

一、緒論 (INTRODUCTION) (Chapter 1 & 2)

劉仁沛教授

國立台灣大學農藝學研究所生物統計組

國立台灣大學流行病學與預防醫學研究所

國家衛生研究院生物統計與生統資訊組

jpliu@ntu.edu.tw



【本著作除另有註明，網站之內容皆採用 創用CC 姓名標示-非商業使用-相同方式分享 3.0 台灣 授權條款釋出】

緒論

- 統計學（Statistics）
 - 收集資料
 - 分析資料
 - 解釋意義
- 研究工具與資訊應用的科學

緒論

- 應用領域
 - 生物：生物統計（Biometry）
 - 工業：工業統計（Industrial Statistics）
 - 醫學：生醫統計（Biostatistics）
 - 遺傳：遺傳統計（Genetic Statistics）
 - 基因體：基因體統計（Genomic Statistics）
 - 基因序列：生物資訊（Bioinformatics）及計算生物學（Computational Biology）
 - 心理學：心理統計（Psychometrics）
 - 經濟：經濟計量（Econometrics）

緒論

- 收集資料方法
 - 試驗設計 (Experimental Designs)
 - 抽樣方法 (Sampling Methods)
- 分析資料方法
 - 資料種類
 - 連續性資料 (Continuous Data)
 - 類別資料 (Categorical Data)
 - 設限資料 (Censored Data)

緒論

- 連續性資料
 - 分立資料 (Discrete Data)
 - 心跳
 - 細菌在 1 C.C. 中的個數
 - 某種鳥在一平方公里的個數
 - 連續資料
 - 年齡
 - 血壓
 - 肝功能指數

緒論

- 類別資料
 - 質的類別資料 (Attribute)
 - 性別
 - 種族
 - 品種
 - 順序類別資料 (Ordinal)
 - 症狀
 - 無 輕微 中等 嚴重
 - 滿意度
 - 非常不滿意 不滿意 無意見 滿意 非常滿意

緒論

- 設限資料
 - 種子發芽的時間
 - 產品的壽命
 - 癌症病人存活時間
- 在五年內不是每個癌症病人均死亡所以不是每人癌症病人的存活時間均可觀測到
- 在五年內未死亡的癌症病人存活時間至少為五年☒其資料設限於五年

緒論

- 敘述統計學 (Descriptive Statistics)
 - 將蒐集的資料整理為簡單的數據 (如平均數或百分比) 或圖表，以說明此資料的統計學
- 推論統計學 (Inferential Statistics)
 - 以樣本資料的結果推論族群 (母體) 該資料的全體性質的統計學

緒論

- 族群 (母體, Population)
 - 一般性質相同事物所測量觀測值的全體資料, 其中最基本單位稱作元素 (Element)
 - 例子
 - 台大學生的體重
 - 某牛奶工廠所產生鮮奶每瓶之容量
 - 台北市 30 至 40 歲工作人口之收入
 - 目標族群或研究族群 (Targeted or Study Population)- 研究對象的族群
 - 例子
 - 台灣的全體居民

緒論

- 樣品（樣本 Sample）
 - 族群中一小部份的觀測值
- 代表性樣品（Representative Sample）
 - 樣本結構及特性與族群相同
- 隨機樣品 (Random Sample)
 - 族群中每個觀察值獨立且機會均等地被選取的樣品

好樣本的特性

- ☒ 母體有定義
- ☒ 母體裡面的每一個個體都有被抽選機會
- ☒ 樣本是母體的縮影
- ☒ 樣本具有代表性

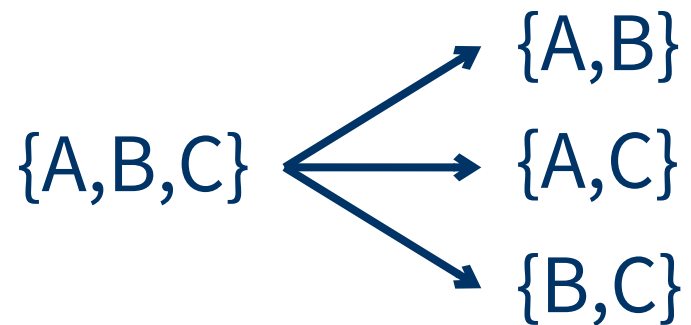
抽樣設計的基本概念

1. 決定調查的母體
2. 從母體中抽取樣本
3. 確保樣本能正確的代表母體

目標：樣本是母體的縮影

緒論

- 不歸還抽樣法 (Sampling without Replacement)
 - 抽出的元素不歸還族群後再抽取下一個元素



緒論

- 歸還抽樣法 (Sampling with Replacement)
 - 抽出的元素歸還族群後再抽取下一個元素

$\{A, B, C\} \longrightarrow \begin{array}{l} \{A, A\}, \{B, A\}, \{C, A\} \\ \{A, B\}, \{B, B\}, \{C, B\} \\ \{A, C\}, \{B, C\}, \{C, C\} \end{array}$

