

統計學緒論及資料 蒐集與整理

第一章





內文教學區

一、統計的意義

利用數據變動的資訊及經過資料蒐集、整理、表現、分析和解釋，再依分析出來的結果，做出較為正確且合理的決策，並大幅降低結果之不確定性的科學方法。簡言之，統計就是將大量的資料轉化成有用的資訊。

二、統計學分類

(一) 敘述統計學

事件資料之蒐集、整理、分類和簡易計算，目的在於描述事實。這類資料的彙總可能是表列的、圖示的，或是數值的。

(二) 推論（歸納）統計學

將蒐集整理後的資料加以分析、估計、檢定和預測，目的在於求得全體之一般化結果。

三、統計學常用之名詞解釋

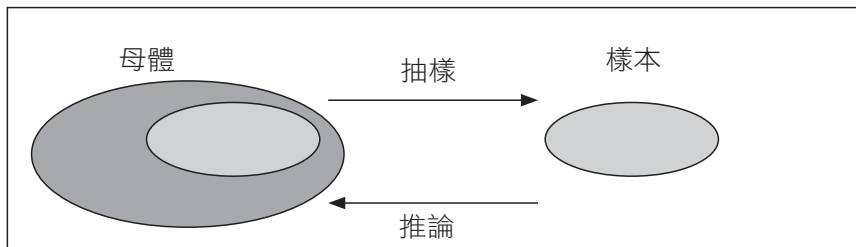
(一) 母 體

人們所要研究的對象全體稱為母體（population），它是一群具有某種共同特性的人或事物之個體所構成的群體，而每一群體通常含有多種特性，可依某種特性歸納為一母體，亦可依另一種特性歸納為另一母體。例如，台大醫院想瞭解 98 年度該院新生兒之平均體重，則 98 年度台大醫院所有新生兒所成的集合，為此研究的母體。

(二) 樣 本

樣本由抽自母體中之某些個體測量所構成。例如，從 98 年度台大醫院新生兒中隨機抽出 200 位，則此 200 位新生兒所成的集合，為此研究的樣本。

母體與樣本的關係如下圖所示。



母體與樣本之關係圖

(三) 母體參數

用來代表母體特徵的量值稱為母體參數。例如，98 年度台大醫院所有新生兒之平均體重（稱為母體平均數 μ ），為此處所謂的參數。

(四) 樣本統計量

用來代表樣本特徵的量值稱為樣本統計量。例如，從 98 年度台大醫院新生兒中隨機抽出 200 位之平均體重（稱為樣本平均數 $\bar{X}$ ），為此處所謂的統計量。

(五) 變數與變量

統計資料一般是由一個或多個變數的值所組成，所謂變數是一種具有不同值或結果之特徵的衡量，凡一切可計量的特徵都稱為變數，而變數（ X ）的值或量則稱為變量（ x ）。

(六) 量的資料與質的資料

統計資料依據其變量的特性可分為量的資料與質的資料，量的資料亦稱為屬量資料，係依據數字尺度衡量出來的資料，而質的資料亦稱為屬性資料。

(七) 類別資料

依據資料的屬性或類別尺度來區分的資料。

(八) 自變數與因變數

1. 自變數（Independent variables）：被認為是「因」（cause）或用來解釋其他變數的變數。
2. 因變數（Dependent variables）：被認為是「果」（effect）或是被解釋的變數。例如， $y=f(x)$ ，此時 x 為自變數， y 為因變數。

四、企業與統計之關係

(一) 企業決策與統計

1. 企業在做出決策之前須經過蒐集、整理、分析資料的工作，在這



些過程當中即須應用各類統計方法以確保決策方向之正確性及決策之品質。

2. 當決策者面臨數個決策方案與未來事件充滿風險或不確定性時，即可用決策分析決定最佳的策略。例如，時尚服裝製造商可能希望大量生產消費者之接受性與需求量較大的產品。

(二) 企業傑出經理人，應善用統計方法，做出正確的決策

1. 取得與問題有關的統計資料（data）。
2. 使用統計方法，得到統計資料所傳達的訊息（information）。
3. 做出正確的決策（decision）。

五、統計學未來發展與限制

(一) 統計未來發展，包括政府統計、生物醫學統計、人事管理、財務分析、工業統計（含品管及生管）、行銷研究、民意及市場調查、保險精算、經濟分析等。

(二) 統計方法的應用仍有限制：統計應用的領域隨著時代的進步日漸擴大，但統計方法的應用仍有限制，如：

1. 統計分析必須有足夠的統計資料。統計是一門科學，而不是預言家或是魔術師，它必須從「足夠多」的統計資料中尋找研究對象對群體的通則。
2. 統計資料的蒐集必須客觀與周全，否則統計的結果必有所偏差，而導致錯誤的統計推論。

