統計學 Statistics for Business & Economics

參考書籍：《統計學》David R. Anderson等原作；陳可杰，黃聯海譯。

本筆記由 國立台北科技大學 109資工 黃漢軒所撰寫，只用於教育用途，不做任何商業行為。

若侵權請聯繫t109590031@ntut.org.tw 或 sigtunatw@gmail.com，非常感謝。

> 這份筆記是看書自學並且撰寫下來的筆記，因此內容可能不是100%正確。
>
> 若你覺得/肯定哪邊有錯，歡迎聯繫我。

Bug Report

reporter	issues	approve
amyxinrou	Section 1.8 電腦與統計「方法」應該是「分析」	☑

目錄

Chapter 1 資料與統計

Section 1.1 商業與經濟上的應用

應用	遇到的問題/想要達成的事情	利用統計來解決遇到問題的方式
會計上	為客戶稽核帳目時，因為應收帳款資料數量龐大，逐筆驗證勢必耗時費力且昂貴。	審計員選擇一部份的帳目，稱之為樣本，檢閱樣本帳目的正確性後，便可決定是否接受資產負債表的應收帳款總數。
財務上	給出投資上面的建議	檢閱包括本益比及現金殖利率在內的各式財務資料，藉由比較個別股票和整體股票市場平均值的資訊，就能決定此股票是否為好的投資標的，來幫助財務分析師針對股票做出買進買出或繼續持有的建議。
行銷上	對於行銷做研究	電子掃描器可以蒐集資料，透過購買雜貨店的銷售點掃描資料，處理資料後再將匯整的統計資料出售給製造商，品牌經理檢視銷貨及促銷活動的統計資料後，就能夠分析在眾多商品建立未來的行銷策略。
生產上	監控製成的產出	透過x-bar圖可用來監控平均產出，只要樣本平均值在管制圖的管制上限與管制下限之間，表示生產製程在管制內，可以繼續生產。
經濟上	預測未來的經濟狀況或發展相關趨勢	運用許多統計資訊進行預測，例如物價指數或失業率和產能利用率來預估通貨膨脹率，將這些指標輸入可以預測通貨膨脹率的電腦預測模型，就能夠得到預測值。
資訊系統上	管理組織內電腦網路的日常運作	利用統計資訊可以協助評估電腦網路的效能，有助於系統管理者更瞭解電腦網路。

Section 1.2 資料

TABLE 1.1 Data Set for 60 Nations in the World Trade Organization

Nation	WTO Status	Per Capita GDP ($)	Fitch Rating	Fitch Outlook
Armenia	Member	3,615	BB−	Stable
Australia	Member	49,755	AAA	Stable
Austria	Member	44,758	AAA	Stable
Azerbaijan	Observer	3,879	BBB−	Stable
Bahrain	Member	22,579	BBB	Stable
Belgium	Member	41,271	AA	Stable
Brazil	Member	8,650	BBB	Stable
Bulgaria	Member	7,469	BBB−	Stable
Canada	Member	42,349	AAA	Stable
Cape Verde	Member	2,998	B+	Stable
Chile	Member	13,793	A+	Stable
China	Member	8,123	A+	Stable
Colombia	Member	5,806	BBB−	Stable
Costa Rica	Member	11,825	BB+	Stable
Croatia	Member	12,149	BBB−	Negative
Cyprus	Member	23,541	B	Negative
Czech Republic	Member	18,484	A+	Stable
Denmark	Member	53,579	AAA	Stable
Ecuador	Member	6,019	B−	Positive
Egypt	Member	3,478	B	Negative
El Salvador	Member	4,224	BB	Negative
Estonia	Member	17,737	A+	Stable
France	Member	36,857	AAA	Negative
Georgia	Member	3,866	BB−	Stable
Germany	Member	42,161	AAA	Stable
Hungary	Member	12,820	BB+	Stable
Iceland	Member	60,530	BBB	Stable
Ireland	Member	64,175	BBB+	Stable
Israel	Member	37,181	A	Stable
Italy	Member	30,669	A−	Negative
Japan	Member	38,972	A+	Negative
Kazakhstan	Observer	7,715	BBB+	Stable
Kenya	Member	1,455	B+	Stable
Latvia	Member	14,071	BBB	Positive
Lebanon	Observer	8,257	B	Stable
Lithuania	Member	14,913	BBB	Stable
Malaysia	Member	9,508	A−	Stable
Mexico	Member	8,209	BBB	Stable
Peru	Member	6,049	BBB	Stable
Philippines	Member	2,951	BB+	Stable
Poland	Member	12,414	A−	Positive
Portugal	Member	19,872	BB+	Negative
South Korea	Member	27,539	AA−	Stable
Romania	Member	9,523	BBB−	Stable
Russia	Member	8,748	BBB	Stable
Rwanda	Member	703	B	Stable
Serbia	Observer	5,426	BB−	Negative
Singapore	Member	52,962	AAA	Stable
Slovakia	Member	16,530	A+	Stable