緒論

- 抽取機率

	<u>元素</u>	<u>樣品</u>
不歸還抽樣法	$1/3$	$1/3$
歸還抽樣法	$1/3$	$1/9$

抽取好樣本的方法

- ☒ 簡單隨機抽樣 (Simple Random Sampling)
- ☒ 系統抽樣 (或稱等距抽樣)(Systematic Sampling)
- ☒ 分層隨機抽樣 (Stratified Sampling)
- ☒ 集體抽樣 (Cluster Sampling)

簡單隨機抽樣 (Simple Random Sampling)

簡單隨機抽樣：母體裡面每一個個體都有相同的機率被抽選成為樣本

例如：調查台北市某個社區內的一千戶居民每個月自來水的平均用量，隨機挑選其中10戶進行調查

Excel 建立的隨機亂數表

列							
1	05185	43887	18753	00311	94100	35566	16430
2	87995	14282	78996	32280	30600	92256	87427
3	53765	31248	59731	64065	19196	34096	18851
4	74286	77413	80564	19953	43834	29275	17835
5	51866	18117	39231	78767	21008	82150	14765
6	42556	15969	13194	90604	90822	74266	44568
7	17392	16653	40766	68158	96706	93531	18871
8	54424	51373	89873	91075	69086	31522	95614
9	65630	71684	74375	39613	17960	30618	85780
10	16333	68856	79164	21804	53523	18127	87763
11	59005	40282	11900	03934	98586	32059	32526
12	67512	30406	80866	85616	77086	70590	21962
13	25919	40719	78240	71439	02121	68917	66949
14	95552	33071	03501	16466	46385	19839	18546
15	87783	39335	37777	12269	35192	56460	00033
16	71606	53549	49756	63348	10379	84045	67438
17	83320	63004	89476	88550	74555	86302	32738
18	59791	97981	11/15/2022	77174	08813	54717	90860
19	37262	29207	10036	34405	63685	50424	84057

簡單隨機抽樣 (Simple Random Sampling)

- 利用隨機數字進行簡單隨機抽樣
 - 自有 120 位新生的班上抽取 5 位同學，參加與校長座談
 - 先將 120 位新生自 1 至 120 編號
 - 取隨機亂數表中任二列，因有 120 位新生，故取連續 3 個數

字代表一組編號

- 忽略大於 120 的數字，取出前 5 個小於或等於 120 的數字

列 1 051 854 388 718 753 003 119 410 035 566 164

列 2 879 951 428 278 996 322 803 060 092 256 874

→ 3, 35, 51, 60, 與 119 號參加與校長座談

11/15/2023

Jen-pei Liu, PhD

系統抽樣 (Systematic Sampling)

