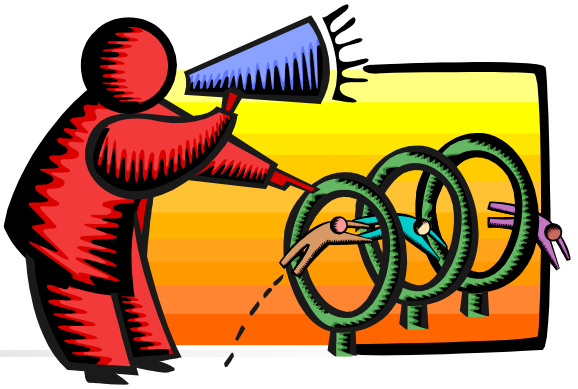
https://pbs.twimg.com/media/CCfGxAqW0AEDr8-.jpg

資料品質的傳統定義



來源：Guide for Checking the Quality of Syndromic Data, International Society for Disease Surveillance

正確性、完備性



- 資料品質通常分為五個層面：
  - **正確性**：正確紀錄資料的內容。透過比對編碼簿(Codebook)與各欄位數值、是否符合問題定義、出現異常數值等方式判斷資料正確性。
  - **完備性**：所有相關資料都被完整記錄下來，可被數量化的特性都忠實地被儲存。

→遺漏值通常含有豐富資訊！是否刪除？

→抽樣偏差(Biased Sampling) ex: 敏感性議題

唯一性、即時性、一致性

- **唯一性**：紀錄資料時不可模擬兩可。實際上記錄錯誤出線機會不低，特別是資料出自多個來源、輸入者不只一人。為確保唯一性通常會統一編碼簿、舉辦講習，確定資料有相同定義、輸入有一致規範。
  →利用主索引鍵(Primary Key)串聯不同資料庫
- **即時性**：資料紀錄為最即時的訊息，而非過時或不符現況；研究議題(對象)會影響即時性的定義。
- **一致性**：資料欄位在分析數據過程中需前後一致，在使用Panel Data時特別小心！

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，而足夠觀察值可看出母體原貌（三人成虎）。

→ 但為了避免「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。

資料品質！（錯覺或是瑕疵？）

https://img.ltn.com.tw/Upload/news/300/2017/12/13/phpZWbIuX.jpg

https://i2.wp.com/cdn.shortpixel.ai/client/to_webp,q_glossy,ret_img,w_1000,h_400/https://www.dataquest.io/wp-content/uploads/2019/08/garbage-in-garbage-out.jpg?zoom=2.5&w=450&ssl=1

