巨量資料與統計分析

政治大學統計系余清祥

2024年10月8日

第五週：探索性資料分析(結構資料)

http://csyue.nccu.edu.tw

大數據資料的類型

結構資料與非結構資料的比較

- 大數據資料可分為結構資料、非結構資料
- **結構資料**又稱為硬資料(Hard Data)，通常以有系統的方法蒐集，且多為量化資料，
- **非結構資料**又稱為軟資料(Soft Data)，則較無系統、偏向質性資料，而且量測的變數與目標值間相關，但未必完全一致。
- 以手機的品質為例，可蒐集的結構資料包括拍照品質、通話及網路品質、待機時間等，非結構資料則有品牌形象、外觀、介面等。

結構資料與非結構資料的範例

- **結構資料**

→Excel之類的試算表資料，除了容易以掃瞄方式輸入，也易於分析欄位之間的關連性，甚至建立網絡類型的關係（標籤？）；可透過搜尋引擎之類的工具找出觀察值的基本特性。

- **非結構資料**

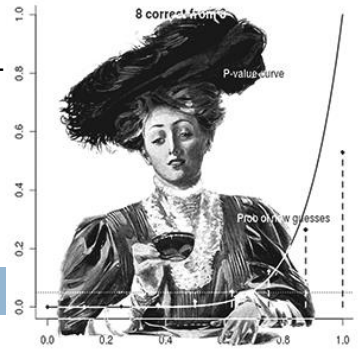
→Email之類的文字資料，通常除了時間、檔案大小、使用者外，不容易找出資料的基本特性（例如：平均數、變異數），需要對議題、內容有相當瞭解，才能將資料結構化（耗時、費力）！

結構資料與非結構資料的比較

	結構資料(Structured Data)	非結構性資料(Un-Structured Data)
別名	硬資料(Hard Data)	軟資料(Soft data)
記錄方式	大多以數量化格式	以文字、影音、圖片等非數量化格式
資料蒐集	較為客觀	較為主觀
資料特徵	■ 容易量化 ■ 且有系統 ■ 容易整理 ■ 容易存取 ■ 容易大量傳輸 ■ 資料測量與蒐集與研究者無關	■ 不易制定量化 ■ 較無系統 ■ 不易整理 ■ 資料本身與研究者有深度連結 ■ 需相關知識才可傳輸與量化資料
資料類型	問卷調查、全民健保資料庫	質性訪談資料、部落格、電子郵件
分析方式	省時、較為直接	耗時費力、必須經過某種方式轉換
範例	性別、年齡、科系、相片畫素	臉部特徵、性格、情緒、品牌

探索性資料分析

https://thedailyomnivore.net/2015/06/08/lady-tasting-tea/



統計分析與假設檢定

- 問題：為什麼假設檢定常讓人感到困惑？
  以下為幾個常見與假設檢定有關的範例：

→溫室效應、極端氣候

https://bookzone.cwgv.com.tw/topic/details/10493

→淑女品茶 (Lady Testing Tea)

- 「觀察」是假設檢定的必要步驟之一，當觀察到超乎預期（Q：定義？）的現象時，我們需要判斷這是偶而出現（罕見）、或是經常出現（常態）。

資料分析策略

- 「觀察」、「推論」、「驗證」三步驟

→檢查資料品質，避免Garbage in, garbage out。

→先進行探索性資料分析(EDA, Exploratory Data Analysis)找出關鍵變數（或特性）。

→接著驗證性資料分析(CDA, Confirmatory Data Analysis)。（EDA、CDA結果應該接近！）

資料分析的類型

- 統計觀點可分為兩類：

→探索性資料分析(Exploratory Data Analysis)

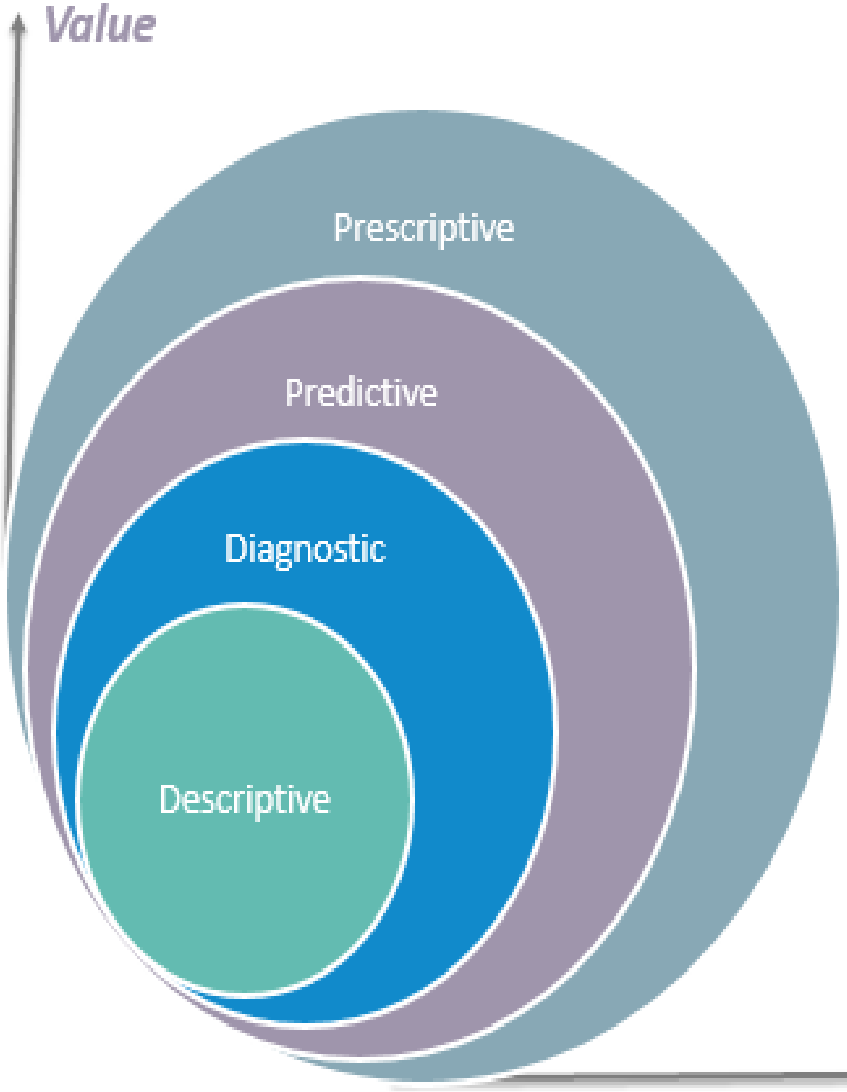
→驗證性資料分析(Confirmatory Data Analysis)

- 機器學習觀點分為三類：

→敘述性分析(Descriptive Analytics)、預測性分析(Predictive Analytics)、建議性分析(Prescriptive Analytics)；

→「發生了什麼事」(What has happened)、「未來會如何」(What would happen)、「我們如何調整」(What should we do)。

4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

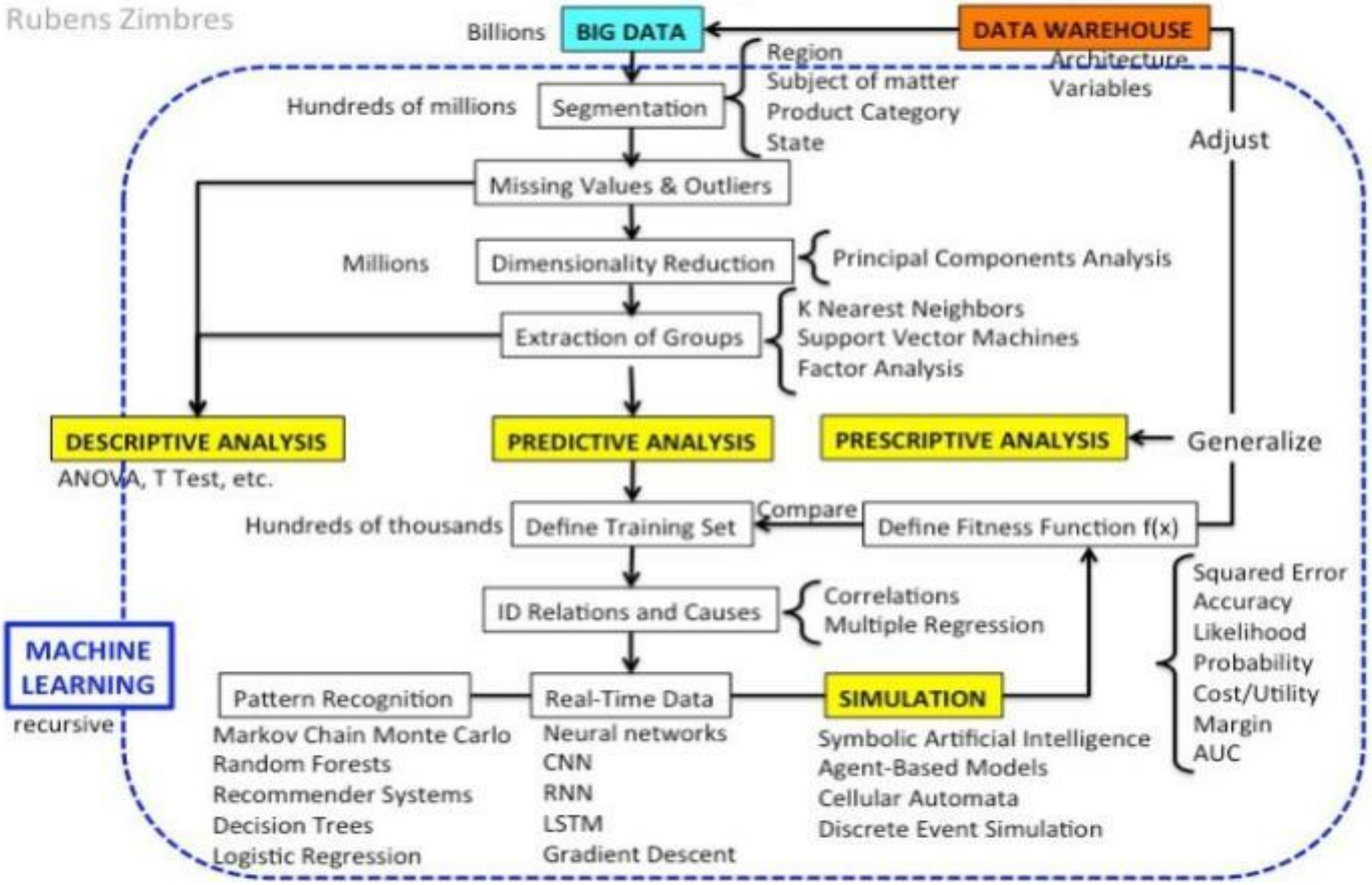
Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

http://www.kdnuggets.com/2017/07/4-types-data-analytics.html

Machine Learning Applied to Big Data



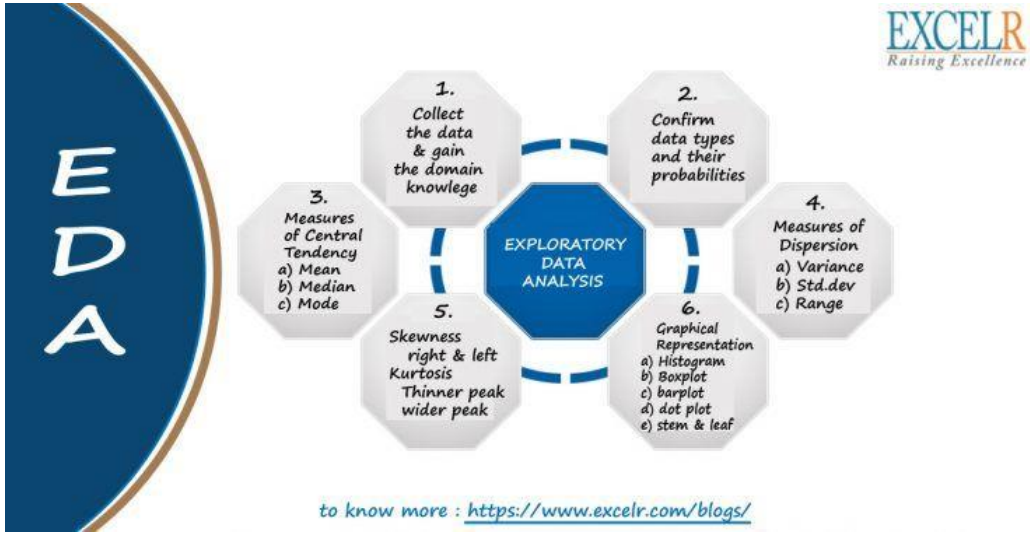
http://www.kdnuggets.com/2017/07/machine-learning-big-data-explained.html

探索性資料分析(EDA)

