

資料的性質與蒐集



統計 * 讓數字說話

母體與樣本

母體(population)與樣本(sample)

母體(population) :

具有共同特質的元素或個體所組成的群體，也是所要探討事件之全體對象。



參數(parameters) :

描述母體特質的統計測量數（ /特徵數 ）。

樣本(sample) :

來自母體抽取部分群體的集合。

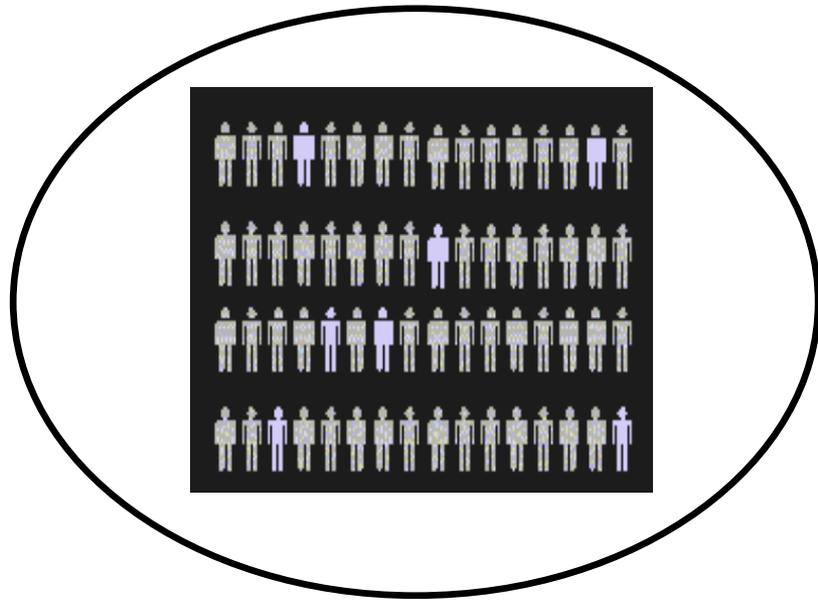


統計量(statistics) :

描述樣本特性的統計測量數（ /特徵數 ）；隨著每次選取的樣本不同，統計量會不同。

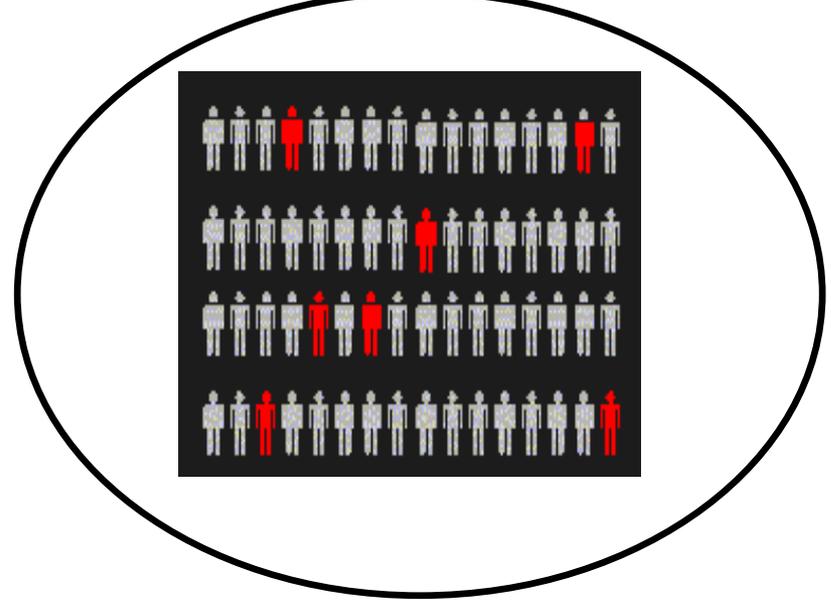
母體(population)與樣本(sample)

