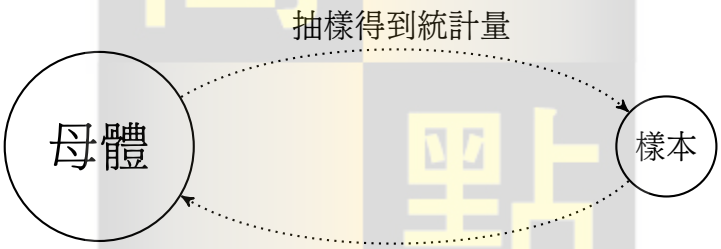
1.1 概論

性質 1.1 統計學 (*statistics*)

統計學是資料科學 (*data science*), 是在資料中萃取出重要資訊的學科。美國華裔統計學家吳建福將統計工作描述爲資料收集、資料建模和分析以及決策制定。以下是幾個重要名詞:

- 母體 (*population*): 所有可能觀察到的樣本形成的群體
- 樣本 (*sample*): 在母體中, 可觀察到的一小部份
- 抽樣 (*sampling*): 從母體中選擇出樣本的過程
- 推論 (*inference*): 利用抽取出的樣本刻劃母體特性的過程
- 參數 (*parameter*): 描述母體資料特性的數值
- 統計量 (*statistics*): 描述樣本特性的數值

1. 統計學意即利用樣本**推論**母體, 其架構圖示請參見下圖:



2. 利用統計量推論母體參數, 稱爲統計推論或推論統計學 (statistical inference)。
3. 對於蒐集到的資料, 利用簡單的表格、圖型或者數值呈現資料的型態, 稱爲敘述統計學 (descriptive statistics)。而根據資料的特性, 選擇適當的圖表幫助用戶理解資料之重要資訊的過程稱作資料視覺化 (data visualization)。
4. 討論統計學背後的數學理論基礎, 稱作數理統計學 (mathematical statistics)。
5. 利用統計方法, 對經濟問題或理論進行實證研究稱作計量經濟學 (econometrics)。
6. 例如有興趣的母體參數爲全臺灣所有人民的平均薪資所得, 但受限於預算與時間的考量下無法進行普查。因此若想要推論平均薪資所得, 可以隨機在街上訪問 100 位民衆他們的薪資所得。這時這 100 組資料, 即是母體資料 (全臺灣所有人民的各別薪資所得) 中的一組隨機樣本。這 100 組資料所計算的樣本平均, 即爲統計量。

例題 1.1 統計學的基本觀念

1. The process of using sample statistics to draw conclusions about true population parameters is called

   (A) statistical inference.
   (B) the scientific method.
   (C) sampling.
   (D) descriptive statistics.

2. Which of the following is most likely a parameter as opposed to a statistic?

   (A) the average score of the first five students completing an assignment.
   (B) the proportion of females registered to vote in a county.
   (C) the average height of people randomly selected from a database.
   (D) the proportion of trucks stopped yesterday that were cited for bad brakes.

3. Which of the following is not an element of descriptive statistical problems?

   (A) Identification of patterns in the data.
   (B) The population or sample of interest.
   (C) Tables, graphs, or numerical summary tools.
   (D) An inference made about the population based on the sample.

4. A study is under way in Yosemite National Forest to determine the adult height of American pine trees. Specifically, the study is attempting to determine what factors aid a tree in reaching heights greater than 60 feet tall. It is estimated that the forest contains 25,000 adult American pines. The study involves collecting heights from 250 randomly selected adult American pine trees and analyzing the results. Identify the population from which the study was sampled.

   (A) The 250 randomly selected adult American pine trees.
   (B) The 25,000 adult American pine trees in the forest.
   (C) All the adult American pine trees taller than 60 feet.
   (D) All American pine trees, of any age, in the forest.