六、統計工作程序

辦理統計工作的程序，包括事前計畫、蒐集資料、整理資料、分析資料、資料表現等。

七、資料的意義

一般而言，試驗依研究母群體特性的目的分為兩種：一為瞭解該特性的現況而做的試驗稱為調查；另一種則是該特性所發生的原因而做的試驗稱為實驗。做這些試驗必須經由一套具體而嚴格的方法，小心計畫並執行所得到適合且有意義的數據，這些數據就稱為資料。

八、資料分類

統計資料的種類繁多，按資料的性質，可分為質性資料與量化資料兩



種；按資料的來源，又可分為初級資料與次級資料兩種。

(一) 質化資料

其特性為不能衡量、不能用點計算的，以分類或排序為主的資料。

例如，色澤、職業等。

(二) 量化資料

其特性為可計點、可衡量，且以數值或測度為主的資料；又分：

1. 離散型資料（discrete data）：為可數的（countable）數值資料。

例如，人數。

2. 連續型資料（continuous data）：為不可數的（uncountable）數值資料。例如，長度、重量、時間。

※判斷何時用何種統計分析方法，第一步要判定：其各個變數（不論是自變數或應變數）之性質是不連續性的（discrete）或連續性的（continuous）。

(三) 初級資料

係指研究者直接由調查或試驗而蒐集得來的資料，亦稱原始資料或直接資料。例如，民意調查機構要預測選舉結果，得直接向選民詢問調查，所得的資料。初級資料蒐集不易，有賴於完善的蒐集技術與籌劃，雖然其所花費的人力、時間、金錢很多，但是蒐集得來的資料比較周詳而合用。

(四) 次級資料

係指其他私人、機關、團體所蒐集整理後的資料。亦稱現成資料，或間接資料。

1. 次級資料的來源：

(1) 政府機關所編制的統計報告。如台灣人口統計年報、進出口貿易統計月報、台北市物價統計月刊等。

(2) 企業機構所發行的統計刊物。如報紙上的證券行情表、物價行情表等。

(3) 學術團體所出版的書刊雜誌。如中國統計學報、各校的學報、品管月刊等。

(4) 個人方面的論文著作。如博士論文、心得報告、臨床經驗等。

2. 利用次級資料的優點：

(1) 既省時、省力又省錢。

(2) 可以獲得與研究問題有關的資料。



(3) 可以獲知其他相關資料的來源。

(4) 對所研究的問題有新的啟示。

九、資料的度量尺度

(一) 名目尺度

資料分類、屬性無大小之分。例如，性別、職業等。根據特徵的性質做分類，每類答案的代表數字只作為分類用，目的在求分類（此類資料可計算眾數）。

(二) 順序尺度

資料有順序大小之分，但無差異程度。例如，名次、喜好程度等。在順序尺度中只能看出高低次序，但無法確定各類別之間的差距（此類資料可計算中位數）。

(三) 區間尺度

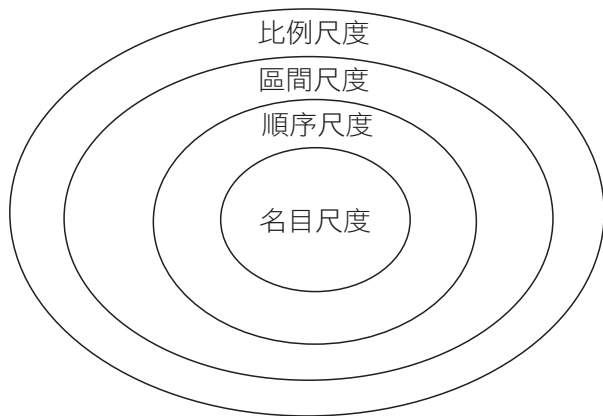
資料有大小、差異程度、無倍數等關係。兩個元素之資料值之間的比較，不僅可以表示順序的關係，尚能測量各順序位置的距離。另外，區間尺度沒有自然原點 0。例如，溫度（此類資料可計算眾數、中位數、算術平均數）。