- EDA(Exploratory Data Analysis)目的在於資料偵錯、獲得資料的大略資訊、驗證已知結果。

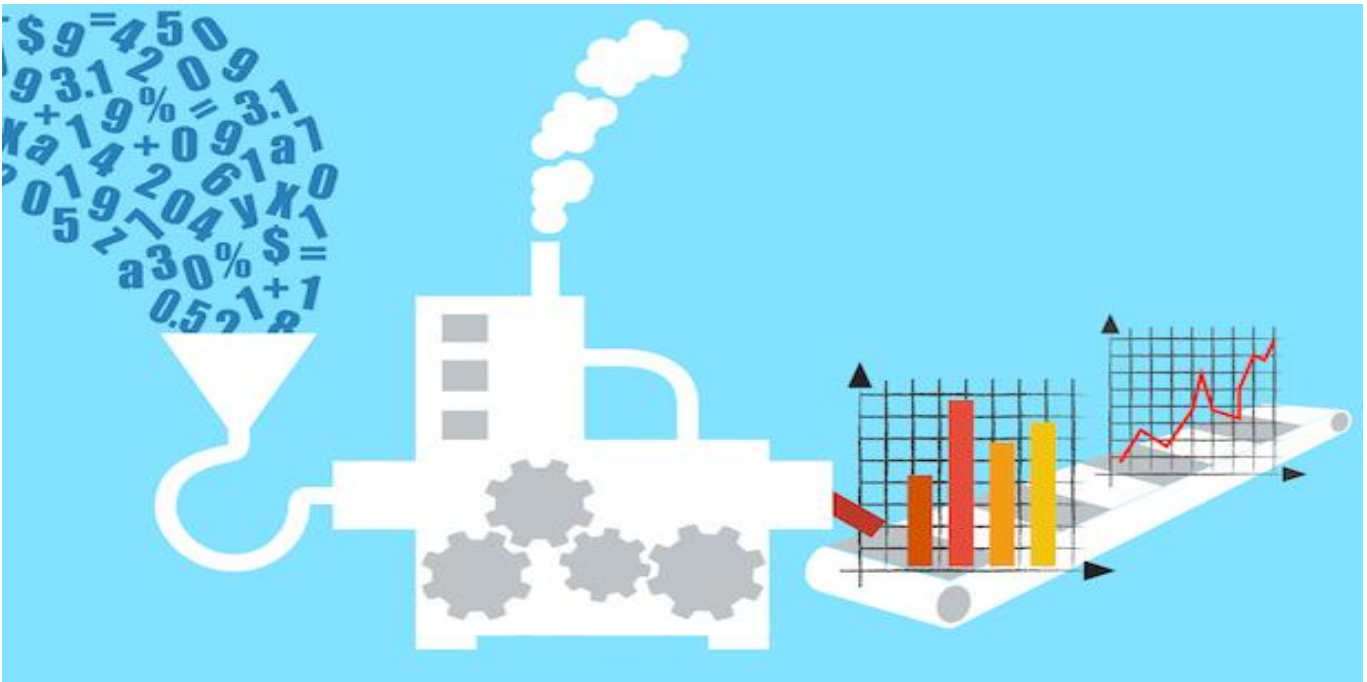
→**圖形**、**表格**在EDA中扮演重要的角色；並由分析結果中尋找合適的下一步分析方法。

→使用統計方法前，先確定該方法的**假設條件**是否滿足。

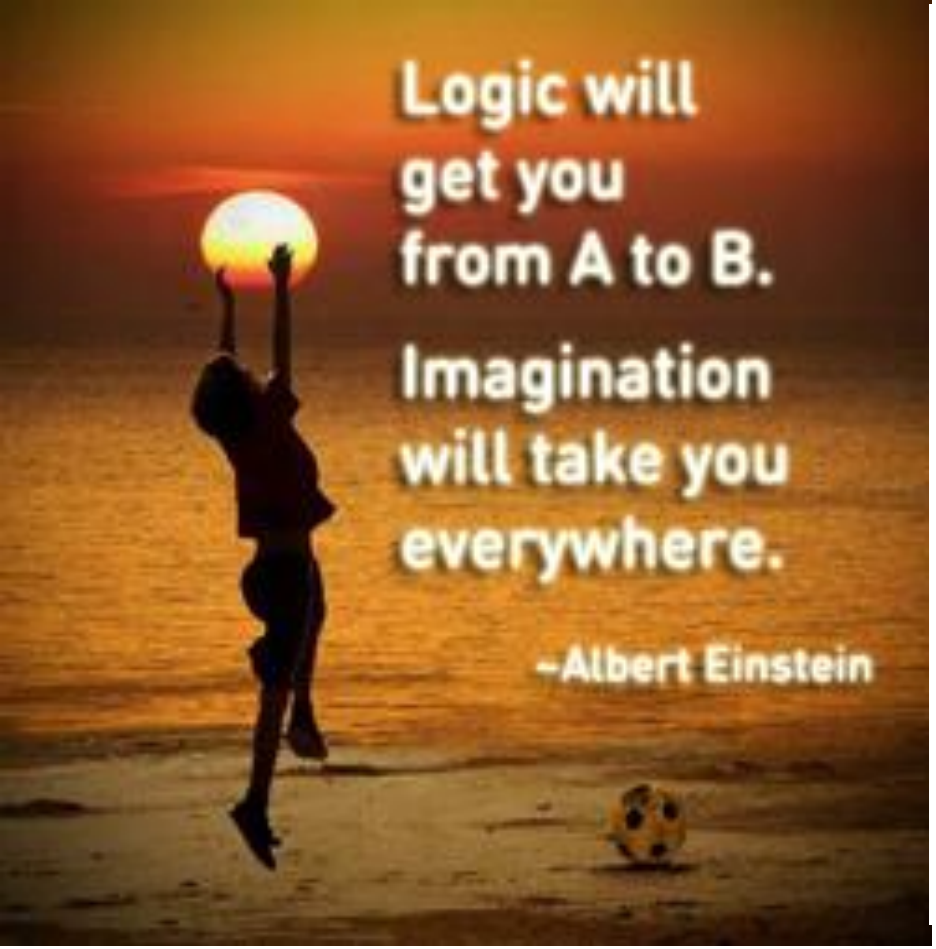


探索性資料分析(資料驅動)

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics … EDA is for seeing what the data can tell us beyond the formal modeling. ---Wikipedia



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.aiche.org%2Facademy%2Fwebinars%2Fapplied-statistics-exploratory-data-analysis&psig=AOvVaw36ZuxAJqz27dLqU5IFzBMO&ust=1570108849384000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCJC1qLXV_eQCFQAAAAAdAAAAABAJ



"To raise new questions, new possibilities, and to regard old problems from a new angle, requires creative imagination and marks real advances in science."

--- Albert Einstein

Imagination and domain knowledge in statistics are both necessary to maximize the likelihood of insight discovery. That is why EDA is challenging, but it's also what makes it fun.

--- Michael Wu

基本資料分析的項目

- 資料偵錯

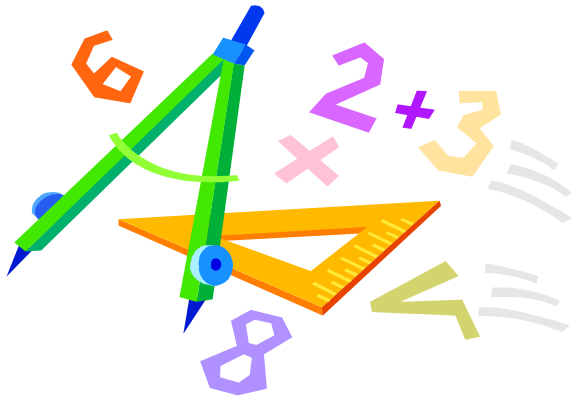
→ 資料輸入錯誤、尋找可能的離群值。

- 初步探索資料的特性

→ 資料的集中、散佈趨勢。

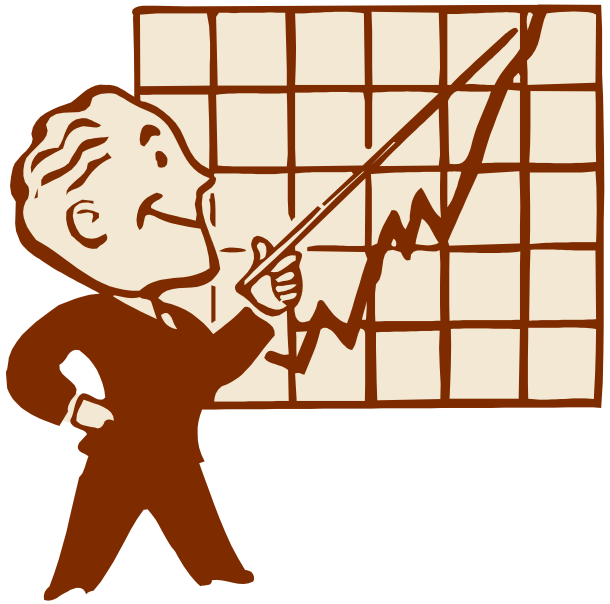
- 驗證已知的結果

→ 是否與已知的結果相同？

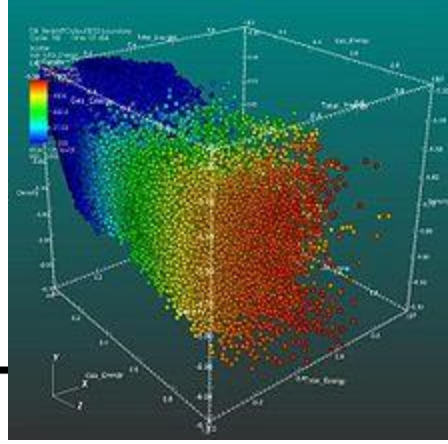


統計分析原則

- 確定問題的定義
- 化繁為簡（反璞歸真）
- 結合相關知識
- 發揮聯想力（大膽假設）
- 勿驟下結論（小心求證）



https://upload.wikimedia.org/wikipedia/commons/c/c4/Scatter_plot.jpg



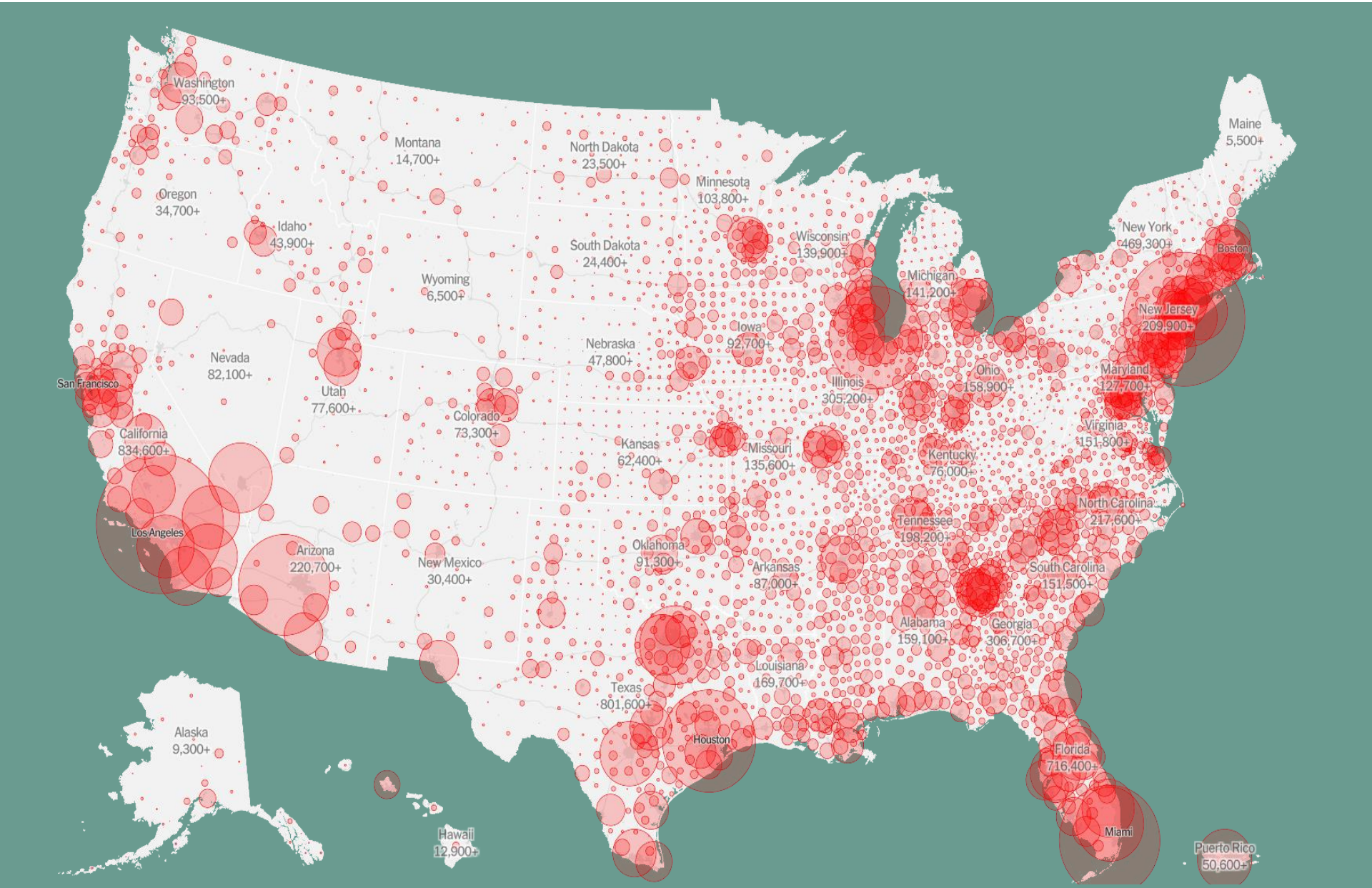
WIKIPEDIA

Data visualization

Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will be represented visually, determining how and to what extent a property of a graphic mark, such as size or color, will change to reflect changes in the value of a datum.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.[1] Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

美國各地Covid-19確診數（紐約時報）



https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

資料偵錯的策略

資料錯誤的可能來源

- 形成錯誤資料的來源包括輸入、整併、儲存、傳送，通常分為下列四種：
  - Data entry errors（輸入）
  - Measurement errors（測量）
  - Distillation errors（標準化）
  - Data integration errors（整併）

註：自動化輸入及偵錯、品管、EDA等可提高資料品質

來源：Quantitative Data Cleaning for Large Databases (Hellerstein, 2008)