系統抽樣：每間隔固定的數選取一個樣本

例如：從五百張訂單中抽取十張訂單作為樣本，
自第一列隨機抽取一個小於等於 500 的數字，

第一個選定的樣本是第 51 號的訂單，
之後每隔五十號 ($500/10$) 選取一個調查樣本，分別為編號

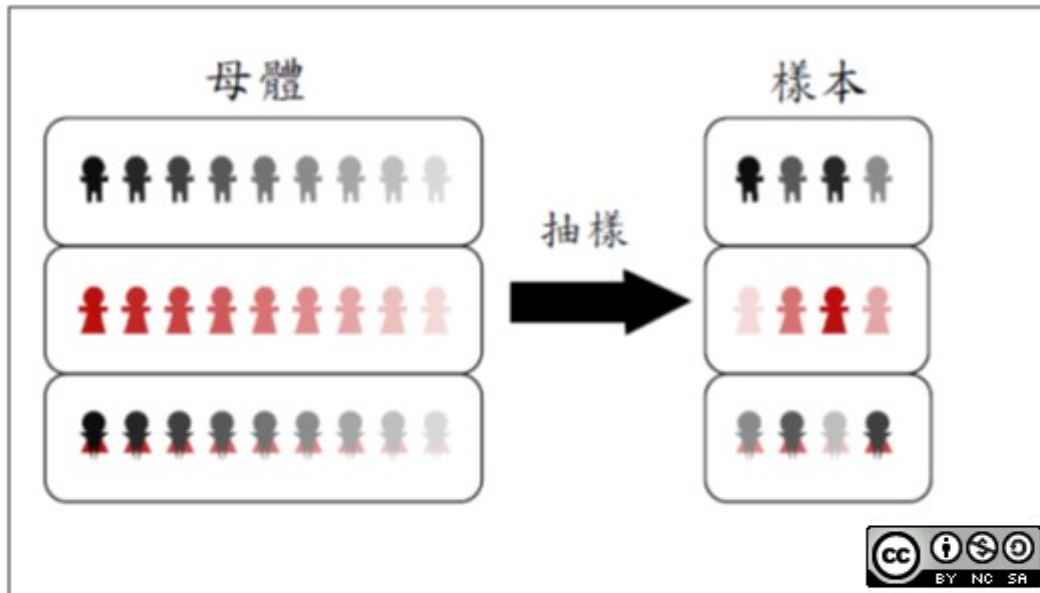
051、101、151、201、251、
301、351、401、451、001 等十張訂單

分層隨機抽樣 (Stratified Sampling)

分層隨機抽樣：根據母體之中每一個個體的特性，分成幾種類型，稱為「層」，然後每層再用機率抽樣方式抽取樣本

例如：估計某個小學全體學生的平均身高，首先根據性別分成男、女兩層，然後每層中各自隨機選取樣本。

分層隨機抽樣 (Stratified Sampling)



原因：可得各層訊息，而且樣本分配較均勻，
提高估計準確度

集體抽樣 (Cluster Sampling)

集體抽樣：先將母體區分為許多個不同的集體，然後隨機抽取少數集體當成樣本，中選的集體全部調查

例如：某市教育局想瞭解三年級數學新教學方法的成效，在該市隨機抽取十個小學當成樣本學校。其中五個小學採用新教學方法，另外五個小學採用目前教學方法。上完一學年後，使用同一份試卷進行測試與比較。