Introduce - 元素、變數及觀察值

觀察上方表格。

名詞	意義
資料	經由蒐集、分析及彙總所得到，作為說明與解釋之用的事實與數值。
資料集	為特定研究目的蒐集的所有資料，由許多元素所組成。
元素	資料蒐集的實體，包含很多變數，例如上方表格的每個國家即為一個元素
變數	元素的某一特性，例如上列表格的每個元素有以下四個變數：WTO狀態、GDP、Fitch Rating、Fitch Outlook
觀察值	對特定元素蒐集的一組衡量值就是觀察值，例如上表的第1個觀察值(Armenia)包含了一組衡量值：Member、3615、BB-及Stable

Introduce - 衡量尺度

資料蒐集需要以下衡量尺度之一：名目尺度、順序尺度、區間尺度及比例尺度。

衡量尺度決定資料包含的資訊量，也指出資料彙整的或統計分析時的最適方法。

名目尺度(nominal scale)

用來表示元素屬性的標記或名稱，比較等於或不等於。

例如上表的國家WTO狀態可以分成「是WTO會員國」與「是WTO觀察員」，因此我們可以以數字1表示這個國家是WTO會員國，2表示這個國家是WTO觀察員，就能夠方便把資料輸入電腦，兩個國家的WTO狀態只能用相同與否來區分。

也因為名目尺度的意義是比較等於或不等於，因此詢問「WTO會員國與WTO觀察員哪個比較大」或者「兩個國家的WTO狀態相加等於多少」是完全毫無意義的行為。

順序尺度(ordinal scale)

與名目尺度不同，順序尺度的類別有一定的大小或順序，比起名目尺度只能比較相等，順序尺度能夠比較大小。

例如上表的Fitch Rating，其中AAA代表最好，F代表最差，因此可以根據評等排出高低，所以是順序尺度。

區間尺度(interval scale)

若變數具有順序資料的特性，且觀察值可以相加或相減，其結果仍有意義，這個變數的衡量尺度就是區間尺度，且一定以數值表示。

例如統測成績就是一個區間尺度，假設有三位學生的統測成績為699、560、350，則我們可以由高到低依序排序來衡量出成績表現的優劣，而他們的差距也存在意義，例如699的學生比560的學生高出139分。

比例尺度(ratio scale)