population



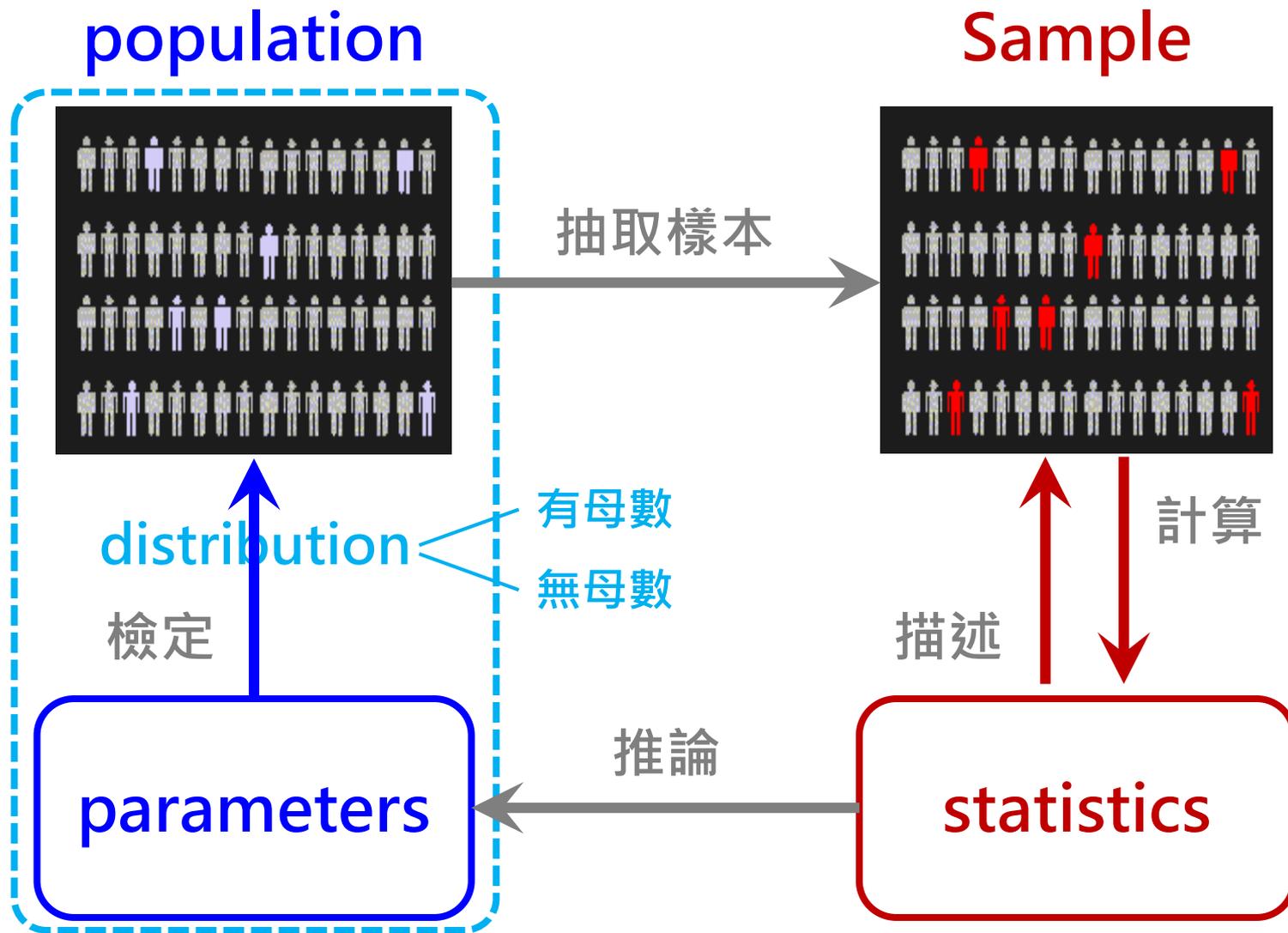
All the items or individuals about which you want to draw conclusion(s)

Sample



A portion of the population of items or individuals

“統計”過程



母體(population)的類型

依母體個數是否可區分：

有限母體(finite population)，個數可數且為有限個。

無限母體(infinite population)，個數無限多個，或者範圍有限但卻不可數（例如介於1到2之間的實數）。

依母體所含資料的特質區分：

質母體(qualitative population)，“類別”資料性質的母體（例如血型、性別、職業等）。

量母體(quantitative population)，屬於數值資料性質的母體（例如身高、體重、收入等）。

樣本(sample)的類型

隨機與非隨機：

隨機樣本(random sample)，透過機率方式抽取樣本，一般利用隨機樣本所推論出來的結果較為可靠。

非隨機樣本(non-random sample)，在某種情況下僅能以非機率方式抽取樣本，例如“方便抽樣法”所抽取的樣本就不具隨機性。

獨立與非獨立：

獨立樣本(independent sample)，樣本與樣本之間不會有互相影響，且無次序、先後的關連。

非獨立樣本(dependent sample)，或稱相依樣本，樣本之間有某種關連或是有時間、先後次序的關係，例如研究父母與子女之間某議題的關連性。



資料的類型

資料可能呈現的各種類型模式

資料的種類1

依獲得資料方式：

一手資料 (primary data)：又稱為初級資料，資料的取得必須透過研究者親自透過問卷調查、實驗或是觀察獲得。

二手資料 (secondary data)：又稱為次級資料，指的是已經存在的資料，研究者可以直接或是間接採用的資料。

依資料存在時間：

靜態資料 (static data)：或稱為橫斷面資料 (cross-section)，通常指的是同時間所收集的資料；也就是不把「時間」的因素考量進統計分析的過程中所收集得資料。