實例應用：2001 年國民健康訪問調查

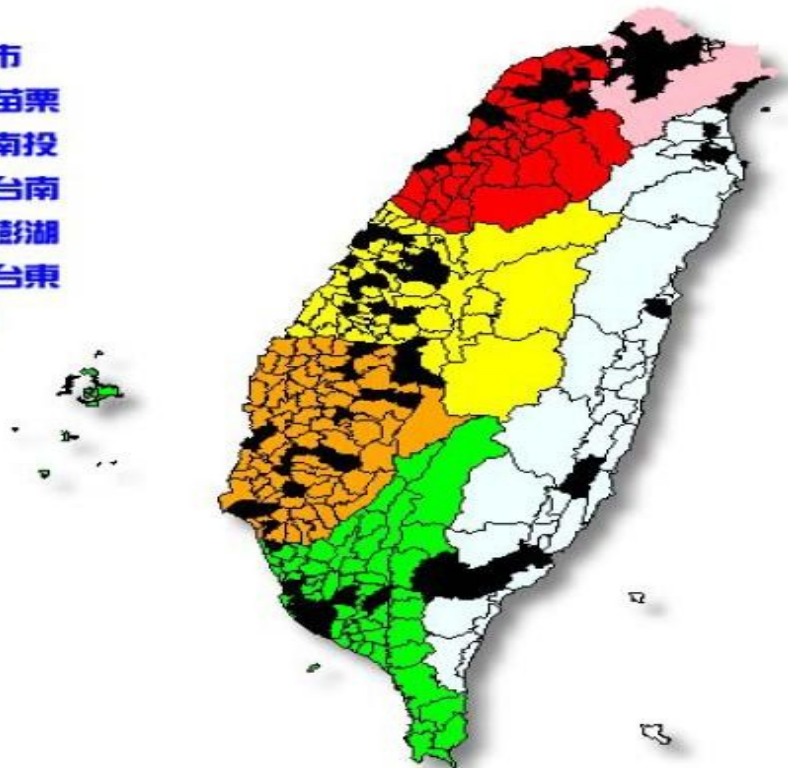
母體：2000 年臺灣地區全體人民

樣本：採用多段分層抽樣設計

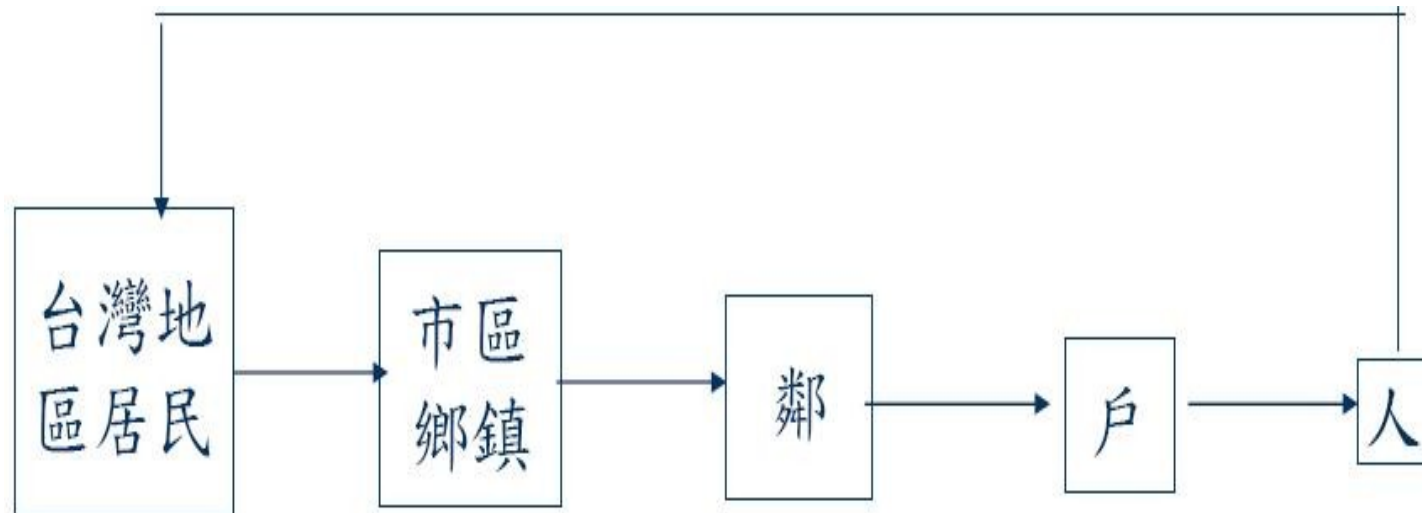
目的：瞭解國人健康狀況與醫療服務利用情形