若變數具有順序資料的特性，且觀察值可以加減乘除，其結果仍有意義，這個變數的衡量尺度就是比例尺度，且一定以數值表示。

與區間尺度的差別在於，**比例尺度要求絕對零點，也就是值必須要大於等於0且在0上必須要是自然的不存在**。

例如年齡不存在0歲，而高度不存在0公分，而可以描述20歲比5歲大4倍。

Introduce - 類別資料及定量資料

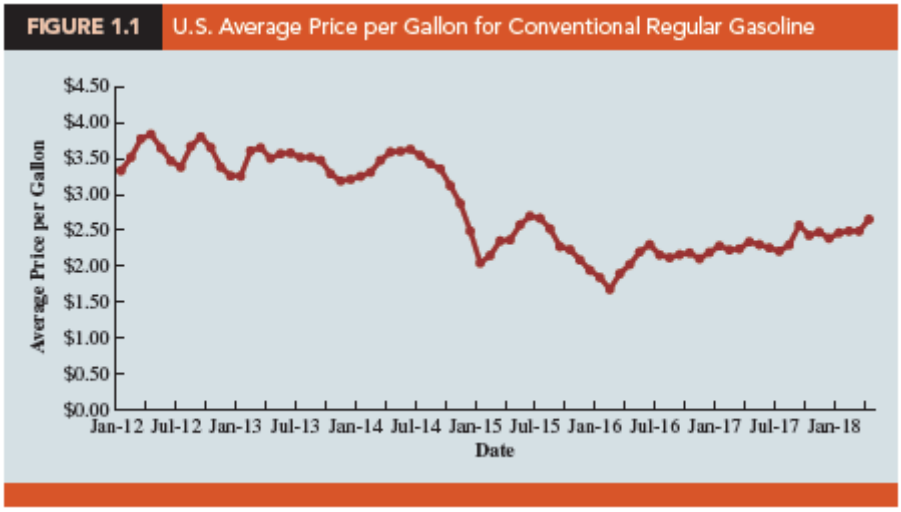
可以把資料分成類別資料與定量資料，類別資料使用名目尺度或順序尺度，而定量資料使用區間尺度與比例尺度。

其中類別變數是類別資料的變數，定量變數是定量資料的變數，算術運算對於定量變數是具有意義的(見以上範例)。

Introduce - 橫斷面資料及時間序列資料

橫斷面資料是在相同或幾乎相同時點所蒐集的資料，例如上表是相同時間點的60個世界貿易組織會員國的5個變數的資料。

而時間序列資料則是數個不同時期的資料，例如以下的折線圖。



這張圖顯示了2012年到2018年傳統普通汽油的每公升平均價格。

Section 1.3 資料來源

Introduce - 資料來源

資料可來自既有資料，或透過觀察研究、實驗設計的方式取得。

既有資料

在有些情況下，可能已有特定應用所需的資料，也可以從專門蒐機與維護資料的組織獲得大量有關商業與經濟的資料。

而網際網路也是資料與統計資訊的重要來源，例如許多公司均已設立網站，並在網站上公布銷售額、員工人數、產品數量等等的資訊。

政府機關也是另一個既有資料的重要來源，例如台灣有公共運輸整合系統流通服務平台，讓大眾可以更容易利用台灣的運輸工具資料。

觀察研究

觀察研究只觀察特定環境發生的事情，對一個或多個感興趣的變數紀錄資料，再對資料進行統計分析。

例如，化妝品銷售業者在街上訪問隨機選擇的顧客，蒐集化妝品的變數資料例如使用頻率，價格，品牌等等。

民調也是一種觀察研究，民調公司透過隨機選擇民眾進行電訪，來預測台灣大選的結果。

實驗

觀察研究與實驗的關鍵差異在於實驗必須在**控制**的條件下進行。

例如：臺灣民調想要根據年齡層與支持政黨的關係進行研究，為了取得這樣的資料，將這群人(樣本)以年齡分成不同的群體並根據提出的答案進行研究。

統計學處理的實驗類型，通常要先找出感興趣的變數，接著找出一個或更多的變數並加以控制，因此可以得到其他變數如何影響研究人員感興趣的主要變數的資料。

Introduce - 時間與成本的議題

想利用資料與統計分析來幫助制定政策，必須清楚取得資料所需花費的時間與成本。

若時間緊迫，則利用既有資料較可行，若重要資料無法得自既有來源，就必須考慮或取資料額外所需花費的時間與成本，取得資料與隨之而來的統計分析所花費的成本，不應超過協助決策時所創造的效益。