《解》

1 選 (A)。統計推論即爲利用樣本統計量推估母體參數。

2 選 (B)。一個區內的投票人中的女性比例是固定的參數。

3 選 (D)。統計推論不屬於敘述統計學的範圍中。

4 選 (B)。此研究的抽樣母體爲 Yosemite 內的 25000 棵松樹。

性質 1.2 蒐集資料的方式

1. 調查 (*survey*): 不控制母體特性下, 直接蒐集資料。
2. 實驗 (*experiment*): 控制母體特性下, 探討單一變數的影響。在進行實驗時, 將受試者隨機分成實驗組 (*treatment group*) 與對照組 (*control group*)。
3. 觀察 (*observation*): 不與研究對象有直接接觸, 透過觀察其活動紀錄獲得資料的方式。

1. 調查的方法可分爲
   (a) 普查 (census): 蒐集母體裡的所有樣本。其優點爲具有普遍性與全面性, 缺點爲所需要耗費的成本 (包含人力與時間等) 較多, 故通常由政府施行。
   (b) 抽樣 (sampling): 抽取母體裡的一小部分樣本。其優點是所需的成本較低, 而常見的抽樣方法將會在第 7 章介紹。
2. 例如欲探討生師比高低對於成績的影響。但是生師比相對高與低的群體之間, 其家庭所得或者學生智商的分配通常具有顯著差異 (通常私立學校有較低生師比)。因此若如果直接抽樣比較, 將無法確認使得成績高低的影響是來自於生師比還是所得或者智商。故爲了控制其他變數不變下 (other things being equal, 或以拉丁文 ceteris paribus 表示), 應將所有學生打散隨機分成實驗組 (低生師比) 與對照組 (高生師比) 的兩個群體。如此一來, 實驗組與對照組除了生師比外, 其他變數都有相同分配。若兩組的成績有差異, 才可判斷差異是來自不同的生師比而不是其他變數。
3. 例如股票價格與匯率等總體資料, 都是實際的經濟活動所留下的記錄, 不需要特別抽樣或訪問即可獲得, 與成績與所得等個體資料蒐集的過程不太相同。

例題 1.2 資料獲得的方式

The personnel director at a large company studied the eating habits of the company's employees. The director watched and recorded whether each employee brought his/her own lunch to work, ate at the company cafeteria, or went out to lunch. What method of data collection was used here?

(A) Direct observation
(B) An experiment
(C) A survey.
(D) A personal interview

《解》

選 (A)。此資料是直接觀察而得。

例題 1.3 資料收集方式

關於資料的蒐集, 下列敘述何者非真?

(A) 使用次級資料時, 應該先檢查一下數據, 看看數字合不合理, 才可以進行分析。

(B) 如何蒐集可靠的資料, 是統計學最初步且最重要的工作。

(C) 資料依據取得的方式可分爲初級資料與二手資料。

(D) 原始資料的蒐集可分爲普查、實驗與觀察。 《104 中正經研》

《解》

選 (D)。

- 初級 (primary) 資料指的是針對欲探討的問題, 由研究者主動收集的資料
- 次級 (secondary) 資料指的是由他人或其他目的所產生的現成資料
- 原始資料的蒐集可分爲調查、實驗與觀察, 其中調查分爲抽樣與普查

例題 1.4 資料收集方式之其二

Which of the following data collection methods is not observational?

(A) A personal interview

(B) A telephone interview

(C) A self-administered questionnaire

(D) An experiment 《104 北科工管》

《解》

選 (D)。實驗資料需由研究者自行制定再收集。

例題 1.5 設計實驗進行因果推論

(10%) A service design researcher is interested in investigating the relationship between alcohol consumption and sleep disorders to develop an innovative app for potential users. Briefly describe how the researcher might design an observational study to investigate the relationship between alcohol consumption and sleep disorders? 《111 台大商研》

《解》

若要衡量酒精使用與睡眠障礙的問題, 則首先先決定如何衡量變數。睡眠品質可以利用熟睡時間衡量 (現在的智慧型手錶即可測量); 而酒精使用量可直接測量睡前或者一整天酒精的攝取量。收集完資料之後, 可利用迴歸模型分析兩者之間的關係, 被解釋變數爲睡眠品質, 解釋變數爲酒精使用量, 即可探討酒精使用量對於睡眠品質的影響。