https://i1.kknews.cc/UUCVcttsu9IA2ikKzpks8sQe8zhiDd5psO6Z4XqRPwo/0.jpg

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。

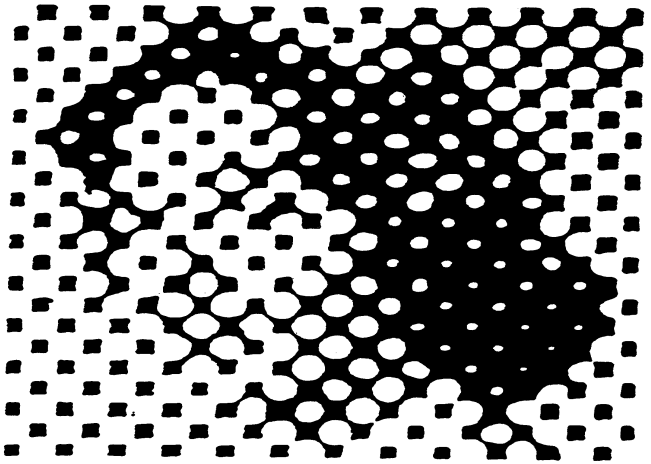
→ ***樣本代表性！！！***

→ 最忌諱「瞎子摸象」

500,000

2,000

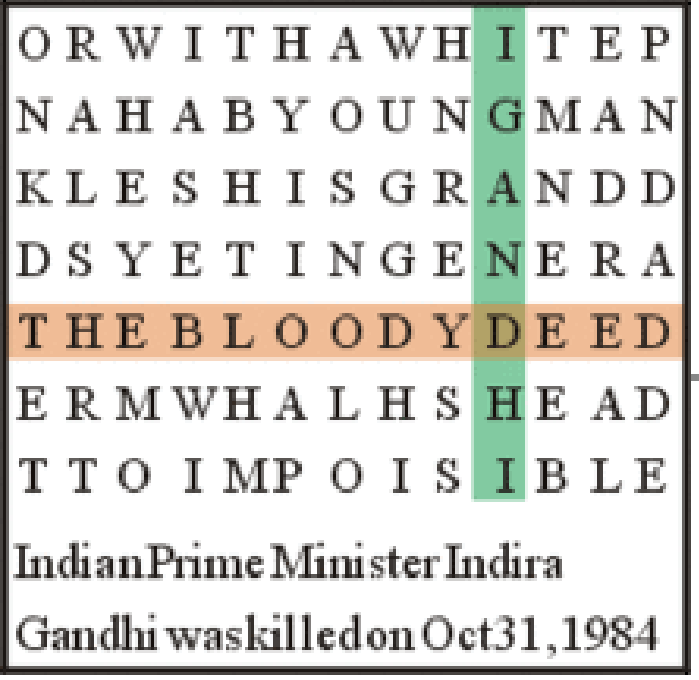
1,000



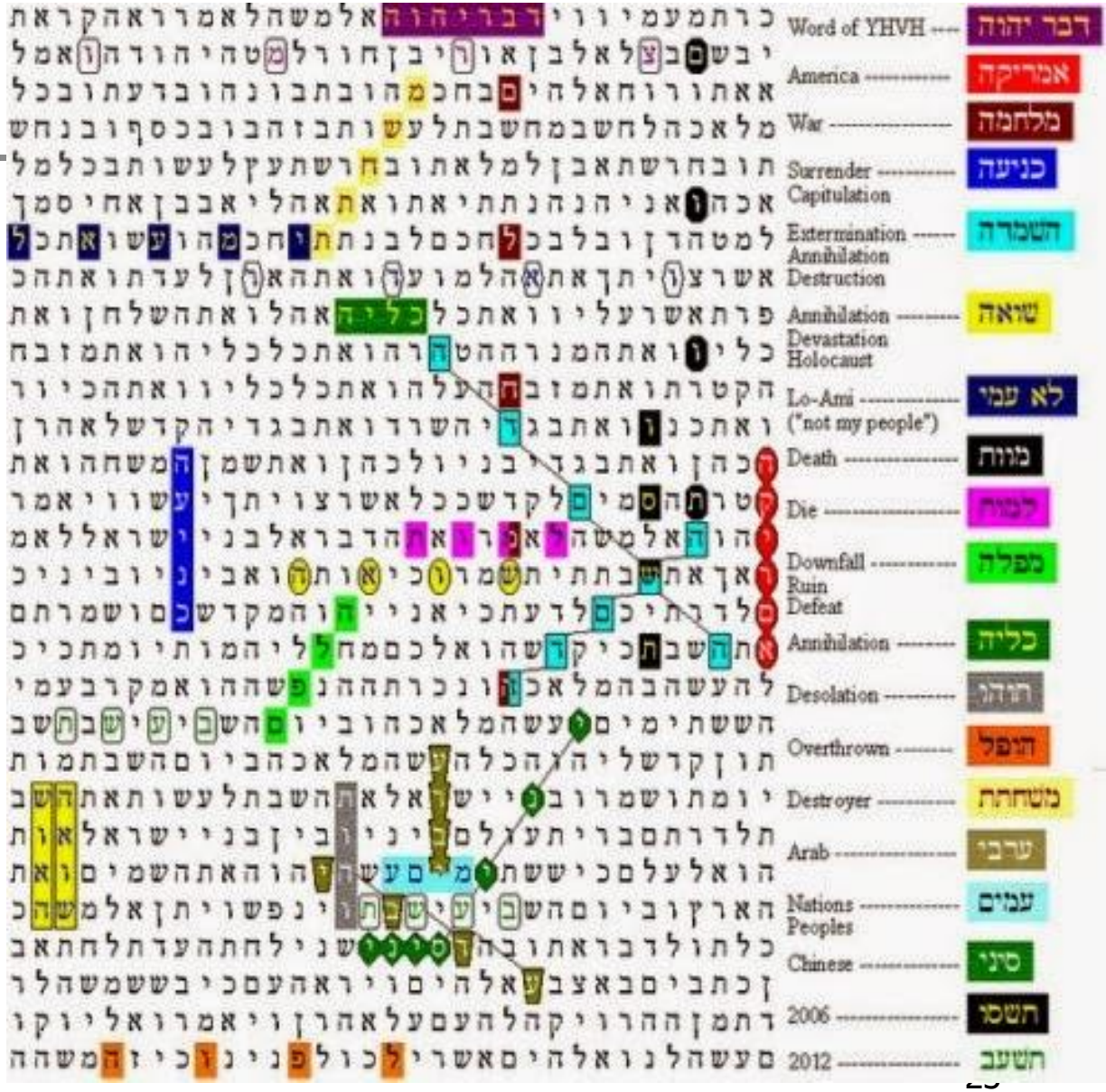
250

樣本對母體之代表性

The Bible Code



https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRbn2zY1y0w1ND8ys4Q5BUJlfzg3QGtE02qG4gGaaeIoIQMNT0KEw



聖經密碼（Torah密碼）

http://hidupgila01.blogspot.tw/2015/04/

博恩夜夜秀 第三季第七集

行前特別聲明

1. 本錄影為商業售票演出，敬請各位媒體朋友尊重智慧產權，切勿觸法。
2. 官方高畫質原音媒體素材免費商業授權請洽：press@strnetwork.cc
3. 在不妨礙現場觀賞體驗與錄影作業的前提下，歡迎觀眾自備應援道具。
4. 韓國瑜先生及團隊約定19:30抵達攝影棚，橋段20:15開始，敬請期待。
5. 假使內容因來賓行程有變，本節目無需亦無法負責，造成不便請見諒。