Introduce - 資料取得的錯誤

管理者應該隨時注意統計研究中資料錯誤的可能性，使用錯誤資料比完全不使用這些資料來得更糟。

只要取得的資料值與經過正確程序取得的真實資料值不符合，就會發生資料取得的錯誤，例如年紀將27歲記成21歲，或者受訪者沒有理解題意就做出毫無相關的回答。

這些資料可藉由特別程序來檢查資料的內部一致性，例如檢查異常大或異常小資料數值(離群值)，或者在檢查程序中找出7歲但職業是大學的資料。

Section 1.4 敘述統計

Introduce - 利用表格與圖表來做敘述統計

考慮表格上刊登所有資料很難讓人看懂某個方向的趨勢。

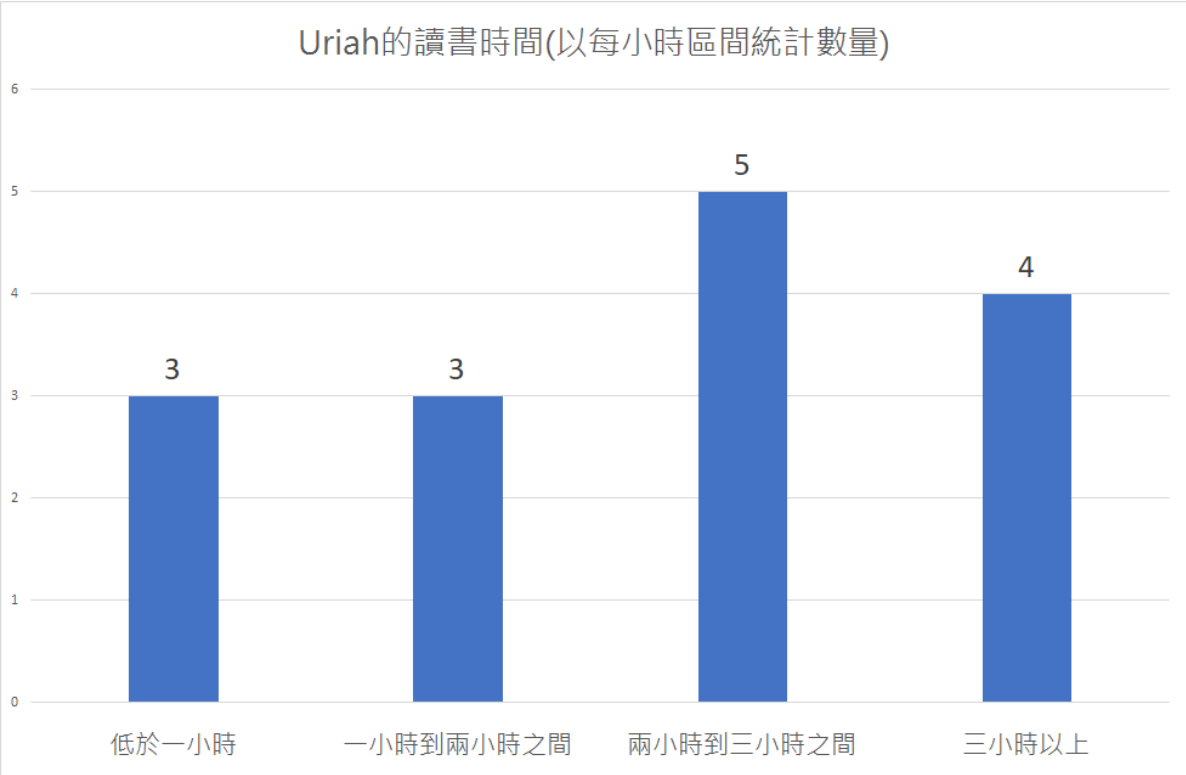
大部分刊登在媒體上，公司年報或其他出版品的統計資訊，是以讓人容易瞭解的資料形式來彙整公布，稱為敘述統計。

例如以下表格呈現了Uriah在15天的讀書時間。

日期	讀書時間(分鐘)
9月1日	249
9月2日	266
9月3日	212
9月4日	289
9月5日	255
9月6日	185
9月7日	191
9月8日	238
9月9日	221
9月10日	263
9月11日	219
9月12日	292
9月13日	261
9月14日	227
9月15日	218