(四) 比例尺度

資料有大小、差異和倍數關係，且比例尺度必包含零值。例如，身高、體重（此類資料可計算眾數、中位數、算術平均數）。

※資料所含的資訊量因衡量尺度而異。資訊量由小排到大的順序是：名目>順序>區間>比例資料。區間與比例資料的算術運算為有意義，名目與順序資料的算術運算沒有意義。

※資料的四種度量尺度關係如下圖所示：



**【例題 1】**

試說明以下變數之測量方法屬於何種測量尺度？

- (A) 紅色 (0)，白色 (2)，藍色 (4)
- (B) 非常滿意 (2)，滿意 (1)，普通 (0)，不滿意 (-1)，非常不滿意 (-2)
- (C) 薪水 (以元為單位)
- (D) 智商 (IQ 表示)

【解答】

(A) 名目尺度；(B) 順序尺度；(C) 比例尺度；(D) 區間尺度。

十、資料的蒐集

統計工作所依據的全是統計資料，因此蒐集的資料是否客觀而周詳，關係整個統計工作的成敗。在蒐集統計資料時，須注意以下幾點：確定調查的對象、時間及範圍，選定調查項目、實施方案、訓練與督導調查員等。例如，國家的人口普查，不適合在冬季的耶誕節與過年期間舉行。

資料的蒐集來源，包括：

- (一) 已存在資料（次級資料）。
- (二) 調查而得的資料（初級資料）。

十一、資料的整理與表現**(一) 資料整理**

統計資料蒐集得來之後，應該審慎而周詳的加以審核。如果發現資料中有錯誤、遺漏、殘缺、矛盾，或其他可疑的地方，應該迅速設法複查並加以更正。例如，蒐集來的資料單位不一致時，應加以換算統一，不能用的資料應捨棄；審核後的統計資料還是必須依照研究目的加以整理，以求簡化，並藉以顯示研究對象全體的通則。

(二) 資料表現

資料經過整理分析後得到一些統計結果，為了便於說明起見，我們常將這些結果用統計表或統計圖陳示出來，或另編統計報告詳加說明，以供各有關方面參考使用。如此統計工作程序即告完成。

十二、定量資料的統計圖表**(一) 百分比、比例與比值**

1. 百分比 (percentage)、比例 (proportion)：百分比與比例都是一



種將資料標準化的做法。百分比是以 100 為基底來標準化，比例則是以 1.00 為基底。比例及百分比可用在名目尺度及等級尺度測量的變項。

2. 比值 (Ratio) :

$$\text{Ratio} = f_1 / f_2$$

f_1 : 第一個類別或項目中之件數 ; f_2 : 第二個類別或項目中之件數。

(二) 次數表的編製

次數表 (frequency table) 是將所蒐集的原始資料，依照數值的大小順序分成若干組，並統計每一組的次數，主要的目的在於顯示原始資料的次數分配 (frequency distribution)。步驟如下：

Step 1. 由小到大排序 (sorting)

Step 2. 求全距 (range)

Step 3. 定組數及組寬 (距) (class interval, class width)

令 k 為組數，而 n 為資料總數。

Sturge's Rule : $k = 1 + 3.322 \log n$ 或 $k = 1 + \log_2 n$

Step 4. 定組限 (class limit)

1. 所有資料中的最小值即為第一組之下限。
2. 組距 = 相鄰兩組下限的差。
3. 每一組的上限與次組的下限相差一個基本單位。

Step 5. 定組界 (class bound)

1. 每一組之下界 = 該組之下限 - $\frac{1}{2}$ 基本單位
2. 每一組之上界 = 該組之上限 + $\frac{1}{2}$ 基本單位

Step 6. 求組中點

$$\text{組中點} = \frac{1}{2} (\text{上限} + \text{下限}) = \frac{1}{2} (\text{上界} + \text{下界})$$

Step 7. 劃記各組次數並完成次數表 (frequency table)

Step 8. 累加次數表 (cumulative frequency table)

若 f_i 代表第 i 組的次數，則第 i 組的 (以下) 累加次數為所有比第 i 組上界小的數值資料之總次數。即：

$$\text{第 } i \text{ 組的 (以下) 累加次數} = \sum_{j=1}^i f_j = f_1 + f_2 + \cdots + f_i$$



Step 9. 相對次數表 (relative frequency table)

$$\text{第 } i \text{ 組相對次數} = \frac{\text{第 } i \text{ 組次數}}{\text{總次數}} = \frac{f_i}{n}$$