問卷資料輸入與偵錯

- 以問卷資料輸入為例，

→先撰寫編碼薄(Codebook)，並將選項填寫在題號之前，篩選可能問題；

→檢查是否有輸入錯誤(1% ~ 5% Error)；

→翻閱原始問卷。

- 其他注意事項

→檢查離群值及其合理性；

→其他選項(文字、項目歸類)的整理；

→複選題及排序題。

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	□	□	□	□	□	□	□	□	□	□

衛星影像驚見台中冒「煙流」？ 鄭明典致歉誤判

鄭明典
10月9日 22:10 ·

追蹤

煙從哪裡來？

這是昨天上午10點左右的衛星影像，箭頭所指的煙流明顯來自地面單點燃燒的釋放源，到底是甚麼東西可以燒到讓衛星看出完整的煙流？
影像來源：NASA/EOSDIS

讚　　留言　　分享

衛星拍到台中有不明煙流 專家質疑、網友罵、環保局追查

今日新聞NOWnews
生活中心 / 綜合報導 2017年10月10日 下午4:02

衛星拍到台中有不明煙流！氣象專家鄭明典9日晚間在臉書上傳一張台中衛星照片，只見畫面中有一道明顯白煙，讓他問：「到底是甚麼東西可以燒到讓衛星看出完整的煙流？」消息一出，網友紛紛大罵工廠汙染，台中市環保局正朝火災、工廠排放、露天燃燒等方向追查中。

綜合媒體報導，9日晚間，鄭明典在臉書貼出一張衛星照片，時間是8日上午10時左右，只見畫面中箭頭所指處，有一道濃濃的白煙飄向天際，讓鄭明典疑惑道：「到底是甚麼東西可以燒到讓衛星看出完整的煙流？」經網友利用Google Map比對後，發現釋放源可能在台中后里區。

看過這張衛星照片的網友們，紛紛大罵「工廠汙染」、「環保單位有在查嗎？」、「檢舉好幾次都沒用」、「大二大城歡迎您，空氣品質逐年提升」，也有人反應「難怪台中這兩天空氣非常髒」，被其他人反駁「每天空氣都很髒」；對於「老天有眼，汙染現形」一事，網友們呼籲將訊息散布出去，讓更多人關注台中空汙問題。

台中環保局得知後，正朝大型火災、工廠排放、大型露天燃燒等方向進行追查。環保局表示，8日當天並無獲報有大型火災事件；后里區的焚化廠和其他工廠的汙染源數據也無異常、超出標準；依衛星照片的煙量，8日查獲的4起大型燃燒也非煙流主因。


HOW'S THE BIG DATA PROJECT COMING ALONG, HOSKINS?
© D.Fletcher for CloudTweaks.com

利用EDA進行資料偵錯

利用EDA進行資料偵錯

- 可根據資料類型設計偵錯方式，大致分成量化資料、類別資料。
- 使用EDA的基本原則為：

→先找出資料的基本特性，像是大略的數值及其範圍(Center and Dispersion)，這些統計量的計算方式不唯一，可能需要反覆嘗試；

→接著根據上述資訊定義離群值，進而判斷哪些資料可能需要修正（偵錯）。

註：（成長）率及指標需要謹慎處理！

- 資料類型將直接影響分析方法的選取，並非所有資料都適合常見的統計方法，任意使用分析方法可能會得出令人啼笑皆非的結果。

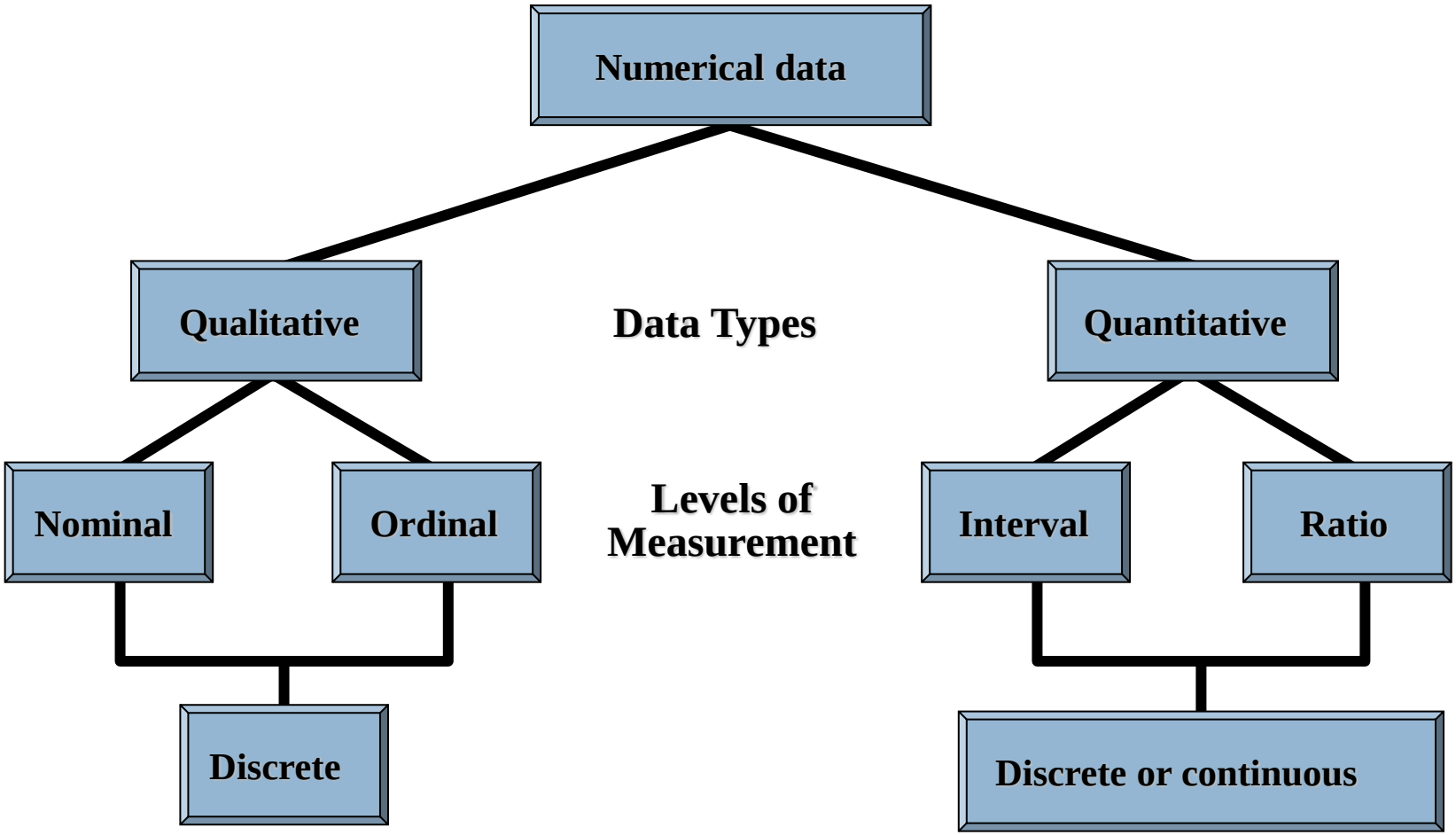
  →已知A > B, B > C 是否代表A > C？

	甲城市	乙城市	丙城市
A候選人	1	2	3
B候選人	2	3	1
C候選人	3	1	2

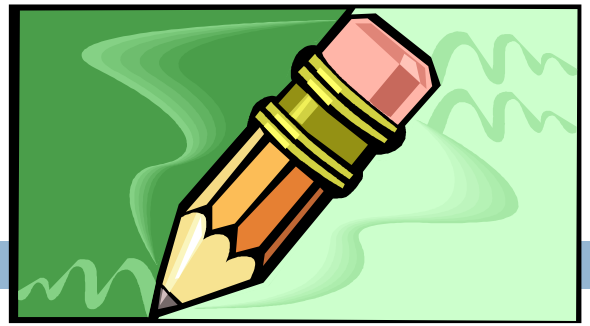
註：1代表最喜歡，3代表最不喜歡。

資料類型與整理方式

Types of Data(資料類型)



計算描述型統計量



- 探索性資料分析(Explanatory Data Analysis)是資料分析中最基本、也是非常重要的一個步驟，資料分析的成敗往往在這個步驟中決定。

→敘述性統計量包括資料的基本特性，如：平均數、標準差、所佔比例(圖表)等 ，一般的整理方式為：**集中**趨勢量數、**散佈**量數。

→以關聯性(Association)描述變數間關係，再進一步以分類(Classification)、群聚(Cluster)區隔變數及觀察值的類別。

視覺化圖表呈現資料

- 除了基本的敘述統計量外，圖形與表格可以輔助判斷資料的特性。

→常見圖形：Boxplot、Histogram、Scatter plot

- 這些圖表看似簡單，但仔細判讀仍可發現重要訊息，甚至不需進階統計分析，即能約略猜出分析的結論。

→以民國94年大學指定科目考試的成績為例，判斷各科分數的特性。

如何藉由統計獲取資訊？

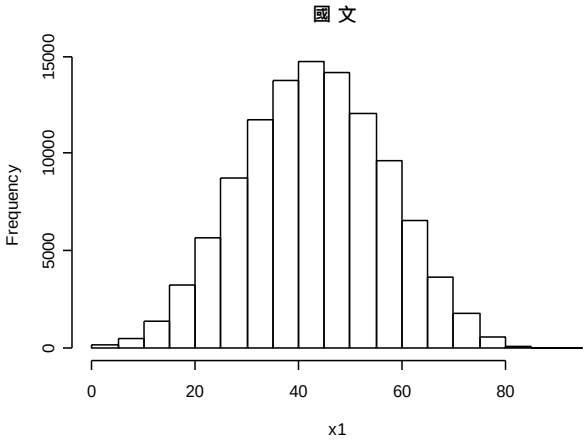
- 如果想瞭解民國94年指定考試各科的特性，可以藉助哪些工具？

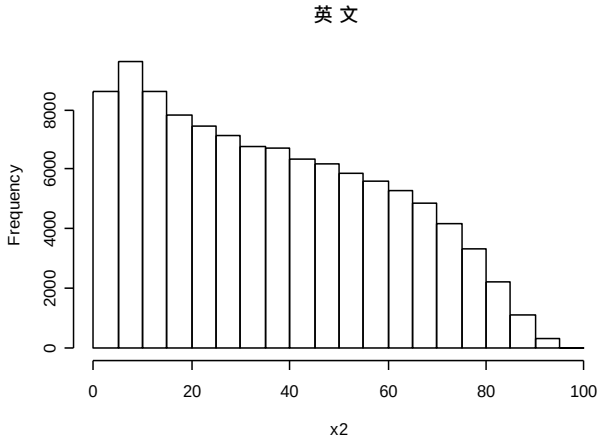
→例如：那一科的分數最不平均，像是哪一科大多數人都考得不好，只有少數人分數分高。

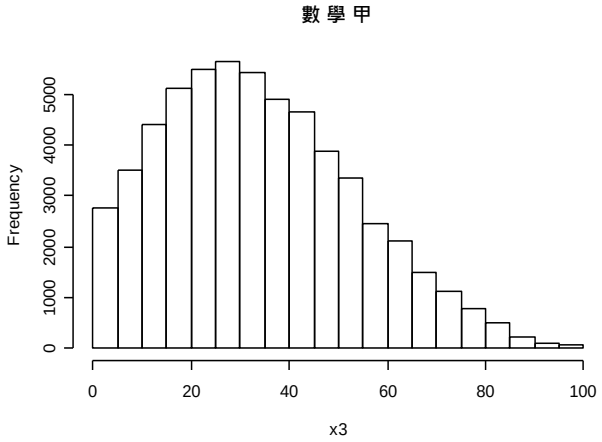
→平均數明顯大於中位數，稱為右偏(skewed to the right)；反之，若平均數明顯小於中位數，稱為右偏(skewed to the left)。平均數等於中位數，則為兩側對稱。

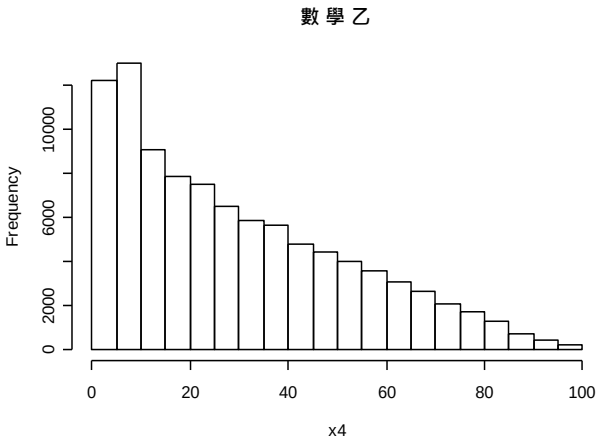
民國 94 年大學指定考試各科成績

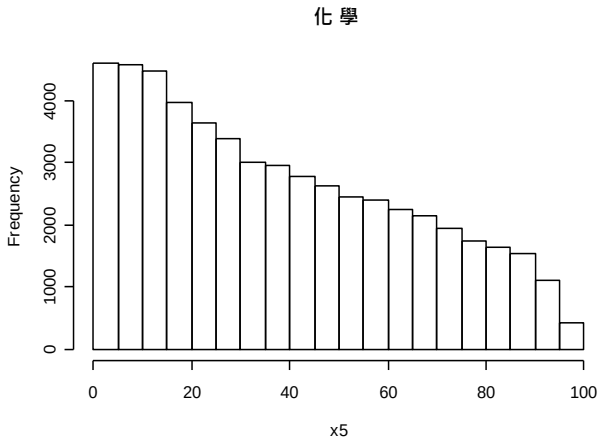
	國文	英文	數學甲	數學乙	化學	物理	生物	歷史	地理
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00
12%	27.00	8.00	11.00	4.00	8.00	6.00	22.00	13.0	18.00
1st Qu.	34.00	16.00	22.00	12.00	15.00	12.00	32.00	28.0	30.00
Median	44.00	34.00	34.00	29.00	34.00	23.00	45.00	39.0	39.00
Mean	43.56	36.68	36.36	34.36	38.88	28.75	46.16	38.7	39.51
3rd Qu.	53.00	56.00	49.00	56.00	60.00	41.00	60.00	50.0	49.00
88%	60.00	69.00	59.00	61.00	76.00	57.00	71.00	56.0	55.00
Max.	93.00	98.00	100.00	100.00	100.00	100.00	99.00	89.0	90.00
st.d.	13.88	23.88	18.72	25.97	27.00	21.50	19.39	16.20	14.46