則透過以表格的統計，我們想要根據每小時的單位來統計Uriah的讀書時間，則可以得到以下長條圖與圖表。

Uriah的讀書時間區間	次數
低於一小時	3
一小時到兩小時之間	3
兩小時到三小時之間	5
三小時以上	4



由此可知Uriah的讀書時間大部分以兩小時到三小時之間為多數。

Introduce - 利用數值來做敘述統計

我們通常也可以利用數值來做敘述統計，最常用的衡量值是平均數。

我們可以將Uriah的讀書時間做加總除以15天，來得到Uriah每天讀書的平均時間。

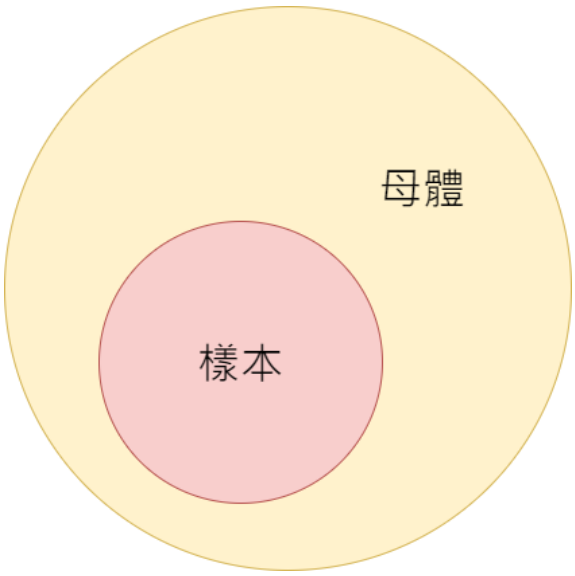
透過計算，可以得出Uriah每天讀書約2個小時24分鐘，平均數可以顯示資料集的中央趨勢或資料集的中央位置。

Section 1.5 統計推論

Introduce - 母體與樣本

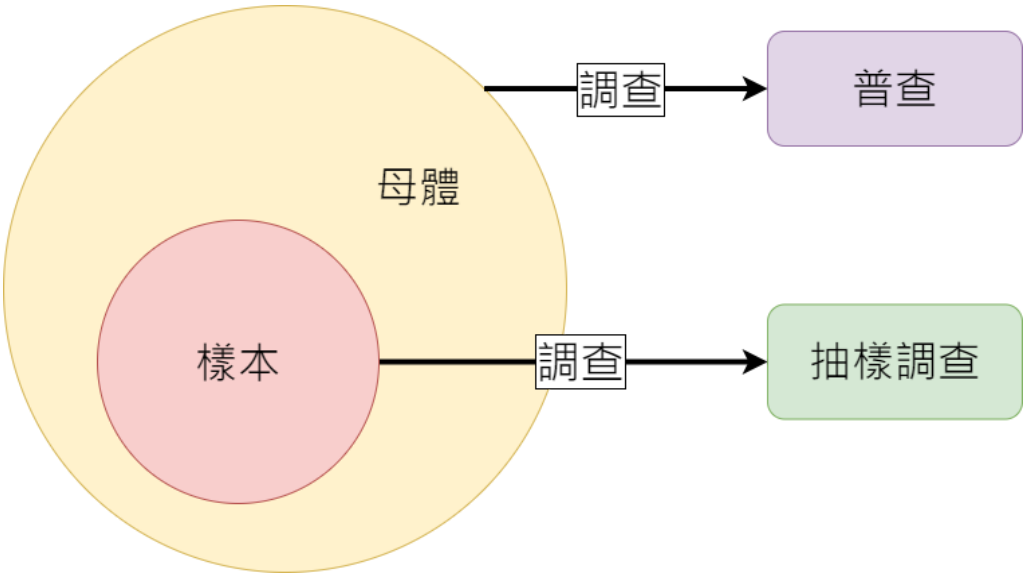
許多情況必須針對大量的元素來蒐集資料，但因為時間與成本的關係，僅僅只能蒐集到資料的一小部分。