國民健康訪問調查：分層隨機抽樣

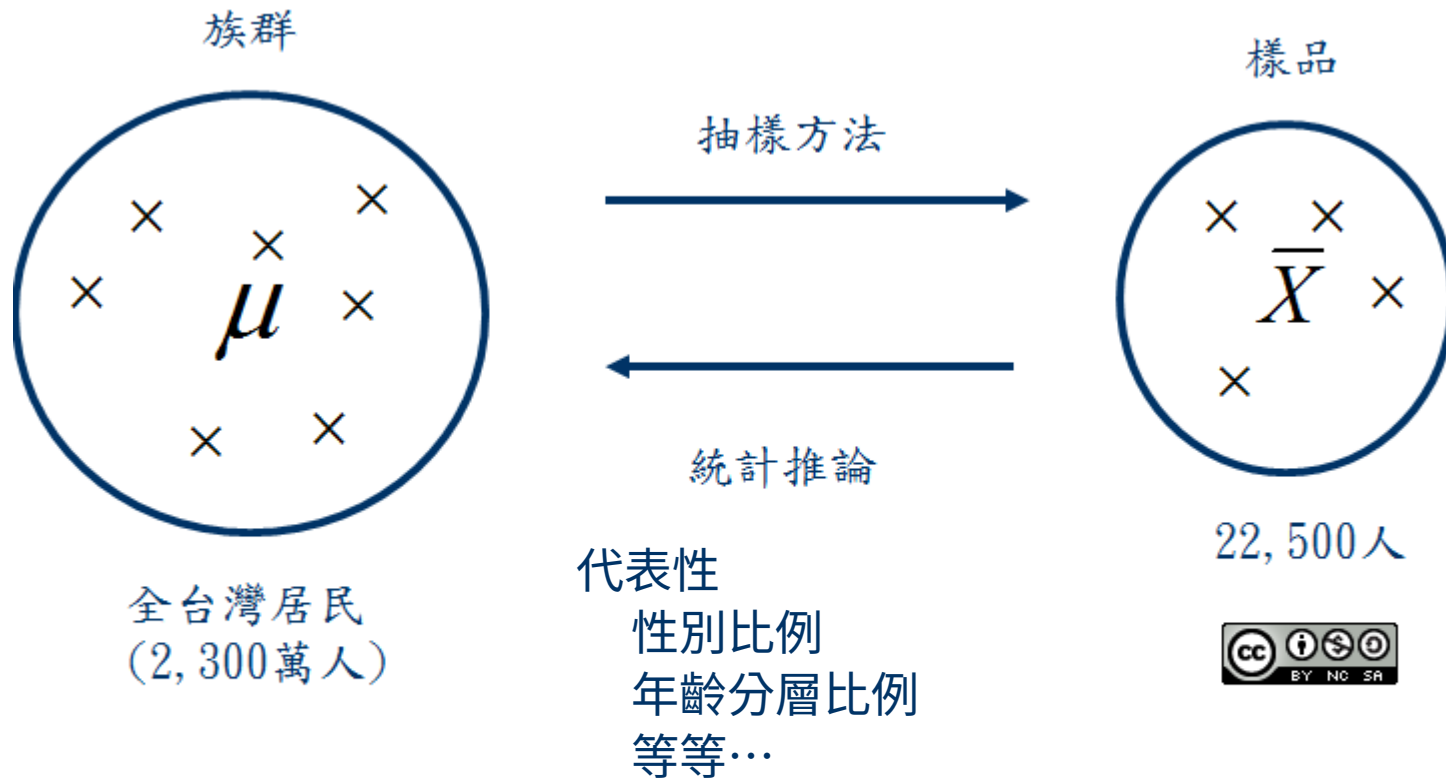
- 大台北地區
- 台北縣、基隆市
- 桃園、新竹、苗栗
- 台中、彰化、南投
- 雲林、嘉義、台南
- 高雄、屏東、澎湖
- 宜蘭、花蓮、台東
- 樣本鄉鎮市區