Step 10. 累加相對次數表 (cumulative relative frequency table)

$$\text{第 } i \text{ 組的累加相對次數} = \sum_{j=1}^i \frac{f_j}{n}$$

※1. 組界的用途：

- (1) 做直方圖。
- (2) 求分位數、眾數與稱「四分位差」。

2. 組中點的用途：

- (1) 做次數多邊圖。
- (2) 求平均數、變異數、標準差與變異係數。

【例題 2】

某飯店西餐廳 20 天晚餐結帳金額 (單位：千元) 如下：

32.65 26.12 42.90 34.26 42.00
 24.10 29.86 39.40 38.75 72.00
 28.40 34.50 69.80 64.45 61.81
 36.62 46.12 53.00 53.06 49.42

試編製次數及相對次數分配表 (分 7 組)。

【解答】

全距 = 72.00 - 24.10 = 47.90，分 7 組，所以取組距 = 7。

組 別	組 限	組 界	組中點	劃 記	次 數	相對次數
1	24.00-30.99	23.995-30.995	27.495	\\ \\ \\	4	0.2
2	31.00-37.99	30.995-37.995	34.495	\\ \\ \\	4	0.2
3	38.00-44.99	37.995-44.995	41.495	\\ \\ \\	4	0.2
4	45.00-51.99	44.995-51.995	48.495	\\	2	0.1
5	52.00-58.99	51.995-58.995	55.495	\\	2	0.1
6	59.00-65.99	58.995-65.995	62.495	\\	2	0.1
7	66.00-72.99	65.995-72.995	69.495	\\	2	0.1
合 計					20	1



【例題 4】

假設下列為民國 97 年台灣地區部分連鎖業平均每坪營業額（單位：萬元/坪），請畫出莖葉圖。

83	45	58	62	23	54	43	39	30
82	17	50	11	33	26	29	69	11
42	53	29	50	43	30	27	14	20
78	37	26	13	89	49	11		

【解答】

莖葉圖如下：

莖	葉
1	111347
2	0366799
3	00379
4	23359
5	00348
6	29
7	8
8	239

實力養成區

一、選擇題

- () 1. 台北市某國小校長想瞭解該校中年級（三、四年級）學生平均每個人可以背誦出多少首唐詩，於是從中年級的名冊中隨機抽出 100 個學生，抽出結果三年級有 48 位，四年級有 52 位，然後記錄這 100 個學生每個人可以背誦的唐詩數目。請問下列何者是此研究的母體（population）？
- (A) 全台灣所有國小中年級學生
(B) 台北市所有國小中年級學生
(C) 該校所有中年級學生
(D) 被抽中背誦唐詩的這 100 個學生
(E) 被抽中背誦唐詩的這 100 個學生及他們的兄弟姐妹



- () 2. 下列敘述何者為真？
- I. 一個統計問題欲研究對象之所有份子所成的集合，稱為母體 (population)。
 - II. 統計量 (statistic) 是母體資料的特性值。
 - III. 參數 (parameter) 是樣本資料的特性值。
- (A) 僅有 I 為真
(B) 僅有 I 和 II 為真
(C) 僅有 I 和 III 為真
(D) 僅有 II 和 III 為真
(E) I、II 與 III 皆為真
- () 3. 下列調查結果何者屬於敘述統計？
- (A) 由於香蕉生產過剩，預計今年香蕉的平均價格將跌到每台斤不到 5 元台幣
 - (B) 由於政府施行禁煙規定，今年二月份的香煙銷售量較去年同期減少了 5%
 - (C) 由於全球金融風暴，紐約地區今年失業率將上升 6%
 - (D) 由於近年少子化的影響，預計自 2010 年開始，台灣的人口將呈現負成長
- () 4. 下列調查結果何者屬於推論統計？
- (A) 因應政府拯救房市方案，今年初各大銀行紛紛調降房屋貸款利率，平均利率下降了 2 個百分點
 - (B) 由於政府去年鐵腕施行掃毒行動，今年上半年的犯罪率較去年同期減少了 5 個百分點
 - (C) 由於台灣地區鰻魚生產過剩，我們預計今年外銷鰻魚的平均價格將跌到每台斤不到 100 元台幣
 - (D) 於全球金融風暴，台灣地區今年一月份的失業率較去年同期上升了 0.6 個百分點
- () 5. 下列何者為樣本統計量 (statistic) 的觀察值？
- I. 台灣地區全體大學生近視比率的抽樣調查結果為 90%。
 - II. 2004 年台北市 20-29 歲產婦，新生兒出生平均體重，普查結果為 3119.6 公克。
 - III. 大港國小 92 學年度全體二年級男生的平均身高為 127.86 公分。
 - IV. 95 年台北市市長選舉，TVBS 民調中心 10 月 17 日的調查結果，郝龍斌的支持度為 44%。
 - V. 會二 B 統計學期中考，全班學生平均成績為 88 分。
- (A) II 與 III 皆是
(B) III 與 IV 皆是
(C) I 與 IV 皆是