在一個研究中，所有元素之集合稱為母體，而母體的部分集合又稱為樣本，如下圖。



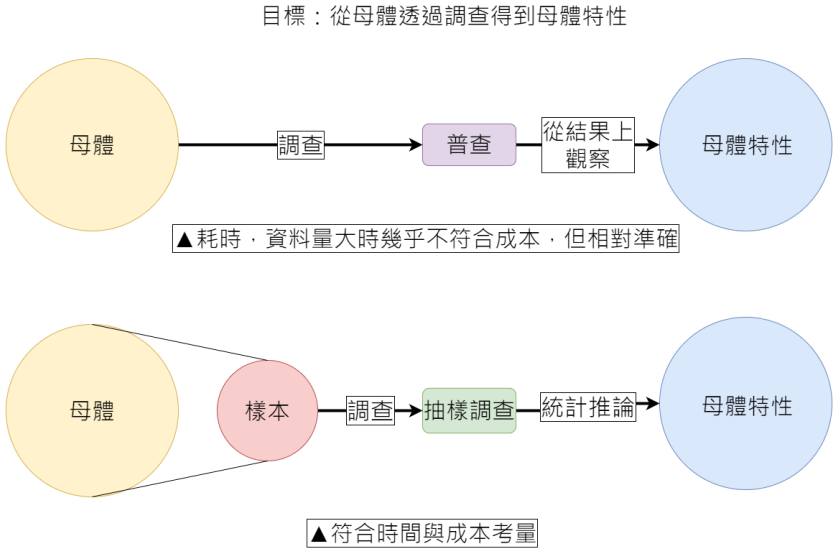
Introduce - 普查與抽樣調查

蒐集整個母體的資料進行調查稱為普查，而從樣本中進行調查稱為抽樣調查。



Introduce - 統計推論

在最省時間與最省成本的情況下，即為希望能夠從樣本中進行調查，進而推估出母體的特性做估計與假設檢定，則稱為統計推論。



Example - 北科學生的生活費

由於北科學生近萬人，如果要逐一調查北科學生的生活費，是一件非常困難的事情。

因此我們可以透過從萬人抽出約100人的樣本調查北科學生的生活費。

18000	7000	6000	7000	7000	16000	9000	8000	14000	18000
13000	10000	16000	18000	17000	11000	11000	8000	17000	7000
8000	16000	17000	15000	17000	18000	8000	14000	12000	6000
18000	6000	14000	12000	18000	17000	9000	11000	17000	13000
14000	15000	17000	18000	13000	16000	8000	6000	9000	7000
10000	15000	12000	12000	13000	10000	13000	9000	13000	14000
18000	13000	12000	9000	8000	7000	18000	17000	10000	8000
17000	17000	14000	12000	14000	15000	17000	8000	7000	11000
16000	18000	16000	8000	11000	14000	10000	16000	18000	12000
12000	12000	8000	13000	11000	11000	14000	9000	17000	13000

因此我們可以知道從這100個學生的平均生活費可得知，北科學生的生活費大概平均都在12590元左右。

而當抽樣數越多則越接近北科學生的真實平均生活費狀況。

Example - Uriah的飲料店

Uriah在北科綠光開了一間專賣鮮奶茶的飲料店，並且因為好喝到爆炸(?)所以每天都湧進上萬人進來購買。

手做飲料幾乎是不可能的事情，所以Uriah把飲料透過機器封裝，每天裝了上萬瓶500ml飲料。

但因為機器封裝飲料的會因為一些神奇的因素，所以讓每瓶飲料都有可能會有一定的誤差，不一定會剛好500ml。

Uriah把其中一箱飲料拿起來當作抽樣，並且統計這箱飲料的容量，得到了以下的數據(精確到小數點第二位)

```
500.48  504.75  501.89  502.51  496.52  495.61  504.84  502.71  495.88  497.75
499.02  503.88  497.43  499.30  496.47  497.89  498.58  502.63  502.39  499.62
497.82  499.92  497.42  495.77  495.28  503.94  498.34  498.37  498.58  495.33
501.10  500.46  502.89  497.31  503.17  503.05  500.94  502.53  496.80  500.98
504.41  497.94  497.89  502.63  504.42  495.37  498.81  499.27  500.50  503.70
496.30  504.74  498.02  500.96  496.48  498.34  498.01  503.21  501.00  504.75
496.28  498.49  499.95  498.31  500.42  497.87  504.14  497.26  498.54  502.98
503.70  503.40  502.88  504.87  502.22  497.78  502.21  496.04  499.54  504.50
500.70  495.77  504.18  497.81  497.19  496.84  495.15  498.87  504.89  501.42
495.10  498.38  500.18  498.76  495.95  496.22  500.90  499.47  499.19  500.81
```