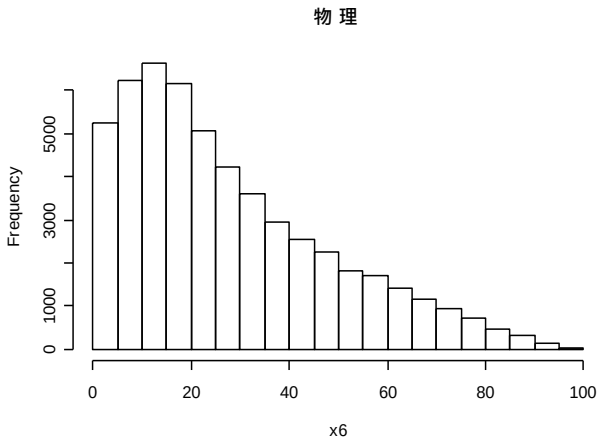

國文
Frequency
0
5000
10000
15000
0
20
40
60
80
x1

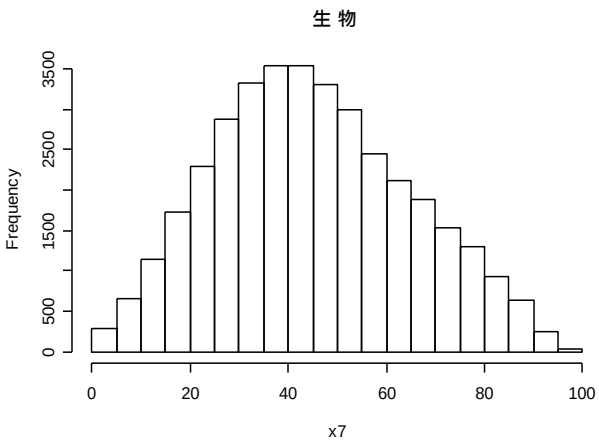

英文
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
x2

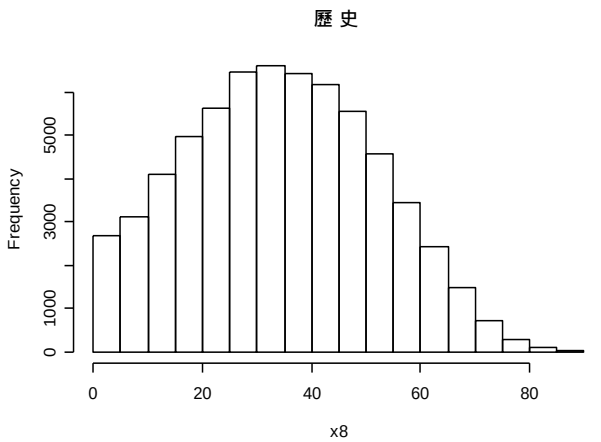

數學甲
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
x3

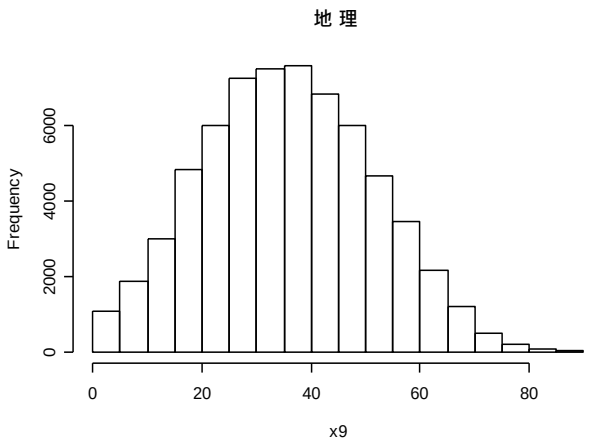

數學乙
Frequency
0
2000
6000
10000
0
20
40
60
80
100
x4

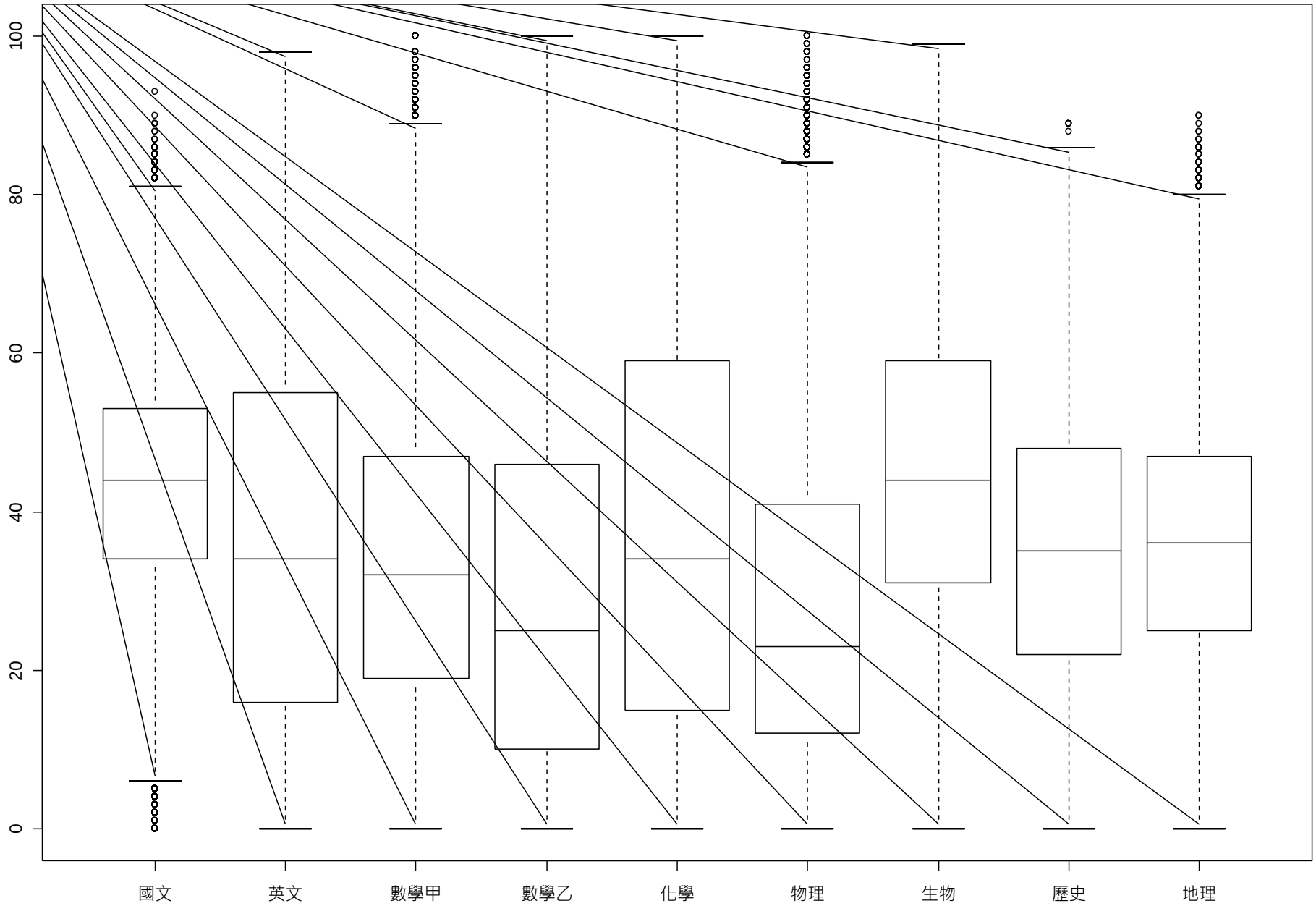

化學
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
x5

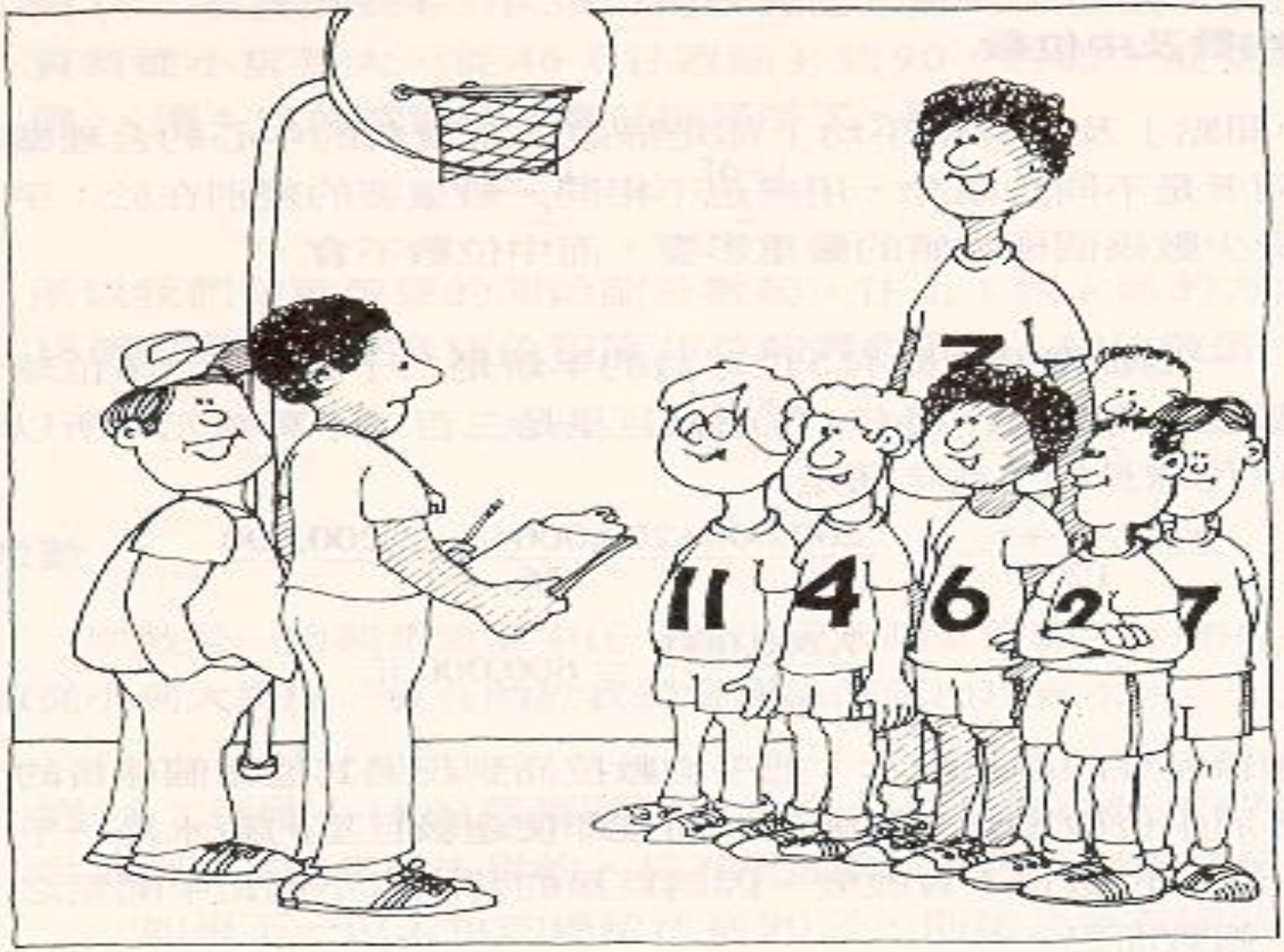

物理
Frequency
0
1000
3000
5000
0
20
40
60
80
100
x6


生物
Frequency
0
500
1500
2500
3500
0
20
40
60
80
100
x7


歷史
Frequency
0
1000
3000
5000
0
20
40
60
80
x8


地理
Frequency
0
2000
4000
6000
0
20
40
60
80
x9


100
80
60
40
20
0
國文
英文
數學甲
數學乙
化學
物理
生物
歷史
地理



「我們是應該宣布我們的平均高度來嚇死對手，還是宣布我們的中位數高度來消除他們的戒心呢？」

資料分析的方向

- 視研究目的，分析大致分為兩個角度：

→尋找資料的整體趨勢；

→偵測較為異常的現象。

- 舉例而言：

→整體趨勢包括平均數、變異數、相關係數等，能反映整筆資料特質的數值。

→異常現象包括異常觀察值（如：離群值）、整體特性的改變、資料是否同質等。

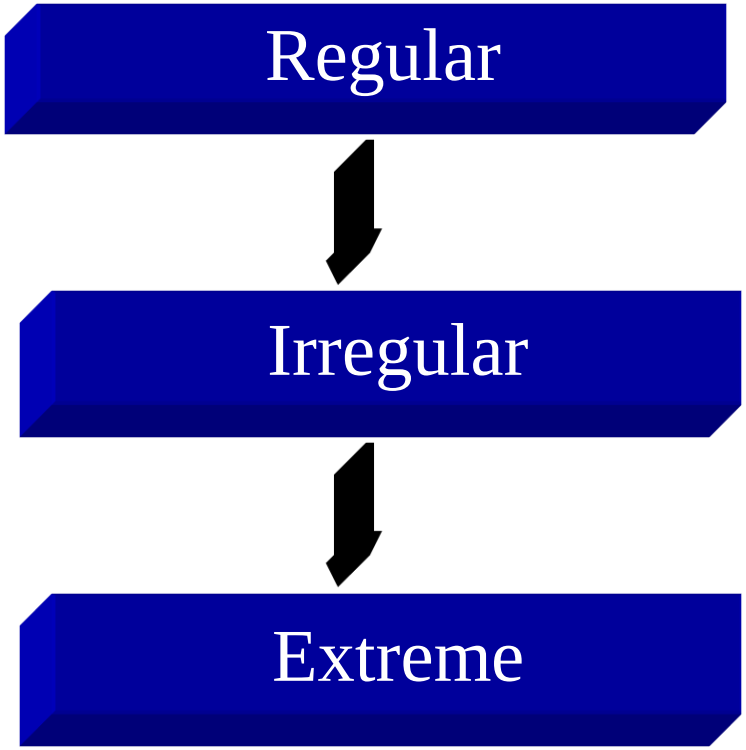
整體趨勢分析

- 教科書資料的數量通常較少，很少出現異常觀察值，但實際資料經常會有「意外」，需要視情況而定，調整分析步驟及項目。
- 整體趨勢的分析可仿造「集中趨勢」及「散佈趨勢」，計算具有代表性的數據，接著再輔以圖形、表格，以另一角度驗證這些結果，作為進一步分析的參考。