備註： 活動過程中，發生人員或道具影響錄影進行之情事，為維護購票觀眾之權益，請遵照維安人員指示離場。

敬請各界朋友拿出民主法治國家的素養
讓博恩與韓市長為大家帶來難忘的夜晚
期待與大家見面

製作人 Hauer

記得要準時

16:45 開始驗票
17:15 開放入場
18:00 正式開始

遲到需等待中場休息才能入場

千萬要帶錢

現場觀眾限定的扭蛋
記得攜帶足夠現金
可使用Linepay

扭蛋四款賣完就沒了

看秀五支箭

進場掃描ＱＲcode，結束會發大合照
請勿於官方上傳前爆雷
請依照現場導播指示
棚內全程禁止飲食拍照錄影錄音
將手機調整為靜音或振動

如何解讀資訊？（推背圖、燒餅歌）

北冥有魚，其名為鯤，鯤之身長過一八○，其重越百斤，一日，化而為巨鳥，其名為鵬，鵬之俊俏，堪稱鳥中之美女，鵬之大，無可比擬，故名之曰：鳥王。鳥王生活無目的可言，故常作無聊狀，天帝見其狀，甚怒，命其改過自新，並予一卵，令鵬孵之，十年歲月，孵出一人，其貌似書呆子，鳥王甚愛之，命名為莊子。某日，莊子為惡獸所包圍，莊子面不改色，引吭高歌，狀甚瀟灑，惡獸懼，皆沒入林中逃遁，莊子大笑，返以告鵬，鵬弗信，以為此乃水中撈月之事，絕無僅有，遂怒逐莊子，莊子甚感悲哀，準備遁世，後受仙人感召，出山林，作逍遙遊以告世人。

執編：李文堯 彭日榮 郭李同

藏頭詩「北一女的新書包沒水準」

北冥有魚，其名為鯤，鯤之身長過一八○，其重越百斤，一日，化而為巨鳥，其名為鵬，鵬之俊俏，堪稱鳥中之美女，鵬之大，無可比擬，故名之曰：鳥王。鳥王生活無目的可言，故常作無聊狀，天帝見其狀，甚怒，命其改過自新，並予一卵，令鵬孵之，十年歲月，孵出一人，其貌似書呆子，鳥王甚愛之，命名為莊子。某日，莊子為惡獸所包圍，莊子面不改色，引吭高歌，狀甚瀟灑，惡獸懼，皆沒入林中逃遁，莊子大笑，返以告鵬，鵬弗信，以為此乃水中撈月之事，絕無僅有，遂怒逐莊子，莊子甚感悲哀，準備遁世，後受仙人感召，出山林，作逍遙遊以告世人。

執編：李文堯 彭日樂 郭李同

EPOCHTIMES.COM

大紀元 - 台「北一女書包沒水準」楊照憶年少輕狂歲月

(大紀元記者江禹嬋台北報導) 17歲的青春歲月，該如何尋找自我？知名媒體人楊照說：「高中是擁有自我的開始，但卻還得活在別人給你的框架之中」所以他想盡辦法打破現有的規定，甚至在校刊上嘲弄隔鄰的女校，在文中嵌入「北一女的新書包沒水準」文字，他憶起，「那是我們

逍遙之音

逍遙遊發刊詞

北冥有魚，其名爲鯤，鯤之身長過一八〇，其重越百斤，一日，化而爲巨鳥，其名爲鵬，鵬之俊俏，堪稱鳥中之美女，鵬之大，無可比擬，故名之曰：鳥王。鳥王生活無目的可言，故常作無聊狀，天帝見其狀，甚怒，命其改過自新，並予一卵，令鵬孵之，十年歲月，孵出一人，其貌似書呆子，鳥王甚愛之，命名爲莊子。某日，莊子爲惡獸所包圍，莊子面不改色，引吭高歌，狀甚瀟灑，惡獸懼，皆沒入林中逃遁，莊子大笑，返以告鵬，鵬弗信，以爲此乃水中撈月之事，絕無僅有，遂怒逐莊子，莊子甚感悲哀，準備遁世，後受仙人感召，出山林，作逍遙遊以告世人。

執編：李文堯
彭日燊
郭李同

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」

大數據層級的資料蒐集

- 產業界（如人力資源）大多仍以傳統問卷形式蒐集資料，很容易造成流失及扭曲。

→**冷備份**vs.**熱備份**（高成本！）

→次級資料vs.原始資料（自由心證！）

→樣本vs.母體（抽樣偏差！）

- 註：誘導性文字（「額外資訊」）、敏感性議題（「收入」）等也會有品質問題。

抽樣調查的浮濫與誤用

『在超市與藥房，私人贊助的調查取代了醫師、父母與藥劑師的地位；在法庭上，各類調查已取代律師的功能；在立法院，民意調查是人民的代言人；市調更是廣告與促銷最有用的利器。市調與民意的關係是一種詭異的循環，個人的信念被千百名陌生人的信念左右』

--- 《真實的謊言》，時報文化

表3 匹茲堡睡眠品質量表(13)