例題 1.6 設計實驗進行因果推論之其二

By 2017, the workplace wellness business had ballooned into an eight-billion-dollar industry in the U.S. alone. One report suggests that half of firms with more than fifty employees offer wellness programs of some sort. The exact structures of corporate wellness programs vary, but the approach is grounded in preventative medicine. Wellness programs often involve disease screening, health education, fitness activities, nutritional advice, weight loss, and stress management. As wellness programs raise ethical questions about employers having this level of control and ownership over employees' bodies, employers cannot force employees to participate. In other words, if you work for a large company, you may have the option to participate in such a program. While employers say that they offer these programs because they care about their employees and want to improve their quality of life, the primary rationale for implementing a wellness program is that by improving the health of its employees, a company can lower insurance costs and decrease absenteeism. But there is a fundamental question: Do they work? Meta-analyses-studies that aggregate the results of previous studies – seem encouraging. They compare employees within the same company who did take part in wellness activities with those who did not, controlling for age, gender, weight, and other characteristics. Such studies typically found significant effects that wellness programs reduce medical costs and absenteeism, generating considerable savings for employers.

1. (6%) Can we conclude that offering a wellness program had such beneficial effects? Was there something particularly ineffective in the previous studies?

2. (6%) Find a way to deal with the problem above and investigate whether offering a wellness program has such beneficial effects. 《112 台大國企乙》

《**解**》

1 由資料可看出參加身心健康方案 (wellness activities) 的員工減少了醫療支出, 但問題在於參加身心健康方案與否不是隨機決定的, 僅有大公司的員工可以選擇參與。意即參加的員工與不參加的員工的群體除了年齡性別與其他特徵外, 可能有顯著差異。參加的員工是比較注意身體健康的, 通常飲食或者作息也會比較健康, 因此造成醫療支出減少可能是因爲注重身體健康而不是參加了身心健康方案。

2 首先, 在大公司跟小公司都要有可以參加方案的機會。再者, 參加方案與否不應讓員工自由決定, 而應該採隨機抽籤決定。

性質 1.3 常見的隨機抽樣 (*random sampling*) 方法

1. 簡單隨機抽樣 (*simple random sampling*)
2. 系統抽樣 (*systematic sampling*)
3. 分層抽樣 (*stratified sampling*)
4. 叢式抽樣 (*cluster sampling*)

1. (a) 簡單隨機抽樣: 將每一個案都編成號, 然後利用亂數表, 以抽出不放回的方式, 隨機選出需要之樣本。

   (b) 系統抽樣: 第一個樣本是利用亂數表取得, 其他的樣本則是依次加上一定之抽樣間距取得, 而抽樣間距爲母群數/樣本數。例如母體數爲 1000, 樣本數爲 20, 第一組樣本若爲編號第 27 組樣本, 則第二組爲編號第 77 組樣本, 第三組爲編號第 127 組樣本, 以此類推。

   (c) 分層抽樣: 先依照資料特性 (例如年齡、性別或所得級距等) 將資料分組, 在每一分層中, 依比例做簡單隨機抽樣 (通常比例爲各組中個體佔母體之比例)。分層抽樣主要適用於資料分佈變異程度較大或者分佈較不對稱時, 透過分層可使得層間差異大, 層內差異小之原則時。分層抽樣中, 同組間通常具有相同特性, 因此**分層抽樣組內差異小, 但組間差異大**。

   (d) 叢式抽樣: 將母體分成幾個群集 (例如學校、縣市等), 再從這幾個群集中抽出數個群集進行抽樣或普查。叢式抽樣主要適用於資料母體不易全部觀察到, 並且符合群集內差異大, 群集間差異小之原則時。**叢式抽樣組間差異小, 但組內差異大**。

2. 舉例來說, 若欲進行總統選舉的民調, 樣本數爲 1000:

   (a) 若採隨機抽樣, 則在底冊 (frame, 所有可供抽取樣本所形成的集合) 中隨機訪問。

   (b) 若採分層抽樣, 則先將底冊根據年齡分組, 並根據各組佔全體人民的比例抽取樣本, 此時抽取的樣本可視作是母體的代表性樣本。

   (c) 若採叢式抽樣, 則先將底冊依據地區分群 (例如鄰或者里), 並隨機抽取一里, 對於該里進行普查。

3. 最常見的隨機抽樣方法是**簡單隨機抽樣**。

4. 隨機抽樣需滿足

   (a) 母體中任意一元素均有可能被抽取

   (b) 任何一組樣本被抽取出的機率是相同的

   (c) 每一筆樣本被抽出的過程是獨立的

5. 抽樣過程非隨機的 (non-random sampling), 例子有

   (a) 方便抽樣 (convenience sampling): 僅抽取研究者身邊可觀察到的樣本。

   (b) 判斷抽樣 (judgement sampling): 依據研究者主觀判斷, 抽取認定的樣本。

   (c) 滾雪球抽樣 (snowball sampling): 訪問完該樣本後, 再請該樣本推薦下一個樣本。

   但非隨機抽樣所取得的樣本容易受到質疑, 因此應使用隨機抽樣, 樣本才會具代表性。

6. 常見的造成抽樣結果不隨機的情況有以下幾個:

   (a) 回應偏誤 (response bias) 或不回應偏誤 (non-response bias): 提供回答的受訪者與未提供回答的受訪者之間有意見的差異。例如在選舉中, 若支持 A 候選人的民眾比較願意表態支持 A, 但支持 B 候選人的民眾比較不願意表態支持 B, 則其調查結果會與實際支持率之間有顯著的誤差。

   (b) 樣本選擇偏誤 (sample selection bias): 在選擇樣本的過程中 , 系統性的只收集或未收集到特定群體。例如觀察有小孩的女性其平均薪資高於沒有小孩的女性, 但其主要原因是有小孩的女性若自身薪資較低, 通常不會選擇工作。

   (c) 倖存者偏差 (survival bias): 在蒐集資料的過程 , 遺漏消失在樣本中的群體 , 只抽取還留在樣本中的群體。例如只收集上榜生的心得但卻忽略落榜生的心得, 可能會誤判正取的關鍵因素。

定義 1.1 抽樣誤差 (*sampling error*)
對於有興趣的母體參數 (例如母體平均), 通常利用對應的樣本統計量 (例如樣本平均) 作爲估計。樣本統計量會隨著抽取的樣本不同而改變, 且通常不等於母體參數, 而樣本統計量與母體參數之間的差異稱作抽樣誤差。

例如若母體中拿 iPhone 的真實比例爲 80%, 若抽取 10 個樣本, 其中有 7 位拿 iPhone, 則樣本比例爲 70%, 故此例的抽樣誤差爲 $|0.8 - 0.7| = 0.1$。

例題 1.7 抽樣方式的判斷

A researcher wishes to estimate the average height of all students at National Sun Yat-sen University (NSYSU). He would randomly sample 30 students and measures their heights. He considers the following four ways of random sampling. Which way of random sampling do you think is most appropriate?

(A) Go to the gym of the university and randomly select 30 students in the gym.

(B) Stay at the main entrance of the university and randomly select 30 students who pass the entrance.

(C) Go to one of the classes of the university and randomly select 30 students in that class.

(D) Go to the registration office of the university and take the list of all students and randomly select 30 students from the list. 《104 中山企研》

《解》

選 (D)。(A) 方法只能抽到在體育館的學生, (B) 方法無法抽取到不走正門或者是當天沒課不到學校的學生。(C) 是叢式抽樣。

例題 1.8 判斷抽樣方式

Two data analysts had the following conversation of sampling methods:

Data Analyst A: If the population is grouped first, and then the sample is composed of every member of some groups. This is the clustered sampling method.

Data Analyst B: If the population is grouped first, and then the sample is composed of some members of each group. This is the clustered sampling method.

Which data analyst is correct in a statistical sense?