→問題：若統計數據與圖表不一致，如何處理？

統計與知識

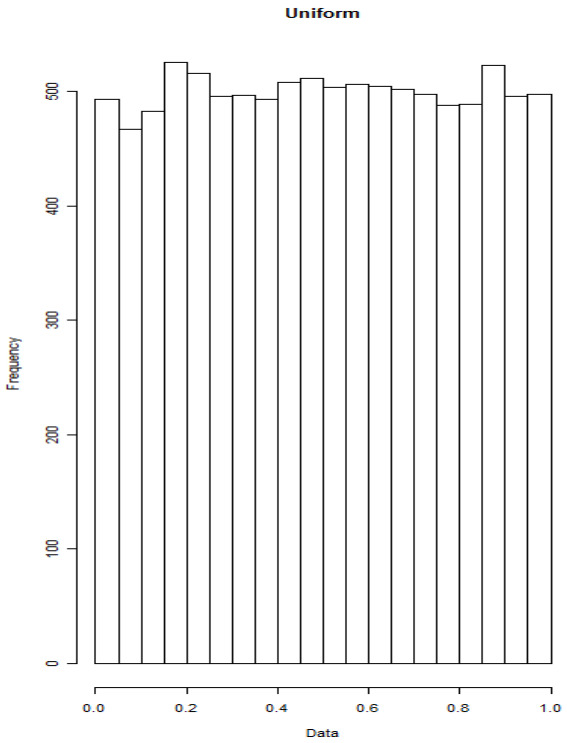
- 統計整理資訊的方法屬於歸納法(Induction)，從龐雜的資料找出共同趨勢，並區分資料具有以下哪一種特性：

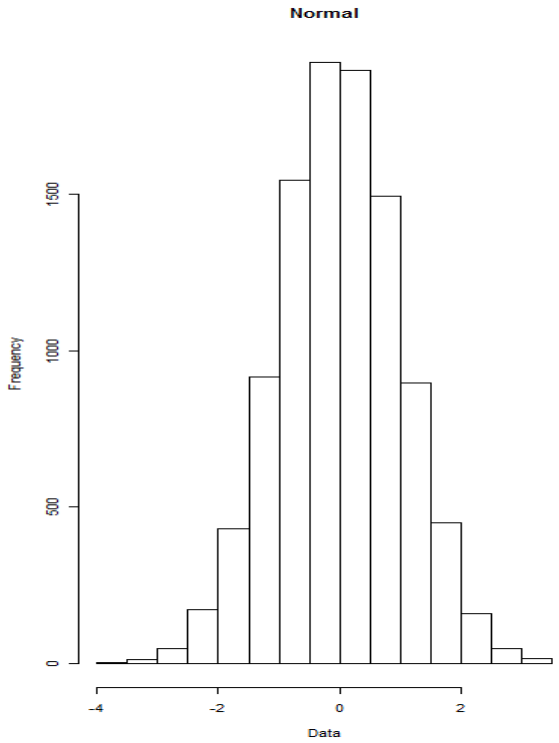


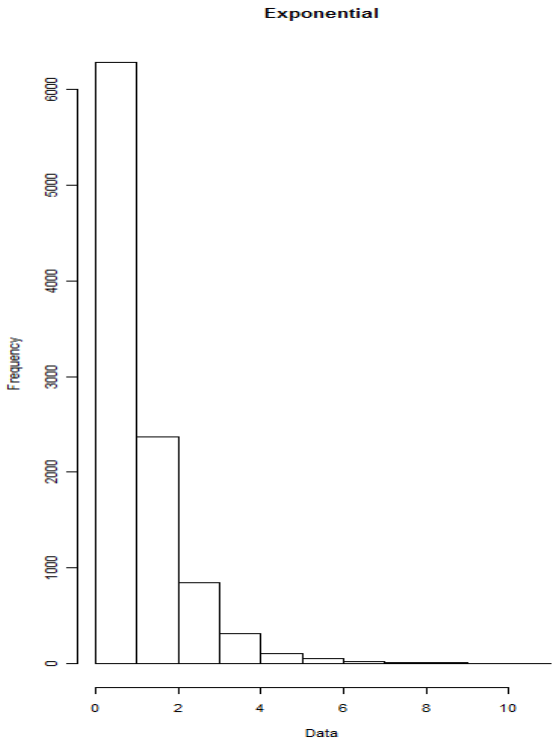
例題：區隔不同分配的資料

- 如何區隔來自連續型均勻分配、常態分配、指數分配的資料？

→下圖為10,000個亂數繪出的Histogram。







選擇有代表性的統計數據

- 如果有足夠觀察值，藉由Histogram足以區分這三個分配；但若資料量不足，可透過統計量確認觀察值的特性。

→首先可比較平均數、中位數，若兩者差異大（以標準差判斷），資料應屬指數分配。

→常態分配較均勻分配更集中，四分位數間距離較為一致(Min, Q1, Median, Q3, Max)。

註：另一種可能是藉助於Boxplot。

基本統計量

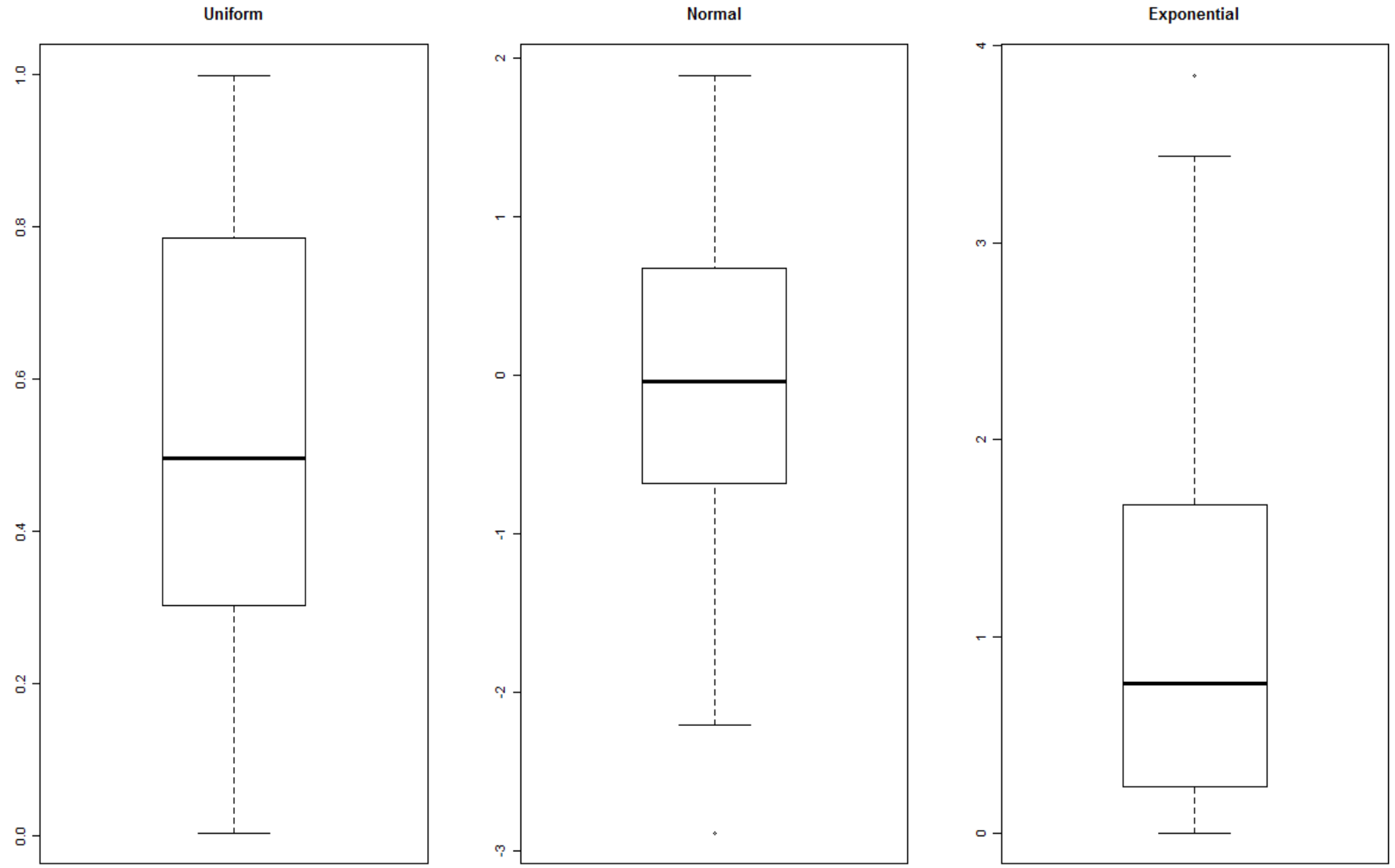
- 以下是三種分配100個亂數的基本統計量：

(1) Min. 1st Qu. **Median** **Mean** 3rd Qu. Max. St.d.
.0037 .1967 **.4577** **.4601** .6920 .9707 .2821

(2) Min. 1st Qu. **Median** **Mean** 3rd Qu. Max. St.d.
-2.2060 -.4512 **.1167** **.1298** .9329 2.2570 .9507

(3) Min. 1st Qu. **Median** **Mean** 3rd Qu. Max. St.d.
.0214 .2483 **.5749** **.9590** 1.2900 4.2170 .9856

註：第三筆資料明顯右偏；第一筆資料比第二筆資料更為「均勻」。



註：上述圖形為100個亂數的結果。

Example: Singer Heights Story

- Each singer in the NY Choral Society in 1979 self-reported his or her height to the nearest inch. Their voice parts in order from highest pitch to lowest pitch are Soprano, Alto, Tenor, Bass. The first two are typically sung by female voices and the last two by male voices.

→What is the best way(s) to describe the heights of these singers?