請針對您過去一個月內夜間睡眠情形之大概狀況，回答最適合您情況的答案

1.過去一個月來，你通常何時上床？ _____時_____分

2.過去一個月來，你通常多久才能入睡？ _____分鐘

3.過去一個月來，你早上通常何時起來？ _____時_____分

4.過去一個月來，你實際每晚可以入睡幾小時？ _____時_____分

以下5、6、7、8題計分方式如下

0分：從來沒有 1分：一週少於一次 2分：一週兩次 3分：一週超過三次以上

5.過去一個月來，您睡眠問題被以下情況所干擾的次數如何？

(1)無法在30分鐘內入睡 ____ (2)半夜或凌晨便清醒 ____

(3)必須起來上廁所 ____ (4)覺得呼吸不順暢 ____

(5)大聲打鼾或咳嗽 ____ (6)會覺得冷 ____

(7)覺得躁熱 ____ (8)作惡夢 ____

(9)身上有疼痛 ____ (10)其他(請說明) ____

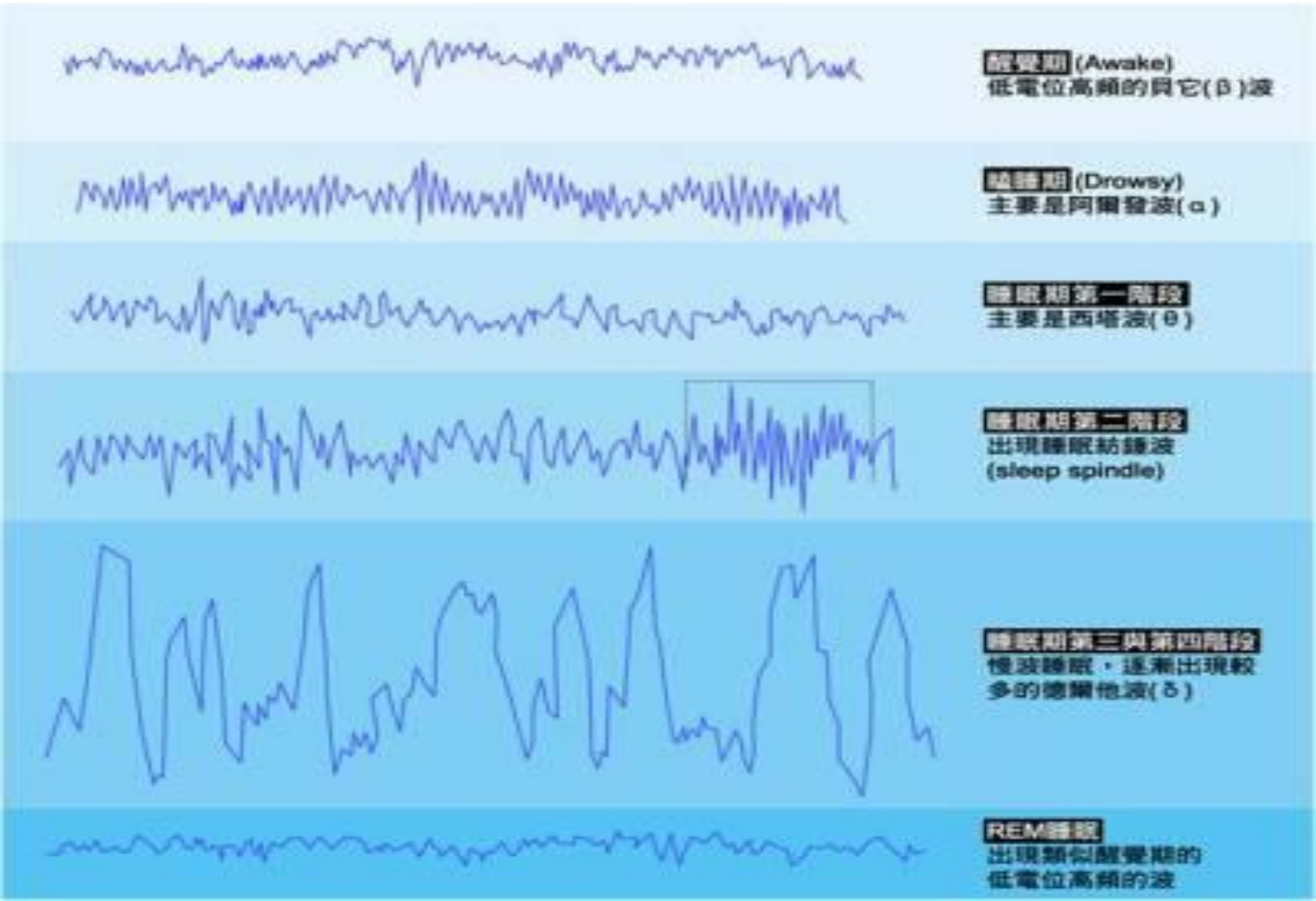
由受訪者填答可能衍生的問題？

表4 睡眠日誌(19)

日期	晚				午夜					早上							中午				下午					
Time	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
1/1	E				◎	X	X	X	X	X	X	X	○	C					S	+				
1/2																								
1/3																								

◎ 熄燈或躺在床上試圖睡著 XXXX睡著的時段 ○ 開燈或起床 ++++半夢半醒 C 飲用咖啡因的飲料(咖啡、汽水或茶)

儀器量測更為精確，但資料蒐集不易！



Soft Data vs. Hard Data

- Hard data常以有系統的方法蒐集，且多為容易量化的資料，Soft data則是較無系統、帶有個人判斷的質性資料，量測變數與目標值間相關（例如：長壽vs.幸福），但未必完全一致。
- 以手機的品質為例，可蒐集的Hard data包括拍照品質、通話及網路品質、待機時間等，Soft data則有品牌形象、外觀、介面等。