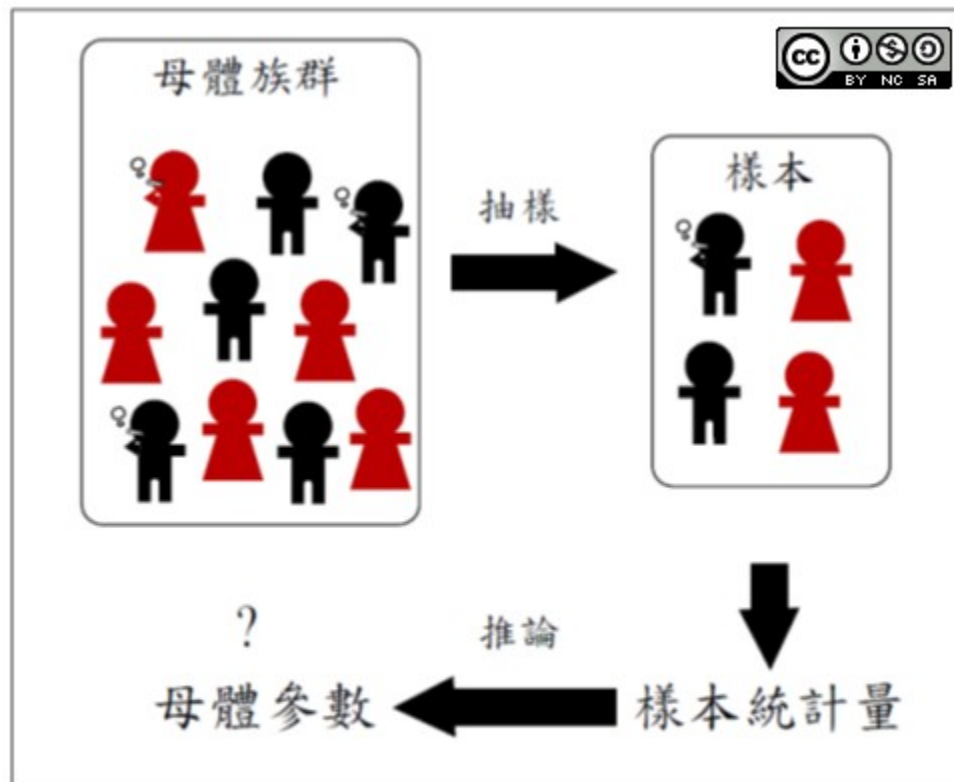
國民健康訪問調查：樣本選取



國民健康訪問調查



母體參數 VS 樣本統計量



緒論

- 母數 (參數, Parameters)
 - 族群全部資料所計算出的數據
 - 例子：
 - 全台灣人口女性的人數及比例
 - 全台北市工作人口的平均收入
 - 族群母數通常以希臘字母代表之
 - 如 μ 代表族群平均數
- 統計值 (Statistics)
 - 樣品資料計算所得的數據，通常以英文字母代表之
 - 如 \bar{x} 代表樣品算數平均數

資料衡量

三聚氰胺值

瘦肉精值

H1N1 病毒量

台灣抽煙的比例

癌症病人的生活品質

身體

心理

社會

福祉

量測值



緒論

- 系統誤差 (Systematic Error) :
也稱偏差 (Bias) ，是測量物體時偏離族群真值的誤差，是一種有原因與方向的誤差。

資料衡量

- 效度 (Validity)：使用的測量工具有沒有效？
測量工具是否可測得要測的真值？
 - 生活品質問卷是否可測得受訪者的生活品質？
 - 溫度計是否可測得真正的體溫？
- 準確度 (accuracy)
- 偏差 (Bias)
- 信度 (Reliability)：每次測量的結果一致嗎？

效度

- 測量游泳池的容量？水桶 VS 皮尺
 1. 水桶是無效度的測量工具
 2. 皮尺是有效度的測量工具
- 測量地球與月球的距離
 - 用拳頭一個一個去量
 - 用皮尺一段一段去量
 - ???

緒論

- 隨機誤差 (Random Error) :
測量物體觀測值，不知原因，是偶然發生的誤差。

資料衡量

信度 (Reliability)：每次測量的結果一致嗎？
用相同的血液檢體，相同的檢測方法，量測膽固醇
五次，是否均得到相同或接近的量測值？

分散度 (Dispersion)
變異度 (Variability)

信度

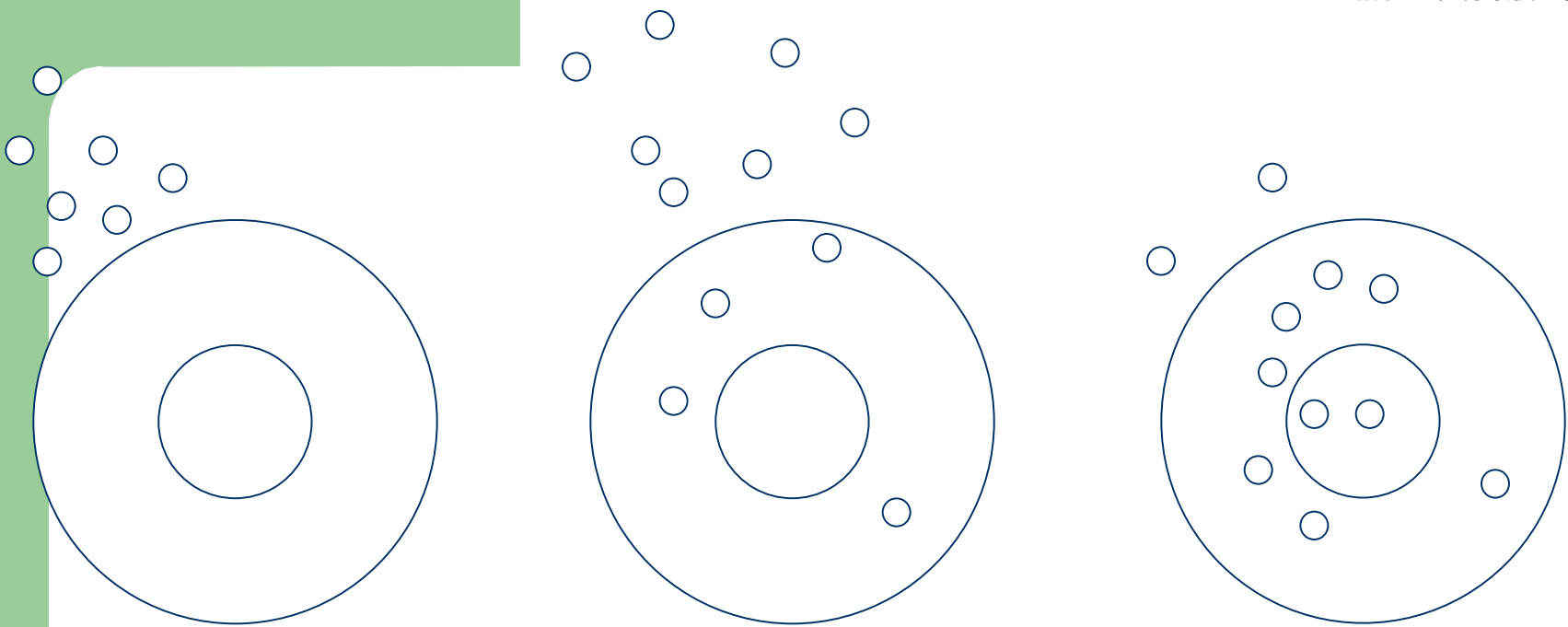
體重機：小寶家 (有信度) VS 小強家 (沒信度)