Note: You may use find the data from web site

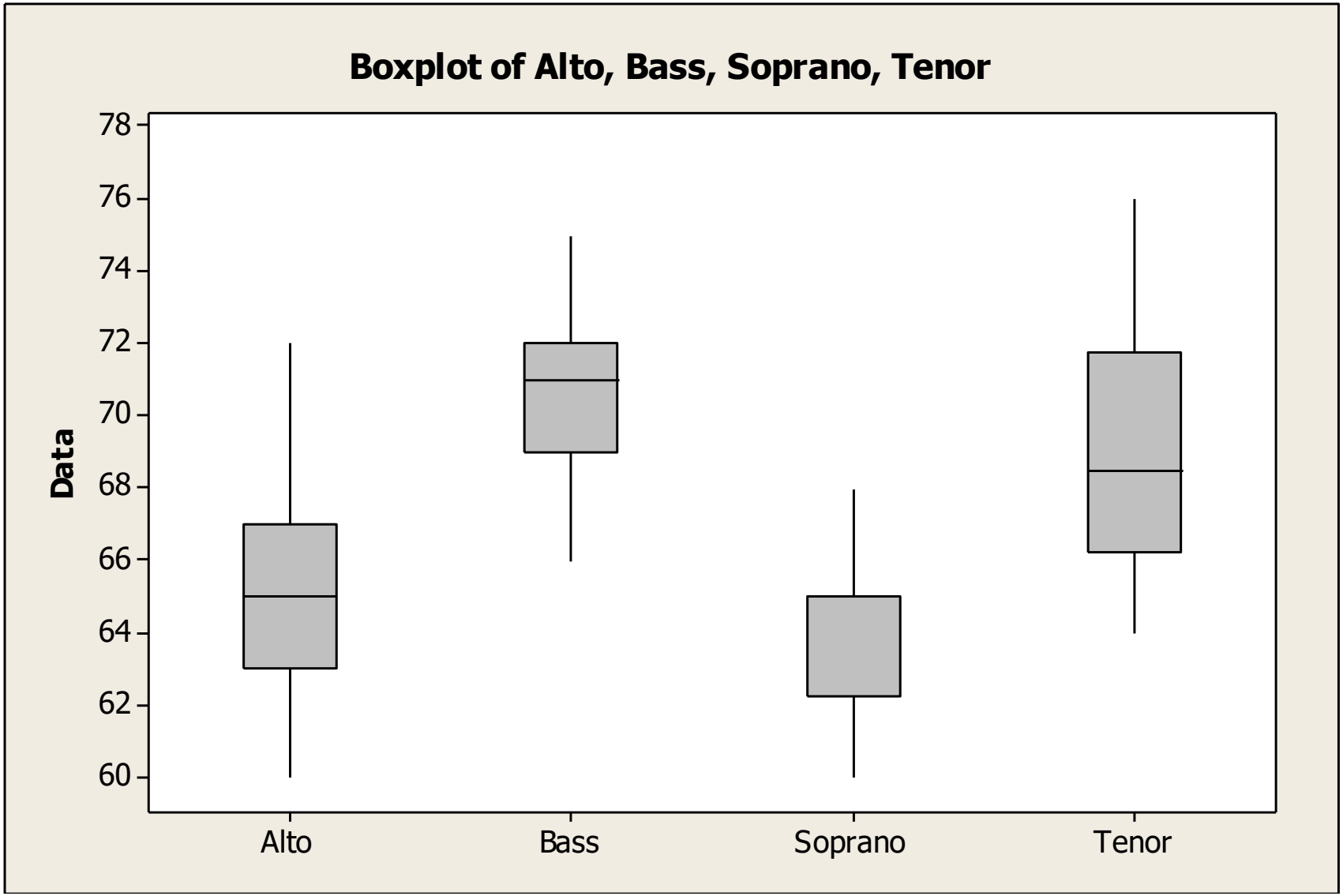
http://lib.stat.cmu.edu/DASL/

Some Descriptive Statistics

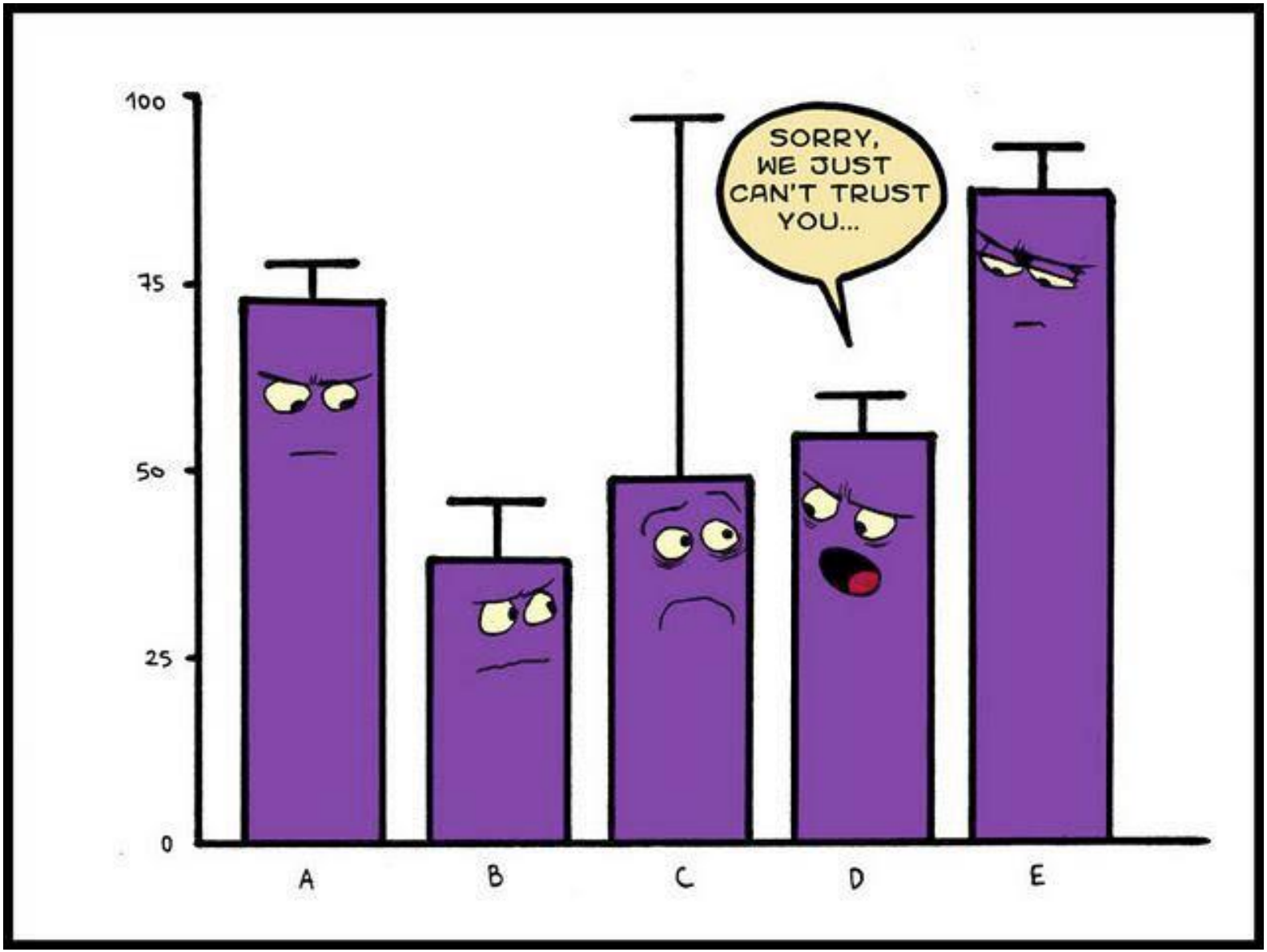
Variable	N	N*	Mean	SE Mean	StDev
Soprano	36	0	64.250	0.312	1.873
Alto	35	0	64.886	0.472	2.795
Tenor	20	0	69.150	0.719	3.216
Bass	39	0	70.718	0.378	2.361

Variable	Minimum	Q1	Median	Q3	Maximum
Soprano	60.000	62.250	65.000	65.000	68.000
Alto	60.000	63.000	65.000	67.000	72.000
Tenor	64.000	66.250	68.500	71.750	76.000
Bass	66.000	69.000	71.000	72.000	75.000

Compare the Differences!



非我族類其心必異！



敘述統計量(範例)

例題一、試以文字詮釋以下隨機抽出某公司業務部門20位員工的年齡：

41	25	25	33	27	31	42
35	36	32	36	41	34	29
34	31	34	35	32	35	

→ 平均數＝33.4，中位數＝34.0，標準差＝4.75，全距＝17。

敘述統計量(續)

例題二、試以文字詮釋以下隨機抽出某公司20位員工去年請假的天數：

0 0 0 0 0 0 0 0 1 1

1 2 2 3 4 5 5 6 7 42

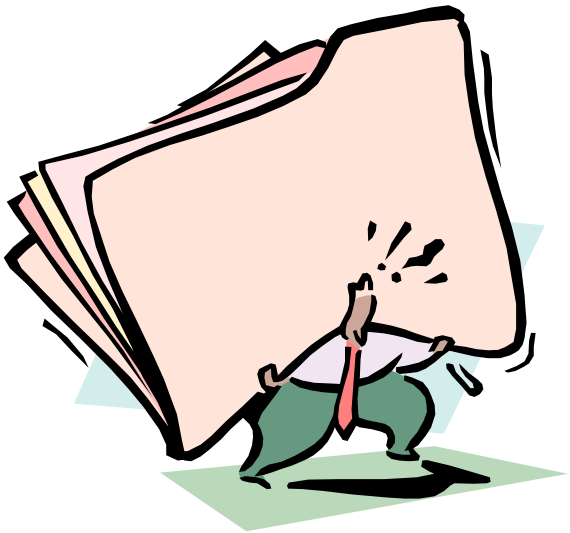
→ 你/妳 看到了甚麼現象？

敘述統計量(續)