目標母體與實際母體

- 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」

資料分析vs.資料品質

- 許多人宣稱**資料量多寡**比**資料品質**重要，但「Garbage in, garbage out」，偏頗資料會扭曲我們的判斷（如：何不食肉糜）。
- 資料科學家**分析前應先確認資料來源可信度**，檢查資料品質是否有重大瑕疵。

→網路有名的「世界四大不能信」：英國研究、臺灣報導、中國製造、韓國發源。

google.org Flu Trends

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

FAQ

Flu activity

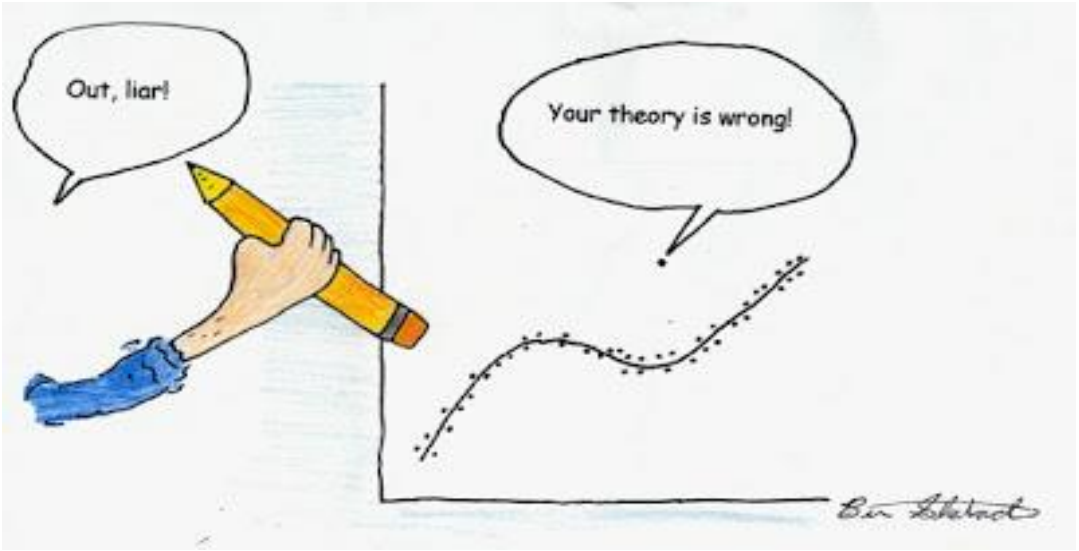
- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

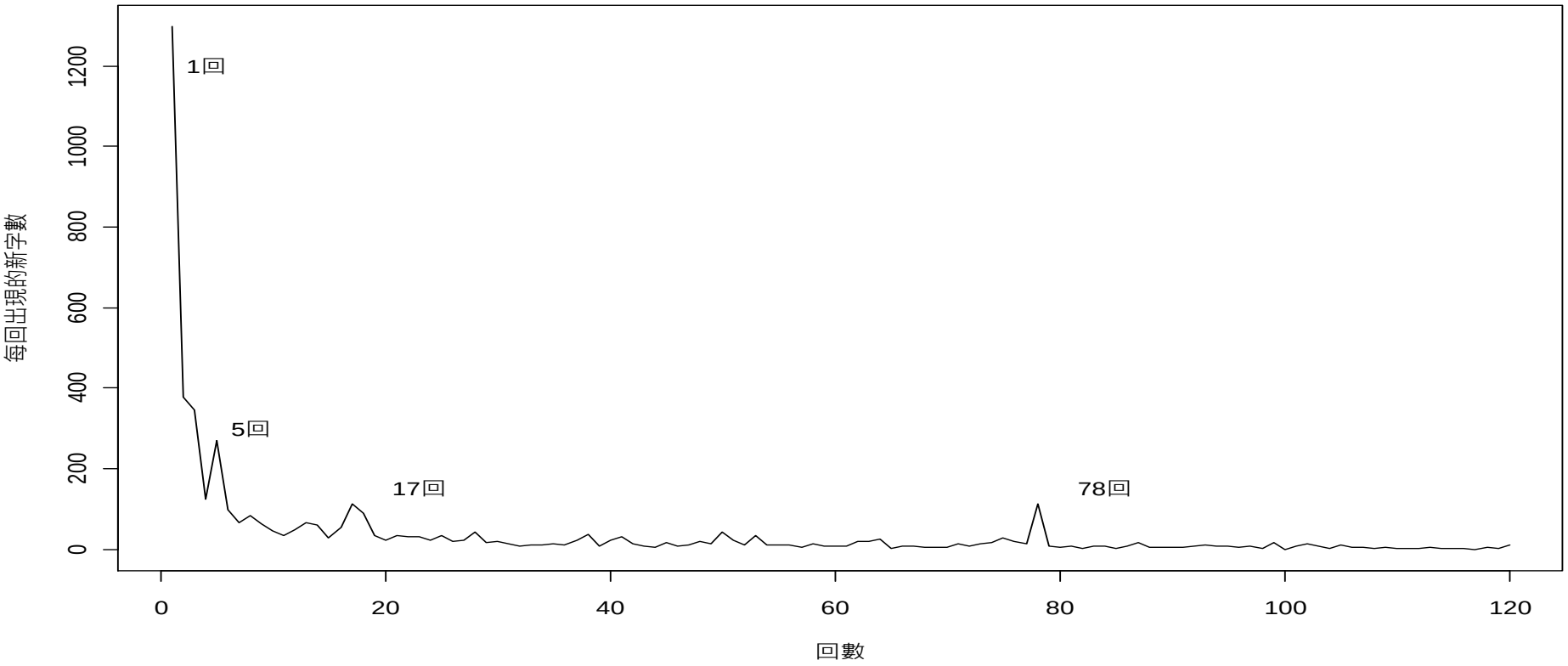
倖存者偏差(Survivorship Bias)