小寶家：70.1、70.5、69.5、70.2、69.8

小強家：57.1、73.5、65.5、77.1、61.8

緒論

- 準確度 (Accuracy) :
抽樣所得樣品中觀測值靠近族群真值的程度
- 精密度 (Precision) :
抽樣所得樣品中觀測值的集中或分散的程度



樣品 A

樣品 B

樣品 C



精密度與準確度比較圖

數據的展示

1. 楊同學考試 58 分
2. 林同學每個月花費 6500 元
3. 新生女嬰重量是 3485 公克
4. 膽固醇值為 175mg/dL

單單一筆數據是毫無用處的，需要有一些參考數據再來研判，解讀數字才有意義

楊同學考試 58 分

如果全班平均 30 分，則楊同學的成績算很高
如果全班平均 80 分，則楊同學的分數不太好

林同學每個月花費 6500 元

根據調查，臺大學生每個月花費的中位數是 5500 元，林同學每個月花費比一半的同學花得多

膽固醇值為 175mg/dL

三高

- 膽固醇值大於 200mg/dL \Rightarrow 高血脂
- 飯前血糖大於 124mg/dL \Rightarrow 糖尿病
- 舒張壓 / 伸縮壓大於 130 mmHg/90mmHg
☒ 高血壓

所有檢驗值均有正常值範圍 (Normal ranges)

蒐集資料的道德規範

1. 通過倫理審查委員會的審查
2. 經過受訪對象的同意
3. 個人資料保密
4. 受訪對象的權益永遠優先

2009年7月17日，行政院通過「人體生物資料庫管理條例草案」

2010年5月26日，立法院通過「個人資料保護法」

經過受訪對象的同意

- 某研究學者，採集葛瑪蘭族原住民唾液進行研究，未完整告知研究目的及接受採檢唾液者應有的權利，違反研究倫理。
- 臨床試驗：受試者在完全告知臨床試驗的利益與風險後，自願簽署受試者同意書後，才可進行臨床試驗。

經過受訪對象的同意

- 臨床試驗：國外藥廠在台灣執行的臨床試驗，常收集國人檢體進行藥物基因體試驗，但並不將結果告知。所以國外藥廠對台灣基因的瞭解程度較我們自己更清楚。
- 國外強權的人造衛星在台灣上空飛來飛去，未知國人同意，以影像方式收集台灣各種資料，但並不將結果告知。所以國外強權對台灣的瞭解程度較我們自己更清楚。

受試同意書

受試同意書 (Informed Consent):

受試者受告知並了解將參與之臨床試驗之相關訊息，且參酌是否參與試驗之所有因素後，自願簽署參加試驗之文件。

事前自願同意並簽署

臨床試驗之受試同意書，應符合赫爾辛基宣言的規定。

除給予受試驗者或法定代理人書面資料外，並應包括口頭說明與雙向溝通，使其了解整個試驗的狀況，並有充裕的時間考慮後，再決定簽署受試同意書。

同意權之行使

- 原則：本人
- 無行為能力人：法定代理人代為
- 限制行為能力人：法定代理人同意
- 無意識或精神錯亂：有同意權人
 - 配偶、同居之親屬

見證人

- 受試者、法定代理人或有同意權之人皆無法閱讀時，應由見證人在場參與所有有關受試者同意書之討論。
- 見證人應閱讀受試者同意書與提供受試者之任何其他書面資料，以見證試驗主持人或其指定之人員已經確切將其內容向受試者、法定代理人或有同意權之人為解釋，並確定其充分了解所有資料之內容。
- 確定受試者、法定代理人或有同意權之人之同意完全出於其自由意願。
- 試驗相關人員不得為見證人。