得到了這箱飲料的容量平均是499.86ml。

Section 1.6 分析

Introduce - 分析

分析是將資料轉換成洞見以制定更加的決策的科學程序。

透過分析可藉由資料創造獨特觀點、提升預測能力、將風險量化、找出更好的決策方案。

一般來說，分析可以分成三大類別：敘述分析、預測分析、規範分析。

Introduce - 敘述分析

敘述分析用於描述過去發生的事實，例如資料查詢、報告、敘述統計、資料可視化、資料儀表板等等。

Example

你手上有一份針對於北科大學生喜歡聽音樂的資料。

資料中每個元素的變數含有學生姓名、學號、性別、喜歡聽的音樂種類(K-POP、J-POP、中文流行...)。

你想要探討性別與喜歡的音樂種類之間的關係，因此你將男生、女生分開討論喜歡聽的音樂種類並做成了長條圖(資料可視化)

做成長條圖的這一個步驟，就是一種敘述分析。

Introduce - 預測分析

預測分析的技術是以模型建構過去資料來預測未來，或評估一個變數對其他變數的影響。

Example

產品過去的銷售資料可以建立成一個數學模型，用來預測未來的銷售，這就是一種預測分析。

Introduce - 規範分析

與敘述分析、預測分析有很大的不同，規範分析是產生最佳行動方案的分析技術。

規範模型(最適化模型)是在已知限制條件下找出可使目標值極大或極小的解答，可用來產生使收益最大化的行動方案。

Example

Uriah的鮮奶茶店賣得很讚，於是Uriah想要知道制定價格與銷售數量之間的影響，用來決定是否漲價。

因此Uriah把銷售資料輸入進規範模型，規範模型會給出一個建議鮮奶茶的售價價格，來使Uriah的收益最大化。

Section 1.7 大數據與資料探勘

Introduce - 大數據

大數據一般指難以用現成軟體來管理、處理與分析的資料集。

可以將其定義成具有3V性質的資料：數量(volume)、速度(velocity)與多樣性(variety)。

其中數量是資料的總立、速度是指蒐集與處理資料的速率，而多樣性則是不同的資料類型。

Introduce - 資料探勘

資料探勘是從大型資料庫中開發出有利決策的資訊的方法，

運用來自統計、數學及電腦科學的綜合程序，分析人員探勘資料轉換成有用的資訊，稱為資料探勘。

資料探勘的技術相當依賴統計方法的技術，例如多元迴歸、相關及羅吉斯迴歸等等，

同時也需要將這些方法與涉及人工知會和機器學習的資訊科學做創造性的整合，使資料探勘變得更有效。

資料探勘也有一定的風險，例如過度配適模型會出現誤導性的關聯或因果關係結論。

因此需要謹慎地解釋資料探勘的結果，並且進行附加檢驗，對此避免這樣的風險會有所幫助。

Section 1.8 電腦與統計分析

統計學家會使用電腦軟體進行統計運算與分析，如果沒有電腦的幫忙，運算會相當複雜。

可以利用相關的軟體(Hadoop、SAS、SPSS)、程式(R、Python)來處理大數據。

Section 1.9 統計實務的倫理守則

統計學在蒐集、分析、呈現與解釋資料時扮演著重要的角色，因此統計學具有一定的倫理議題。

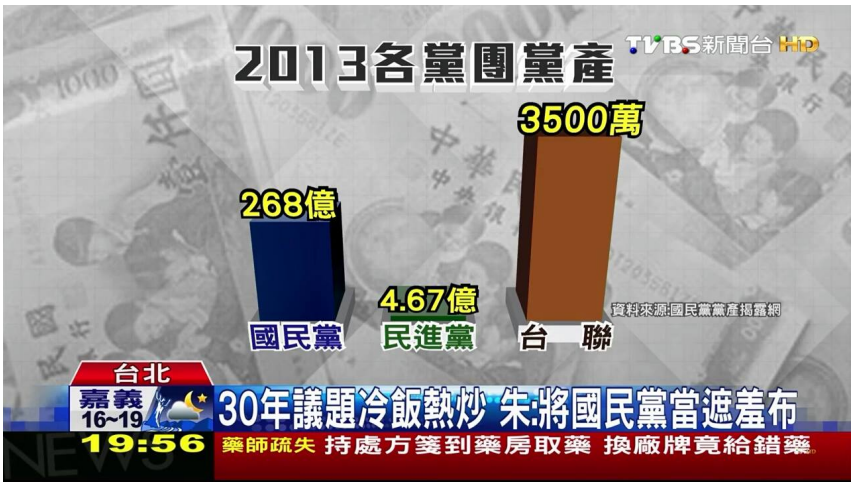
在統計學不適當的抽樣、不適當的資料分析、誤導的圖形、不適當的彙整統計資料及統計結果的偏差解釋，都是一種不道德行為。