- 有時無法判斷蒐集到局部或全體的資料，檢查是否存在倖存者偏差（例如：谷歌流感趨勢預測）；不少人為了省事與得到漂亮結果，而移除離群值！



紅樓夢每回新出現字數序列圖

- 第1回、第5回、第17回與第78回新字彙特別多，可視為**離群值**！但其實含有豐富資訊。



抽樣方法的分類

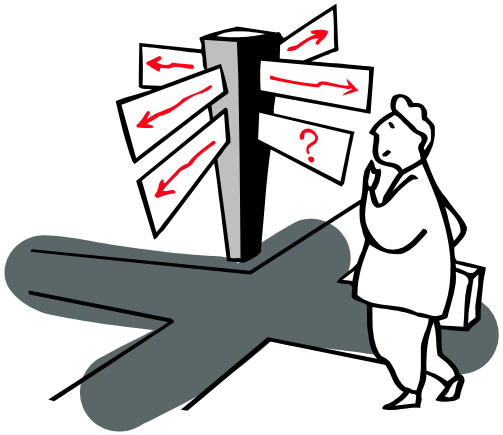
- 抽樣方法可分為隨機抽樣(或機率性抽樣，Random Sampling)及非隨機抽樣，前者不加人為意志，僅以隨機抽取樣本；後者按人為建議選取具有典型代表性樣本。

→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。

註：**市話**、**手機**抽樣的差異、進行方式？

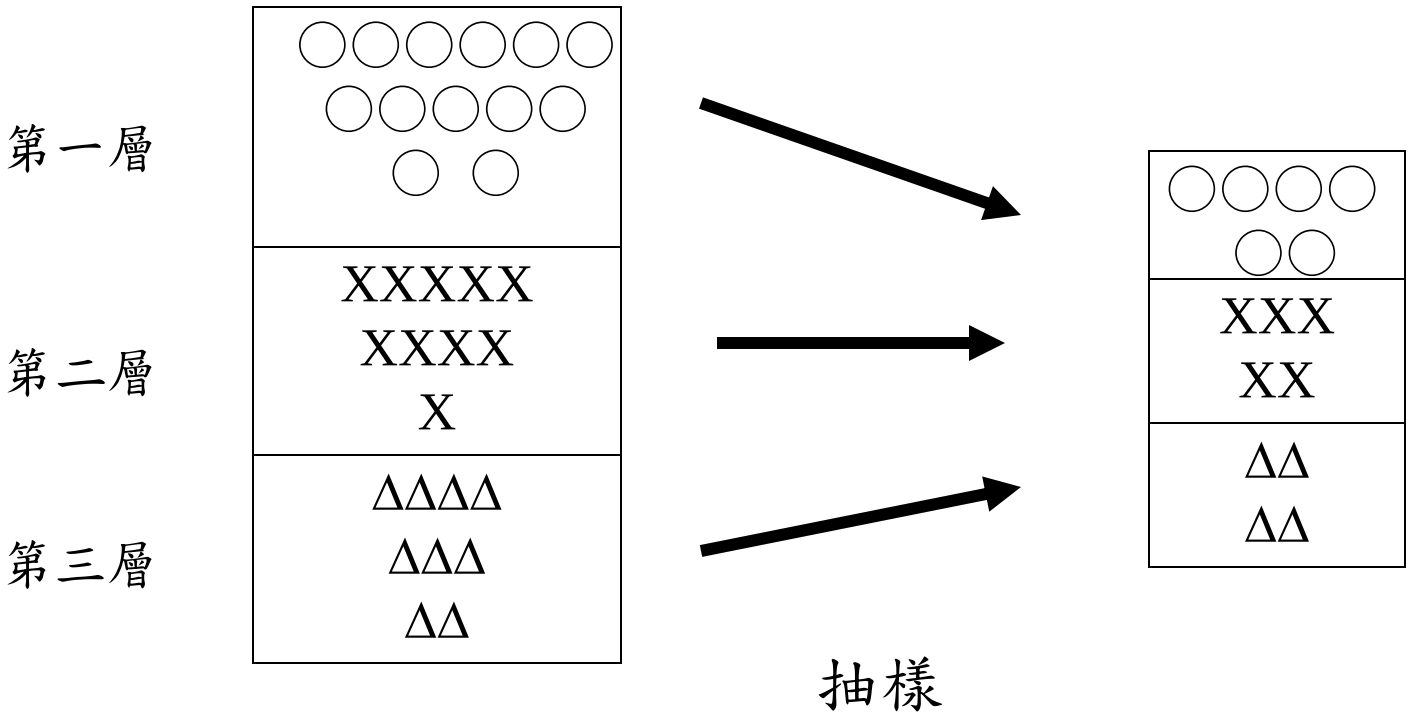
較常見的隨機抽樣法

- 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)