(A) Data Analyst A

(B) Data Analyst B

(C) Both data analysts are incorrect 《108 成大國企》

《解》

選 (A)。叢氏抽樣的是將母體分作類似的數個群體, 然後選取其中一個群體的每個樣本作爲代表性樣本。

例題 1.9 判斷抽樣方式

Manufacturers were subdivided into groups by volume of sales. Those with more than $100 million in sale were classified as large; those from $50 to $100 million as medium size; and those below $50 as small size. Samples were then selected from each of these groups. What is this type of sampling called? 《104 政大財管》

《解》

先分成層之後再抽樣的方法爲分層隨機抽樣。

例題 1.10 判斷抽樣方式之其二

To survey the opinions of the students in a class, a teacher plans to select every twenty-fifth student entering the classroom in the morning. Assuming there are no absences, will this sampling plan result a simple random sample of students attending the classroom?

(A) Yes, because every students has the same chance of being selected.

(B) Yes, but only if there is a single entrance to the classroom.

(C) No, because not every sample of the intended size has an equal chance of being selected.

(D) Yes, because this is an example of systematic sampling, which is a special case of simple random sampling. 《109 成大統研》

《解》

選 (D)。指定每一班的第 25 位即爲系統性抽樣的一個範例。

例題 1.11 抽樣方法的基本概念

Which of the following statements is true?

(A) In stratified random sampling strata are more homogeneous.

(B) In cluster random sampling, clusters are internally homogeneous.

(C) Systematic sampling is also called snowball sampling.

(D) None of the above. 《105 政大財管》

《解》

選 (A)。分層抽樣中, 層 (strata) 之間變異較小。(B) 錯在叢式抽樣中, 不同的分叢之間變異較小 (externally homogeneous)。(C) 錯在系統抽樣與滾雪球抽樣是不一樣的。

性質 1.4 資料類型

蒐集到的資料, 根據其特性可分類如下:

1. 類別 (*categorical*) 或數值 (*numerical*)。
2. 離散 (*discrete*, 或稱間斷) 或連續 (*continuous*)。
3. 時間序列 (*time series*) 或橫斷面 (*cross-sectional*)。

1. 類別資料又稱作屬質 (qualitative) 資料, 數值資料又稱作屬量 (quantitative) 資料。根據其測量尺度 (scales of measurement) 可進一步的細分爲

   (a) 類別 (屬質)

      i. **名目** (**nominal**) 資料: 非數字紀錄統計的資料, 如性別、科系與背號等。屬質的資料亦可以利用數字表示, 例如科系可用科系代碼表示, 或者性別可用 1 表示男性, 0 表示女性。但其相對大小不具任何意義, 因此無法進行排序或運算, 只能進行分類以及計算各類別的累計次數。

      ii. **順序** (**ordinal**) 資料: 或稱作次序資料, 其相對大小有意義, 因此可以排序, 但不可以進行運算。例如債券評比、滿意度以及年級 (大一至大四) 等。

   (b) 數值 (屬量)

      i. **等距** (**interval**): 等距資料的 0 爲相對參考值, 例如**攝氏溫度與智商**。溫度爲 0°C 不代表沒有溫度, 智商爲 0 不代表沒有智商。因此等距資料可以進行加減運算, 但舉例來說, 氣溫 30°C 不爲 15°C 的兩倍熱, 因此不可進行乘除運算。

      ii. **比例** (**ratio**): 比例資料的可以進行任何運算, 其值爲 0 代表沒有, 例如所得。

2. (a) 離散資料:

      i. 離散資料是可計數的 (countable), 爲透過**統計數量**獲得的資料,

      ii. 例如考試成績、一段時間的來客數與擁有的小孩數等都是離散資料。

      iii. **離散資料不可無限細分**, 在有限的區間當中, 其可能實現值的數量是有限的。舉例來說, 來客數 5 人與 10 人之間, 只有可能爲 6 至 9 人共 4 種實現值。

   (b) 連續資料:

      i. 連續是不可計數的 (uncountable), 爲透過**測量或計算**獲得的資料,

      ii. 例如身高、體重、時間、溫度等資料需要透過設備測量, 因此爲連續型資料。