例題三、街頭隨機訪問20位成年受訪者去年閱讀某月刊的期數：

0 1 11 0 0 0 2 12 0 0

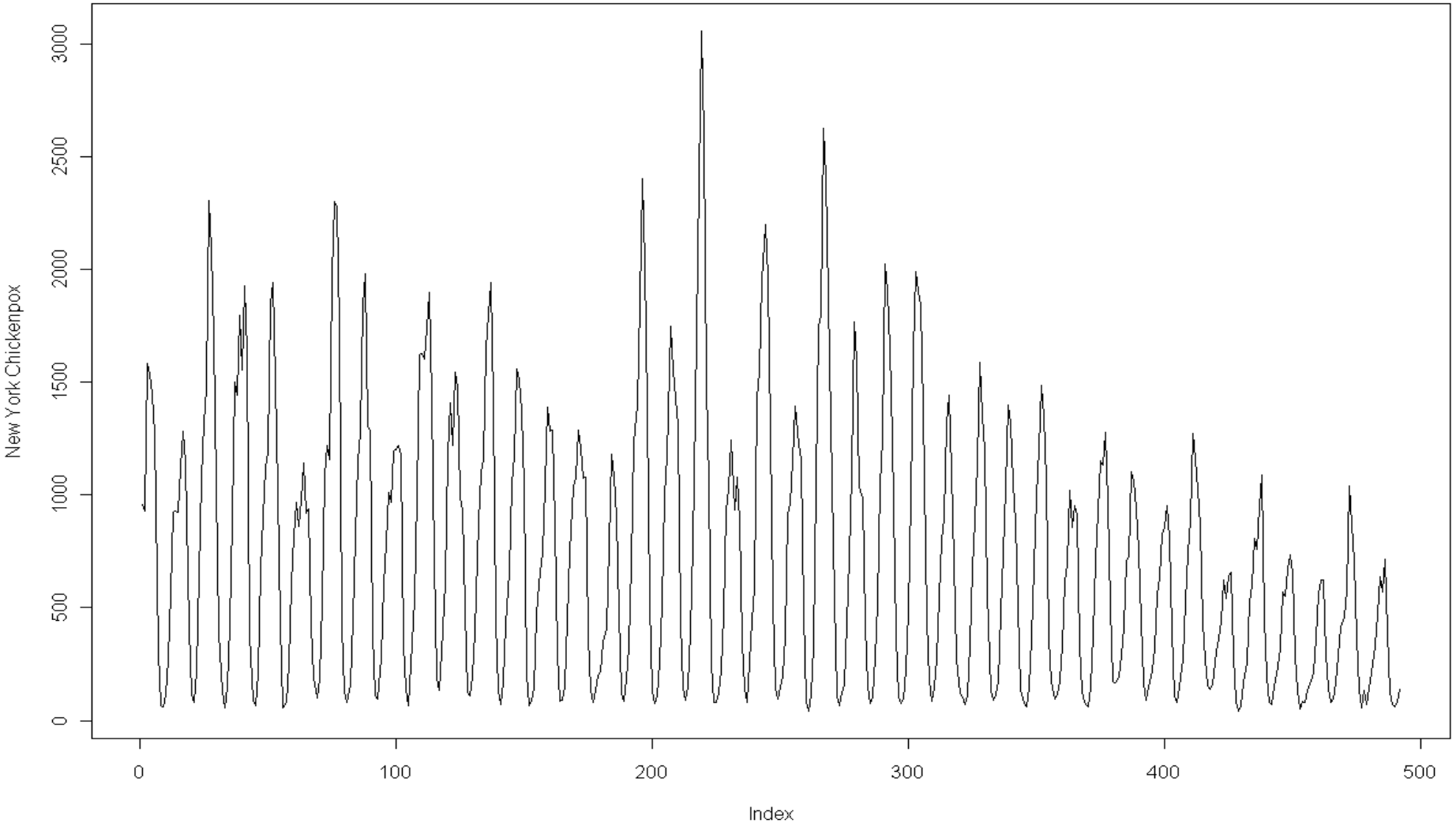
12 1 0 0 0 0 12 0 11 0

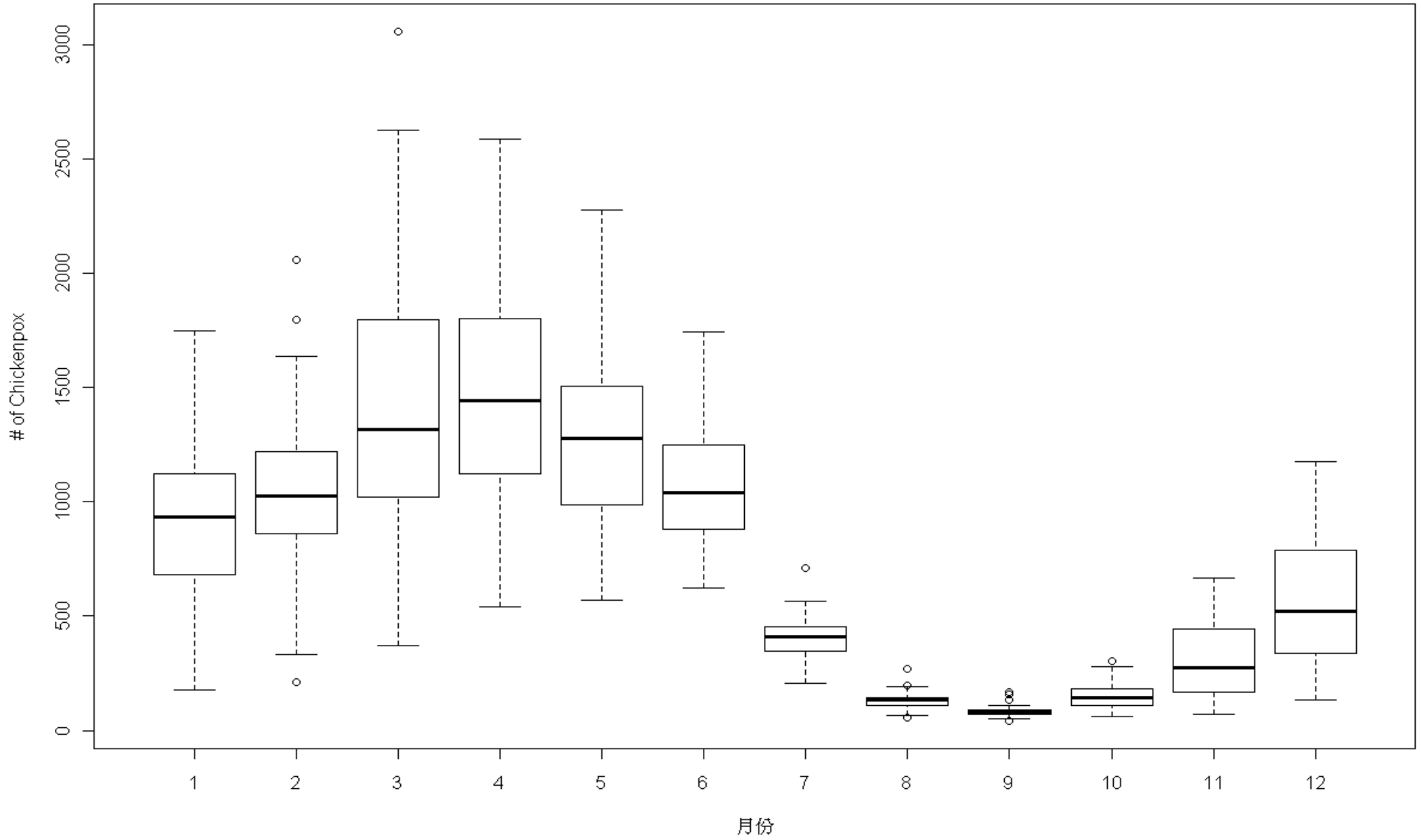


→ 請問這是甚麼樣的月刊？

- 如何分析與時間有關的資料？

→下圖為美國紐約1931-1972每月水痘發病數。





→明顯可知春天病例數較多、秋天較少。

相關性分析

- 相關性分析的主旨在於找出數字大小的差異或關係，例如：

→比較平均數(Z或t 檢定)、變異數的大小

→線性相關係數(Correlation Coefficient：Pearson, Spearman, Kendall) 、獨立性檢定

- 相關性分析需先區隔資料類型，連續型及類別型資料的處理方式不同，混合使用會得到不合理的結果。

- 例如：以下為某項調查的兩個問項：

1.請問您平均一週到現代連鎖餐飲店用餐次數：

__1. 0次 __2. 1～3次 __3. 3～6次

__4. 6～10次 __5. 10次以上

2.請問您覺得現代連鎖餐飲店的價格如何？

__1.高很多 __2.高一些 __3.差不多

__4.低一些 __5.低很多

→計算出兩者的相關係數為0.08，兩者間似乎不相關。

相關性分析(續)

- 有些問卷會要求受訪者填寫0到9(或1到5)的數字，代表喜好(或贊成)的程度。

→填寫9分者是填寫1分者的9倍。(合理嗎？)

第2題：價格

第1題：次數

	1	2	3	4	加總
1	34	47	4	2	87
2	67	199	14	0	280
3	3	22	2	0	27
4	1	3	0	0	4
5	1	1	0	0	2
加總	106	272	20	2	400

卡方檢定

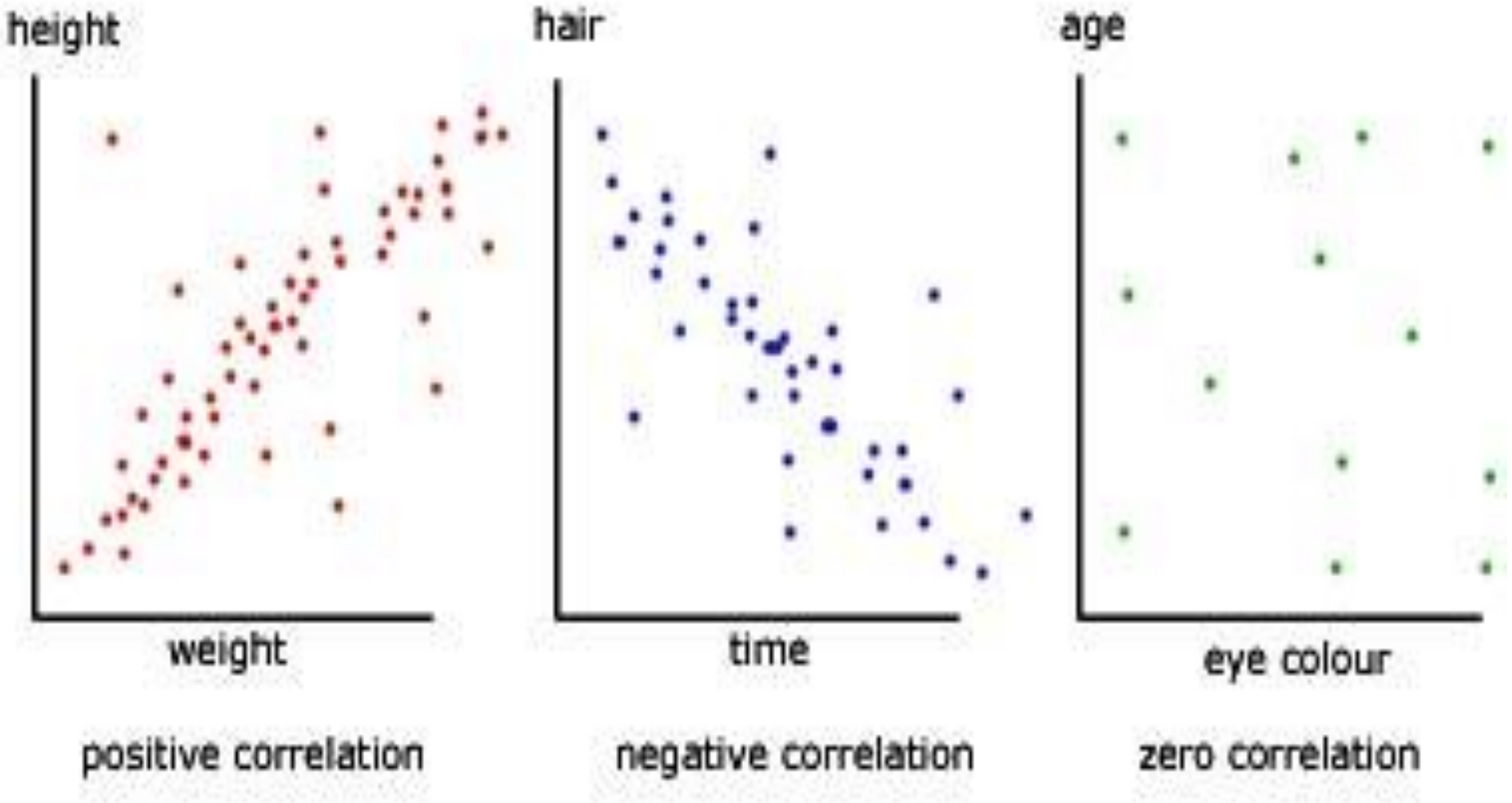
- 可能的合併方式：

	1	2	3以上	加總
1	34	47	6	87
2	67	199	14	280
3以上	5	26	2	33
加總	106	272	22	400

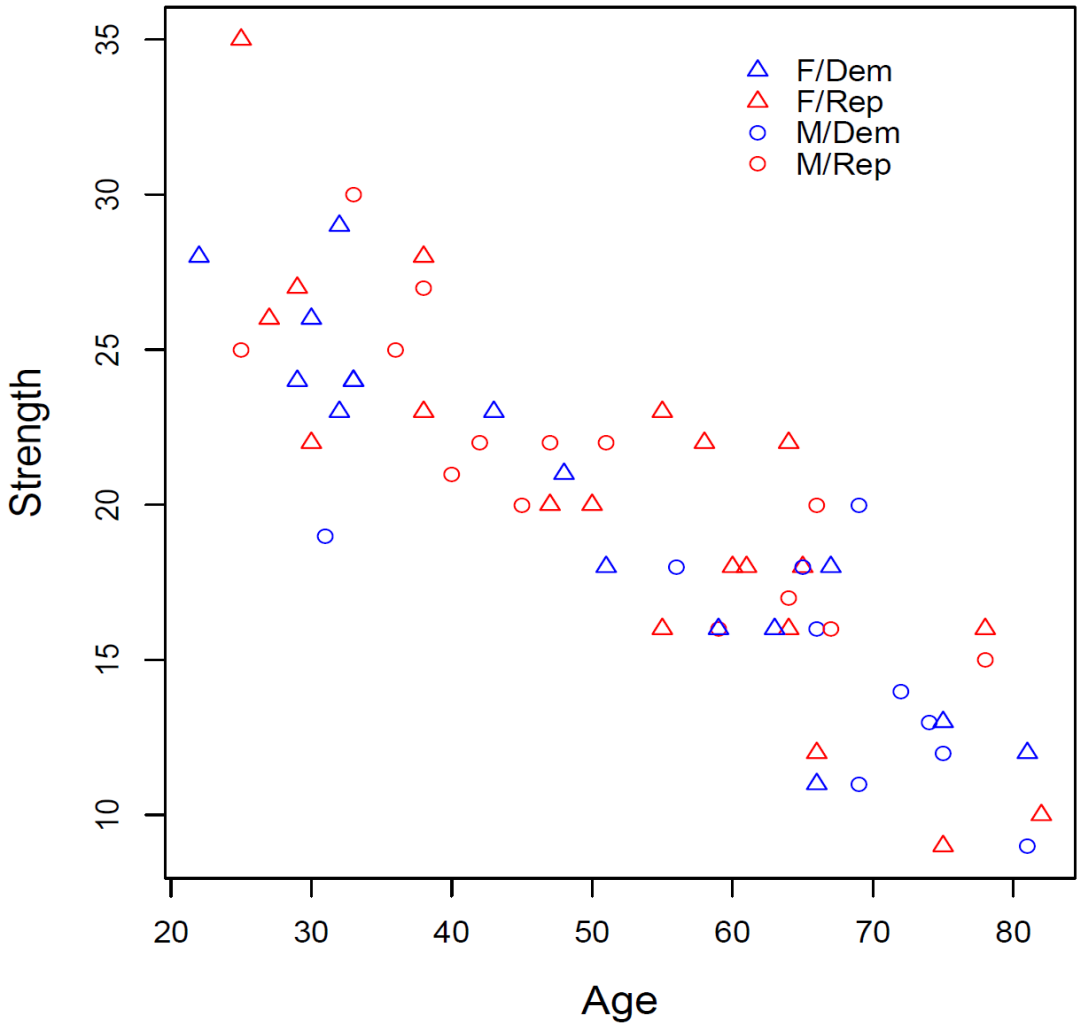
→分析結果：行與列有顯著的關係！

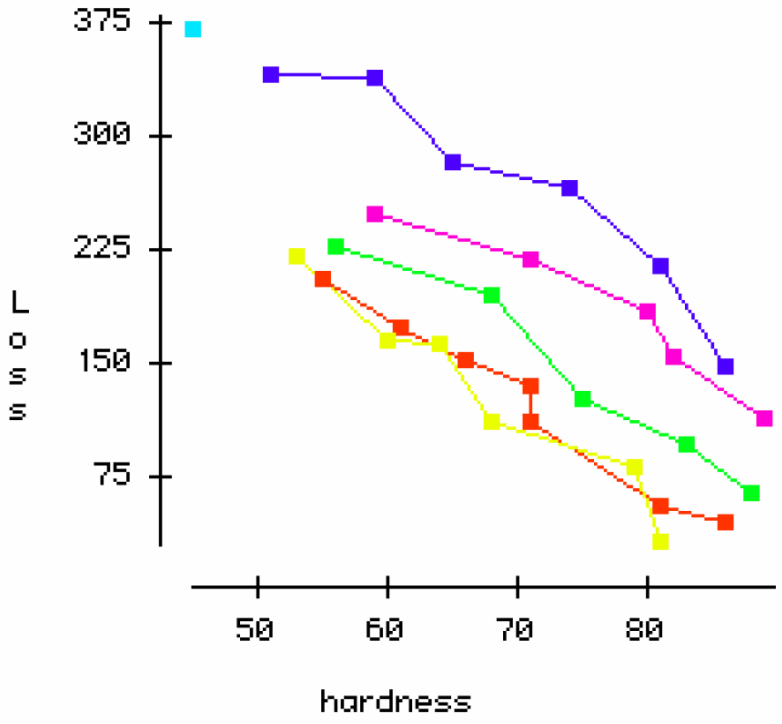
價格 vs.次數

散佈圖(Scatter Plot)