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。

隨機抽樣 vs. 非隨機抽樣

- 隨機抽樣不代表每個個體被抽到的機會都相同，而是樣本選取不受人為因素影響！
- 問題：隨機抽樣有什麼優點？何時會選用非隨機抽樣？

→「隨機」意謂樣本選取與某個機率分配有關，估計及推論較有依據。

→對於罕見事件、蒐集未知領域的資訊，非隨機抽樣往往更合適。(例如：調適型抽樣，Adaptive Sampling)

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

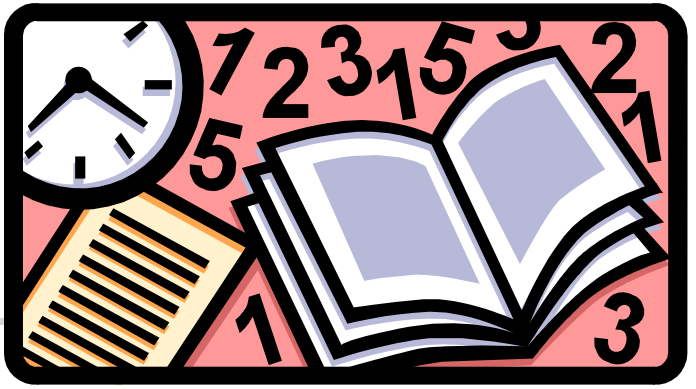
__60,001公里以上

（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)

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

一、基本資料：

捐血日期：_____　　捐血地點：_____

性別：□男　□女

年齡：□17-20 歲□21-30 歲□31-40 歲□41-50 歲□51-65 歲

教育程度：□國小□國中□高中/職□大學/專□研究所以上

職業：□工商□製造業□服務業□軍公教□學生□自由業□其他

二、引發您此次捐血的動機是：(可複選)□師長朋友鼓勵□醫療人員建議□相關宣導活動影響□完全自發性□其他_____

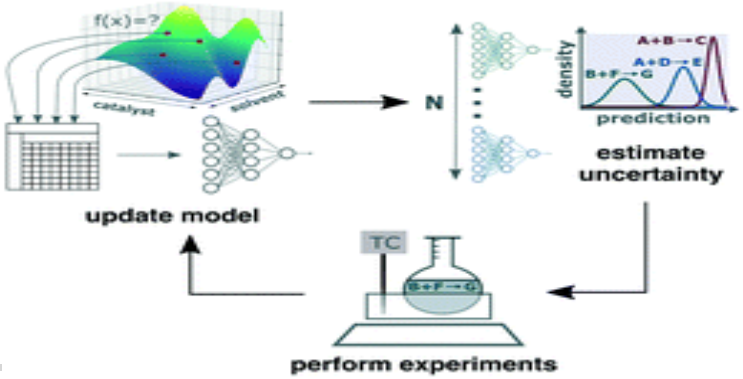
三、您對第一次捐血的感受是：(可複選)□有點緊張□非常緊張□有點不舒服□非常不舒服□很好□沒任何感覺□其他_____

四、您往後會持續捐血？□是□否□不確定；會鼓勵您的親朋好友捐血？□是□否□不確定

五、請就以下項目對捐血人的重要性及您對本次捐血過程的滿意度，加以勾選

項目	很不重要	不重要	普通	重要	很重要	很不滿意	不滿意	尚可	滿意	很滿意
1. 捐血地點是否容易尋找	□	□	□	□	□	□	□	□	□	□
2. 捐血步驟及注意事項於捐血地點是否有清楚的標示	□	□	□	□	□	□	□	□	□	□
3. 捐血環境是否整潔舒適	□	□	□	□	□	□	□	□	□	□
4. 進入捐血地點時，是否馬上有人員招呼	□	□	□	□	□	□	□	□	□	□

https://pubs.rsc.org/en/content/articlelanding/2020/re/d0re00232a



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。

實驗設計與因果關係

- 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。

https://pansci.asia/wp-content/uploads/2022/03/2-6-1020x574.jpg

關於臨床試驗

https://www.quora.com/What-is-a-double-blind-procedure

- 實驗組→處理；處方

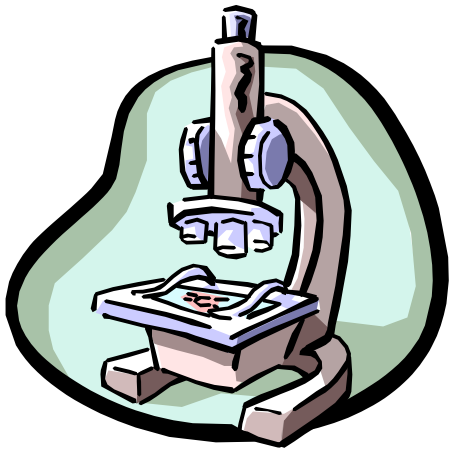
  對照組→安慰劑(Placebo)