- (D) I、II 與 IV 皆是
(E) II、III 與 V 皆是
- () 6. 下列何者為母體參數？
(A) 有五萬名網友參加的網路民調結果
(B) 內政部每年公布的台灣地區新生兒出生數
(C) 行政院主計處調查顯示，上個月台灣地區失業率為 4.91%
(D) 衛生局抽查冰店紅豆冰的不合格率
- () 7. 上學時使用的交通工具，可選擇搭公車 = 1，騎機車 = 2，其他 = 3，這是何種量度尺度？
(A) 順序尺度 (B) 區間尺度
(C) 名目尺度 (D) 以上皆非
- () 8. 「溫度」在量度尺度分類中屬於下列那一種？
(A) 名目尺度 (nominal scale) (B) 順序尺度 (ordinal scale)
(C) 區間尺度 (interval scale) (D) 比率尺度 (ratio scale)
- () 9. 「王建民的球衣號碼」在量度尺度分類中屬於下列那一種？
(A) 名目尺度 (nominal scale) (B) 順序尺度 (ordinal scale)
(C) 區間尺度 (interval scale) (D) 比率尺度 (ratio scale)
- () 10. 某汽車公司想比較在去年與今年，公司銷售的輕型卡車及轎車的數目。下列何種圖形最為恰當？
(A) 兩直方圖 (B) 兩長條圖
(C) 莖葉圖 (D) 圓餅圖
- () 11. 下列何者不是莖葉圖的優點：
(A) 提供原始數值及其取得條件
(B) 較容易製作，就算有數值遺漏或重複，可立即補足或刪除，不必重新繪製
(C) 便利閱讀
(D) 容易讀出各分組的百分比
- () 12. 下列何者不是長條圖的特性：
(A) 各長條依其長度而區別，與其寬度無關
(B) 各長條間應留空隙，勿連接在一起，以利區分
(C) 各長條依某一變數特性的順序排列，使其易於分析，並求美觀
(D) 各長條間連接在一起，以求美觀
- () 13. 調查化妝品市場消費者職業別，可用下列何種尺度來衡量？
(A) 名目尺度 (nominal scale) (B) 順序尺度 (ordinal scale)
(C) 區間尺度 (interval scale) (D) 無法判斷
- () 14. 統計資料依其特性可分為屬量資料 (quantitative data) 與屬性資料 (qualitative data)，下列敘述有幾個為真？



- I. 棒球選手之制服背號，是屬性資料。
 II. 學生之學號，是屬性資料。
 III. 中華隊在亞運比賽，獎牌數名列第七，可知名次是屬量資料。
 IV. 馬拉松賽選手跑完全程所花費時間，是屬量資料。
 (A) 有 0 個為真 (B) 有 1 個為真
 (C) 有 2 個為真 (D) 有 3 個為真
 (E) 有 4 個為真

►►【解析】

中華隊在亞運比賽，獎牌數名列第七，可知名次是屬性資料。

- () 15. 欲將 250 個學生成績分組，若依 Sturge's Rule，則分組組數約為：
 (A) 7 組 (B) 8 組
 (C) 9 組 (D) 10 組
 (E) 11 組

►►【解析】

組數 $k \approx 1 + 3.322 \log 250 \approx 9$ 。

- () 16. 「阿強在 1500 公尺賽跑中得到冠軍」句中的「冠軍」，是何種量度尺度？ (93 初等)
 (A) 順序尺度 (B) 區間尺度
 (C) 名目尺度 (D) 以上皆非
- () 17. 「阿強在 1500 公尺賽跑中跑了 3 分 58 秒」句中的「3 分 58 秒」，是何種量度尺度？ (93 初等)
 (A) 順序尺度 (B) 區間尺度
 (C) 名目尺度 (D) 以上皆非

二、計算題

1. 下列資料表示一家雜貨店在過去兩個月中每日的顧客數：

24	59	80	33	56	75	51	71	81	37	79	78	80	96	40
66	91	86	56	75	53	10	65	74	79	24	81	97	82	66
42	70	84	55	63	73	87	63	73	44	59	54	88	77	83
49	83	91	16	78	35	66	18	81	70	73	64	79	86	60

請將此資料分成 9 組，最小一組由 10 開始。

試做一統計表，內容包括組別、組限、組界、次數、相對次數、以下累加次數、以上累加次數。