動態資料 (dynamic data)：或稱為時間序列 (time series)，指的是資料的收集是透過不同時段所得，也就是統計分析的過程必須考量「時間」的因素。

資料的種類2

依資料涵蓋範圍：

普查資料 (census data)：研究對象必須要全面性調查以獲得所要資料。

抽樣資料 (sampling data)：研究對象可以抽取其中代表性的資料做調查。

依資料來源：

內部資料 (internal data)：存在組織內部的資料，通常不公開。

外部資料 (external data)：可以在組織外部所收集到的資料。

依資料呈現特質：

定性資料 (qualitative data)：或稱為類別資料 (category data)，資料屬性通常只具有名義上的性質，無法做任何數學的四則運算。

定量資料 (quantitative data)：資料屬性為數值類型，通常可以對相同單位的資料作數學的運算。

資料的種類3

依資料數學性質：

離散資料 (discrete data)：資料的屬性會以某一特定數值表示，且兩數值之間無法進行分割，例如撲克牌的點數。

連續資料 (continuous data)：資料之間存在連續性的數值，可以進行無限的分割，例如長度、體積等。

依資料是否有分組：

組距型資料 (grouped data)：整理這類型資料時，通常會按照一個區間作為分類標準，例如以每十年為一個群組。

非組距型資料 (ungrouped data)：或稱為組資料 (raw data)，整理這類資料時是按順序排列。

變數(variable)與變量(variate)