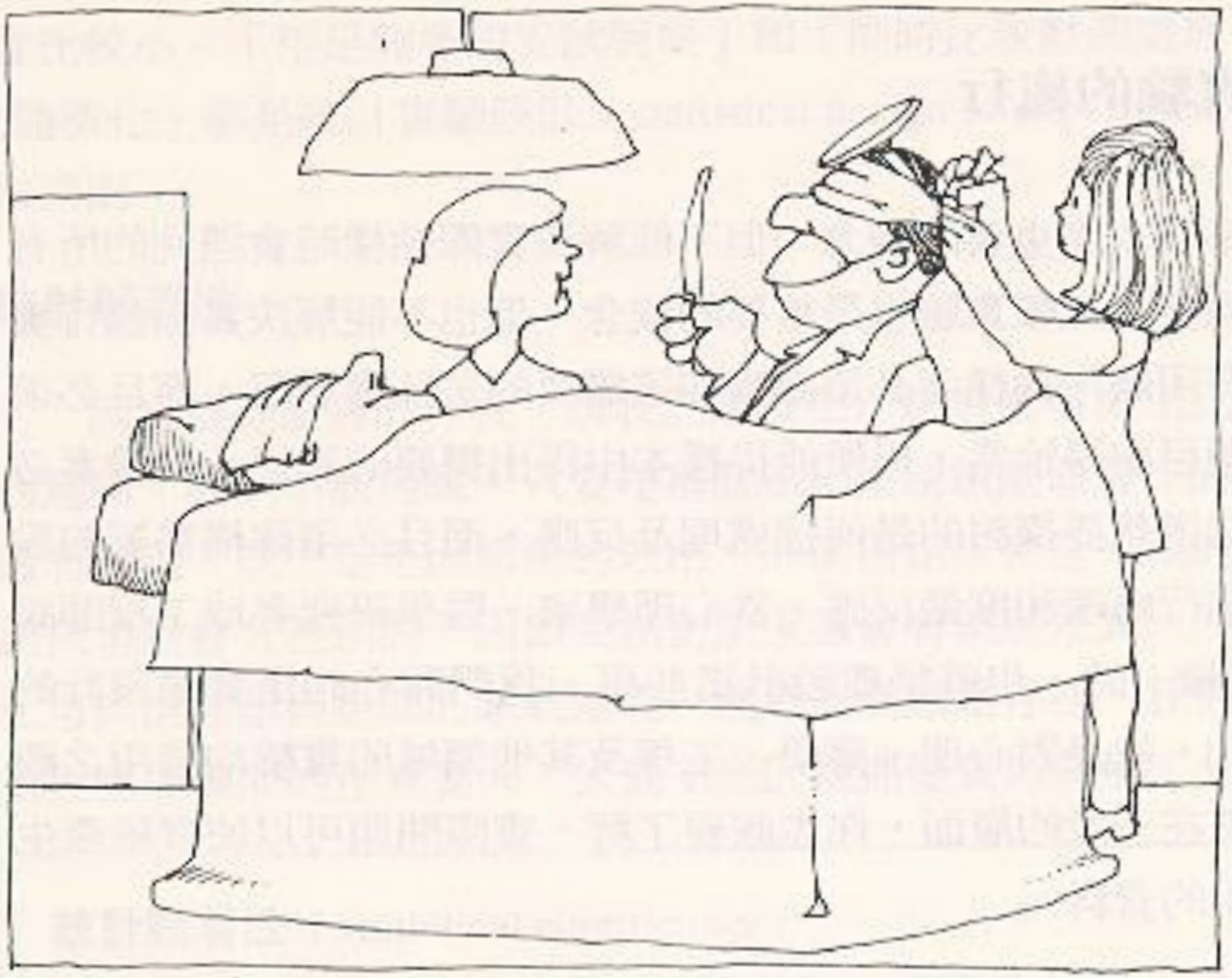
- 單盲與雙盲實驗：

→ 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

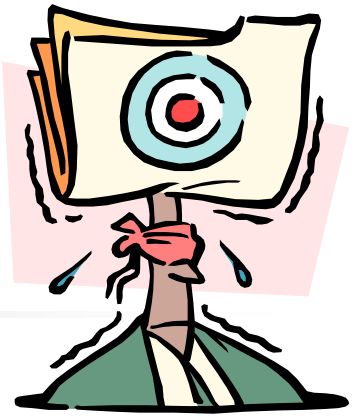
處方的分配





「伯恩斯醫師，您確定統計學家說的雙盲實驗是這個意思嗎？」

實驗成本以外的考量



- 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。

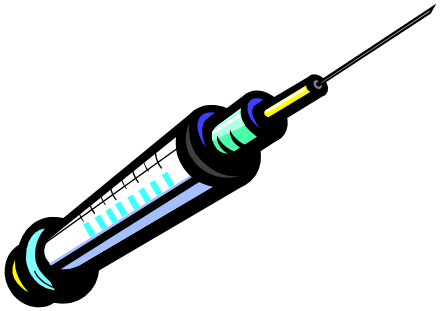
→例如：各國蒐集該國罹患SARS、AIDS等疾病，希望找到共同的特性；選舉研究如何整合不同地區及時間得出的電訪結果，以獲得當前選舉的趨勢。

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：

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

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

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