散佈圖也可用於多重分類！

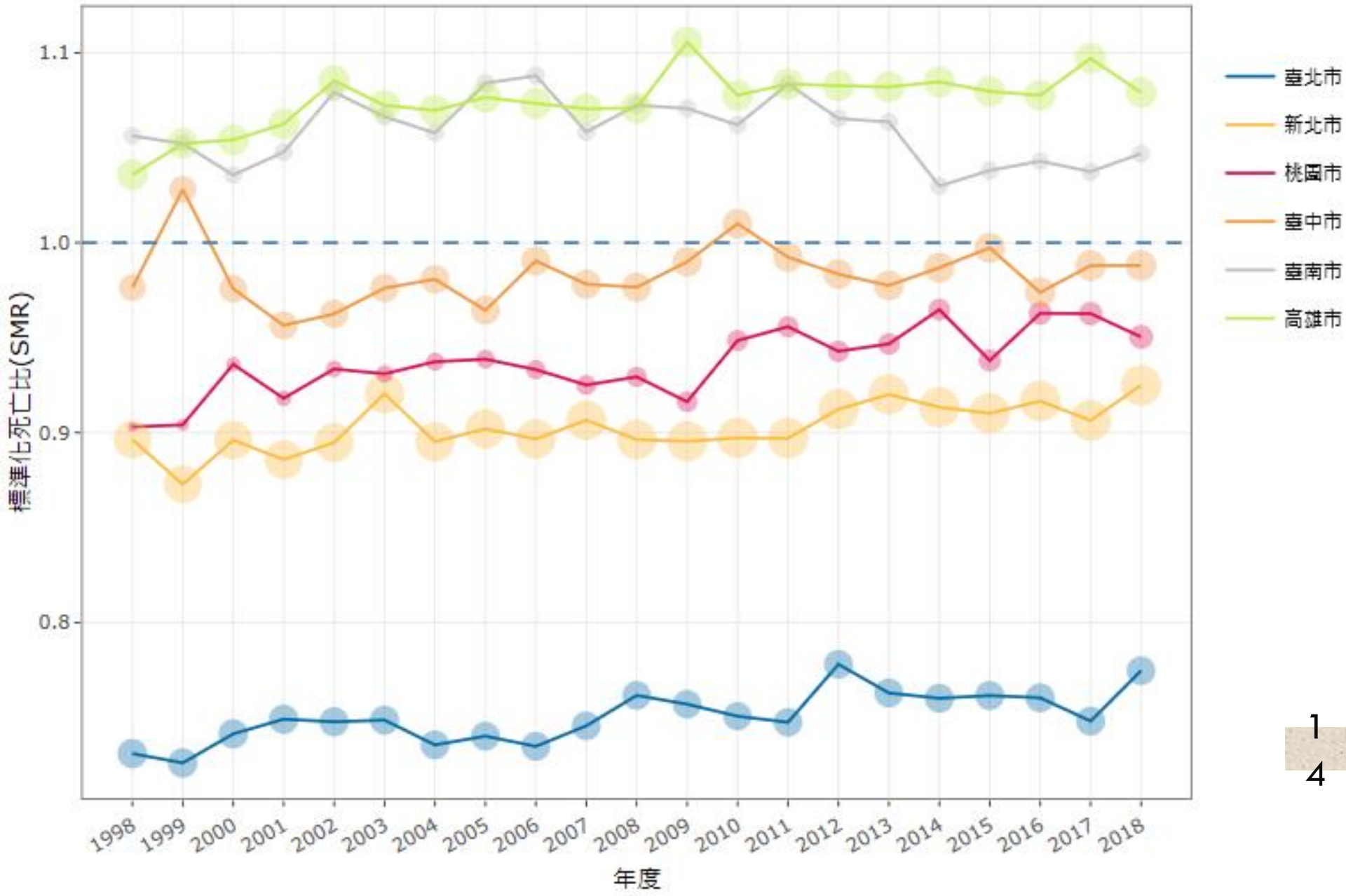




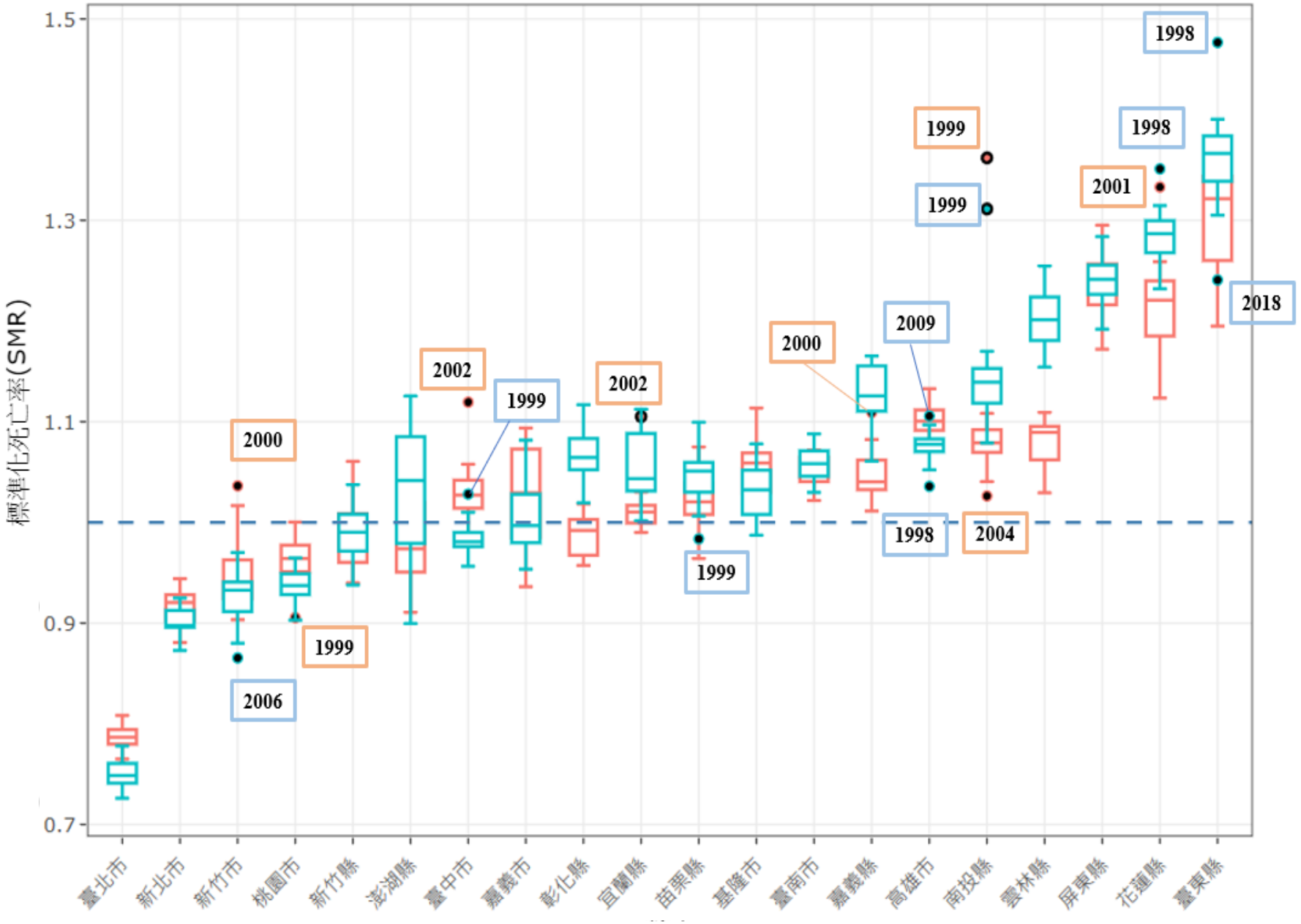
資料視覺化(Data Visualization)

https://boostlabs.com/wp-content/uploads/2019/09/10-types-of-data-visualization-1.jpg

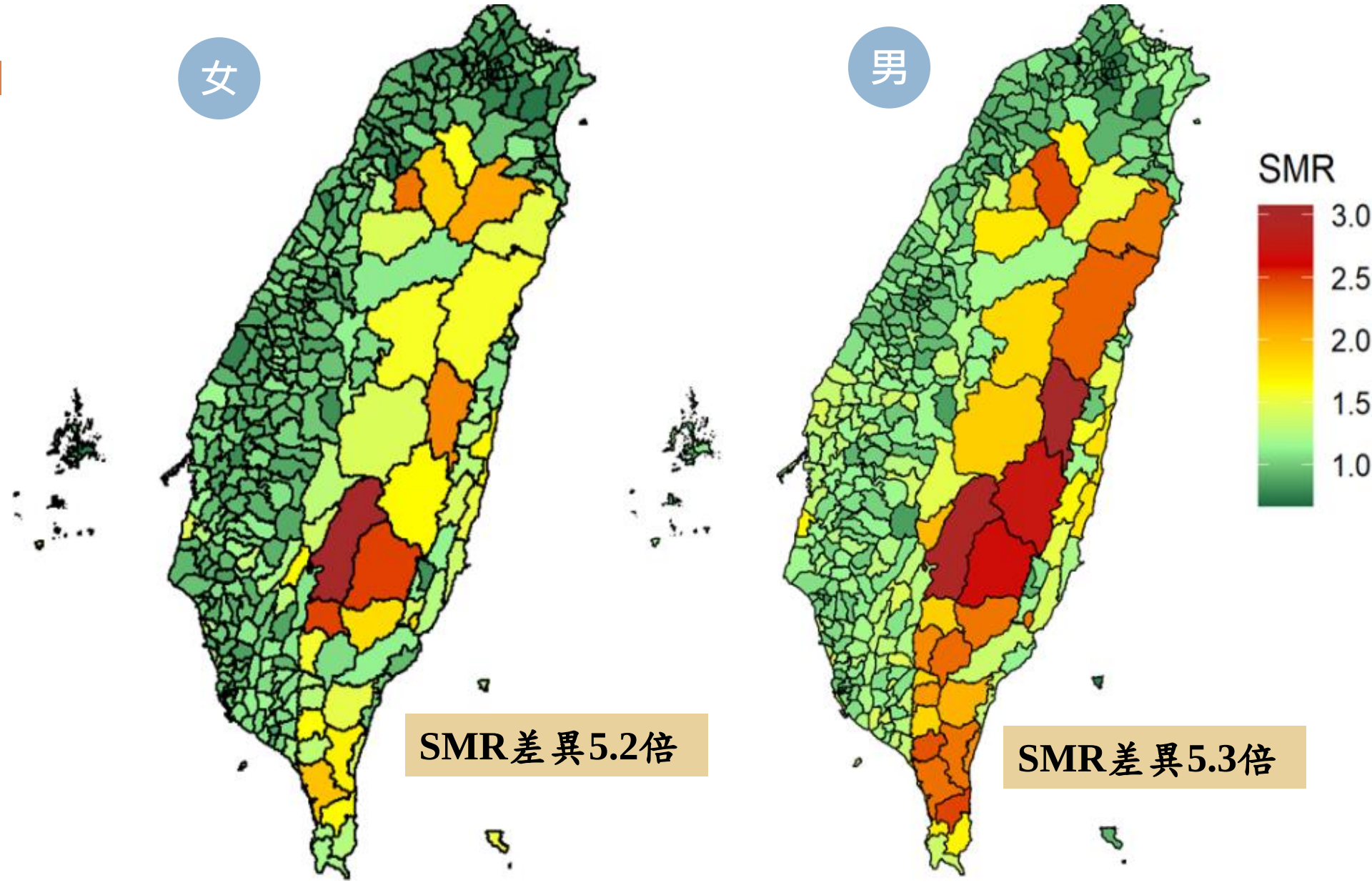
六都1998～2018年男性標準死亡比（Bubble Plot）



1974～2018年標準死亡比（Boxplot）

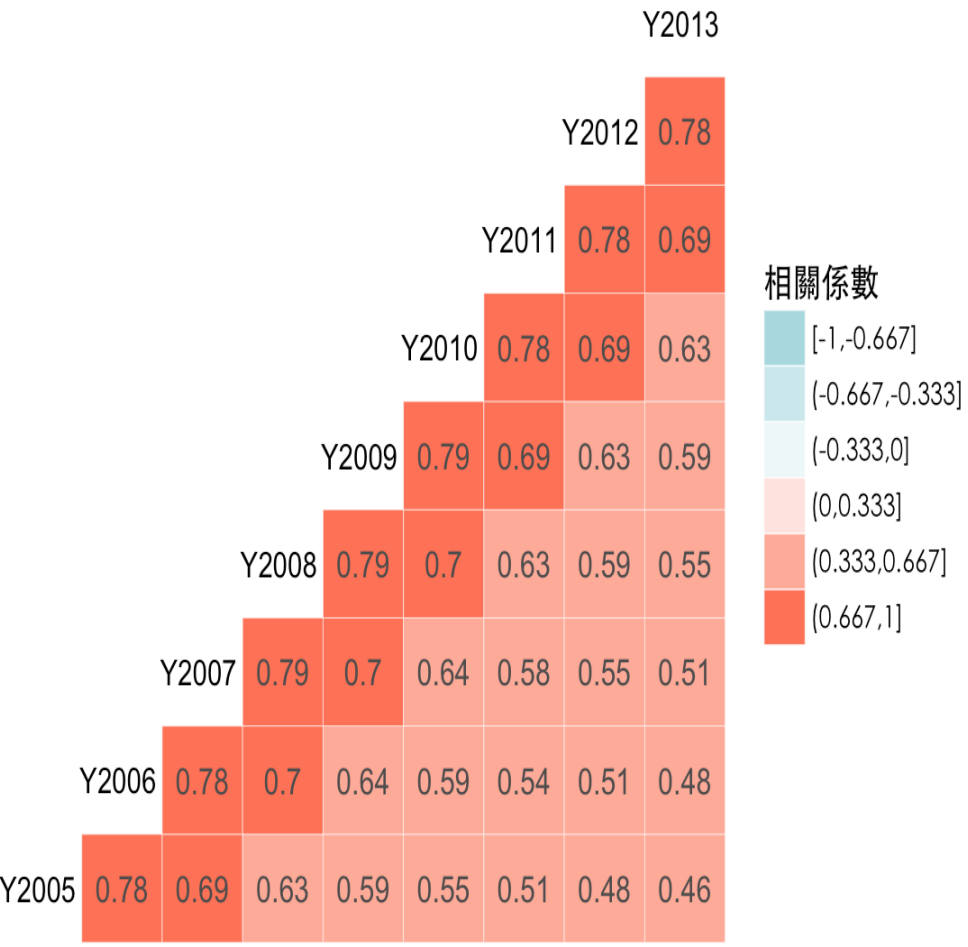


2018年SMR（標準死亡比）差異倍數



臺灣高齡人口的醫療利用

各年度就醫次數相關係數矩陣



各年度就醫金額相關係數矩陣