對受試者之補助

- 試驗委託者對於受試者可獲得之補助及付款方式，不得有強迫或不當影響受試者之情形。
- 受試者之補助，應按臨床試驗進行之進度依比例給付之，不得於試驗完成後方為給付。但小金額者，不在此限。
- 受試者補助之付款方式、金額及付款進度，應載明於受試者同意書及其他給與受試者之書面資料；補助按比例分配付款之方式，應詳細說明。

個人資料保密

- 學生各項考試成績
- 老師教學的評鑑結果
- 全民健康保險研究資料庫:

國家衛生研究院在提供資料之前，一定會先把個案的資訊先做亂碼處理，因為健保局要求保護每一個個案的隱私權。

資料更改或作假

- New York Times (A. Pollack, Sept. 29, 2009)
“Biotech company fires chiefs and others over handling of data.”
“A biotechnology company developing what was expected to be a groundbreaking blood test for Down syndrome fired its chief executive, a top research official and three other employees Monday after an investigation into ‘mishandling’ of test data and results.”

資料更改或作假

- New York Times (A. Pollack, Sept. 29, 2009)
“Biotech company fires chiefs and others over handling of data.”
“The company’s shares, which had been as high as \$28 a year ago, fell sharply in late April when the company said the data could no longer be trusted. In after-hours trading Monday, the stock fell by nearly half to \$3.23.”

Duke Scandal

- Further analyses revealed corruption of multiple datasets compiled by Dr. Potti that had been used as sources of validation of the various chemotherapy sensitivity signatures. These included data derived not only from Duke sources, but also publicly available data. As an example, a dataset of 133 samples from a neoadjuvant breast cancer study at MD Anderson involving patients treated with the combined regimen TFAC was used for validation of an adriamycin signature. The clinical annotation that was assumed to be used by Dr. Potti included 34 responders and 99 non-responders, the same distribution as reported by MD Anderson. However, a detailed comparison of the two datasets revealed that the response information was reversed for 24 cases with 12 labeled incorrectly in each direction. In this case, the corrupted data yielded positive validation results whereas the accurate data did not provide evidence for validation. Similar findings of corruption of data in key validation datasets were observed in other instances.
- As a result, three publications were retracted, a manuscript describing the methods for implementing signatures in the clinical trials that was under review was removed from further consideration, and other publications are currently being analyzed. Dr. Potti issued his resignation statement on November 19, 2010, and a statement of responsibility for the problems with the work. A research misconduct investigation is in progress.

解讀統計數字

極富聲望的〈科學〉（Science）期刊在一篇談論植物病昆害之文章中，提到加州有一塊田每英畝生產 750,000 顆哈密瓜（1 英畝 = 4046.85 平方公尺）。請問你這數據那裡出了問題？

解讀統計數字

某醫學期刊刊登了一篇論文，論文中有一個統計表，裡面有很離譜的錯，這樣的錯，聰明的小學生都可能看得出來。表裡面列有 6 組老鼠，每組有 20 隻生病的老鼠，每隻老鼠接受某種治療後，每組治癒 (成功) 的比率分別為 53% ， 58% ， 63% ， 46% ， 48% 以及 67% 。請問你這組數據那裡出了問題？

解讀統計數字

某高爾夫球俱樂部說明：在 50 歲以上的會員中有 55% 為男性，35% 為女性。請問你這組數據那裡出了問題？

解讀統計數字

某一政策滿意度調查結果如下：

非常不滿意 :5%

不滿意 :6%

無意見 :80%

滿意 :5%

非常滿意 :4%

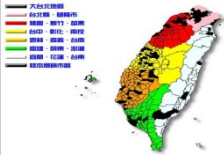

結論：民眾對政策不滿意。

請問你結論這那裡出了問題？

緒論

- 習題 P20 : 1 , 2 , 6 , 7

版權聲明

作品	授權條件	作者 / 來源
		國家衛生研究院出版的 2001 年「國民健康訪問調查」結果報告 (2003)