變數(variable)：

可記量的特徵，用以替代分類項目的名稱，一般以大寫英文字母X、Y、Z 等表示。

變量(variate)：

「變數」的數值，用以代替資料的值。一般以小寫的英文字母 x 、 y 、 z 表示。

X → 男性收入

Y → 女性收入



x_i ：第*i*位男性的月薪收入

y_j ：第*j*位女性的月薪收入

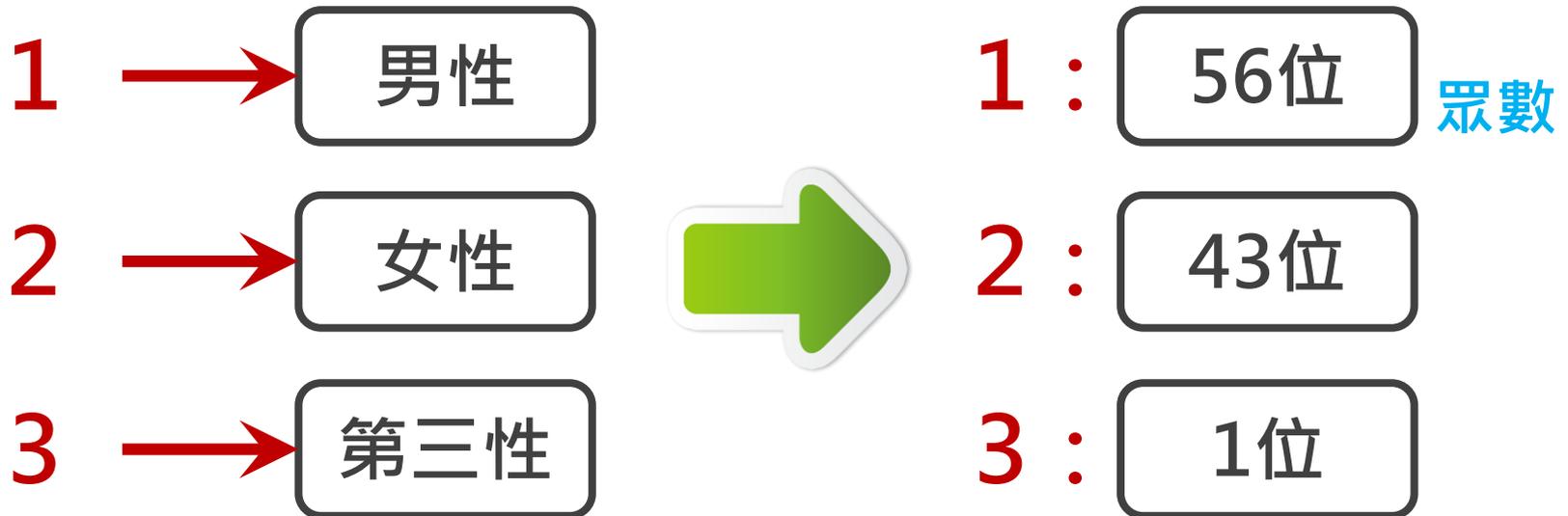


資料的衡量尺度

資料「量化」的方式

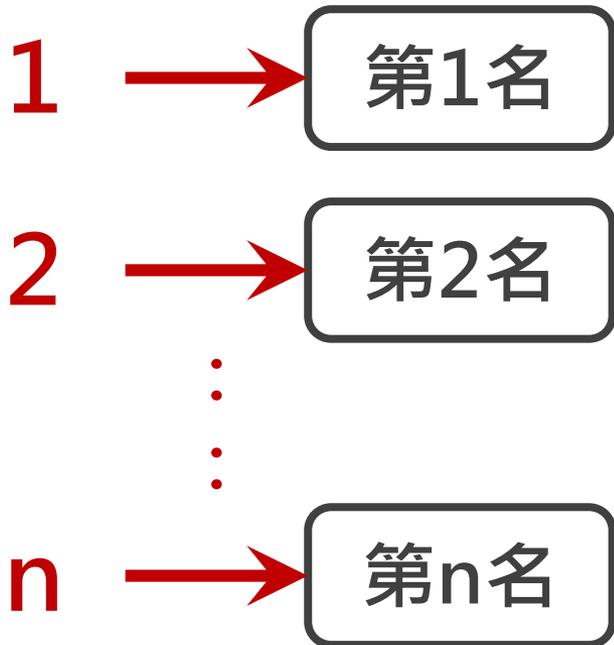
名義量尺 (nominal scale)

主要衡量類別型態的資料，一般使用簡單的數字作為標示其屬性，而這些作為標示用的數字無大小的關係，也不可以做四則運算。



順序量尺 (ordinal scale)

主要衡量有大小、先後、程度的順序資料，其同時也具有名義尺度的性質。



區間量尺 (interval scale)

主要衡量數字與數字間具有差異意義，但不具有倍數關係的資料，例如溫度、智商等。區間量尺同時具有名義量尺與順序量尺的含意。因為數字有意義，所以可以做加減運算，但乘除運算則不具意義。



比率量尺 (ratio scale)

同時具有區間量尺、順序量尺與名義量尺三種的含意。通常來衡量數值型態的資料；兩個資料間的比值具有意義。資料間可以做四則運算，例如成績、體重、身高等。



區間量尺 V.S. 比率量尺

區間量尺

比率量尺

「0」是否代表
不存在？

「0」具有
資料屬性

「0」代表
資料不存在

有倍數意義？

無倍數關係

有倍數關係



錯誤的衡量方式

不要一開始就錯了.....

Types of Survey Errors

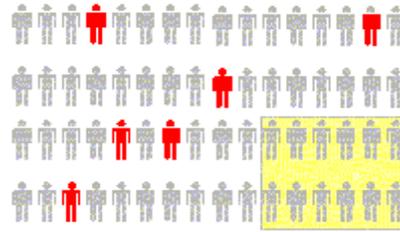
DCOVA

- Coverage error or selection bias
 - Exists if some groups are excluded from the frame and have no chance of being selected
- Nonresponse error or bias
 - People who do not respond may be different from those who do respond
- Sampling error
 - Variation from sample to sample will always exist
- Measurement error
 - Due to weaknesses in question design, respondent error, and interviewer's effects on the respondent ("Hawthorne effect")

Types of Survey Errors

DCOVA
(continued)

- Coverage error



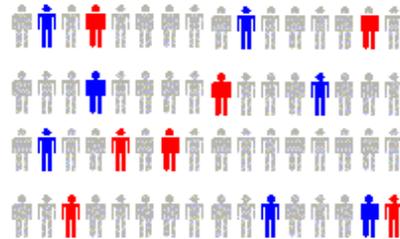
Excluded from frame

- Nonresponse error



Follow up on nonresponses

- Sampling error



Random differences from sample to sample

- Measurement error



Bad or leading question



The End