因此當你開始自己的統計工作時，請使用公平、詳盡、客觀和中立的態度。

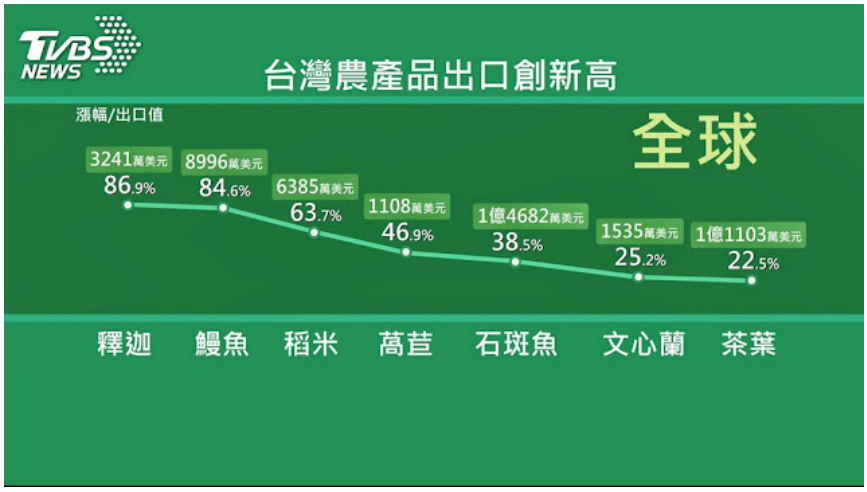
而作為統計資料的消費者，請抱持著懷疑論來檢視資料是好的，並且持續留意並瞭解資料來源，以及所提供的統計資料的目的與客觀性。

Example (誤導的圖形)

(268億 < 3500萬 ???)



(釋迦跟稻米根本毫無關聯，折線圖根本不是這樣用的)



Chapter 2 敘述統計：表格與圖形法

Section 2.1 類別資料的彙總

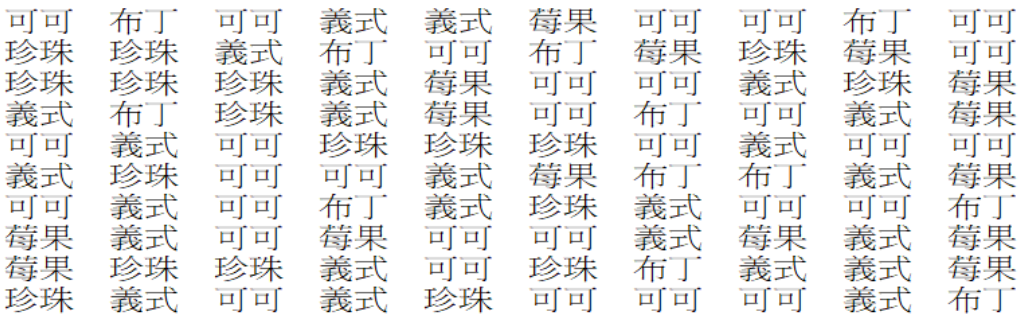
Introduce - 類別資料的次數分配

次數分配是資料的表格化彙總，用以顯示每一個不重疊類別或分組內的觀察值次數。

Example

Uriah的鮮奶茶茶鋪有5種不同的鮮奶茶，分別是莓果、可可、珍珠、布丁、義式。

Uriah彙整了100筆客人的購買紀錄，如下。



為了要建立出這些資料的次數分配，我們算出每種鮮奶茶出現在上表的次數，轉化成下表。

鮮奶茶類別	次數
布丁	12
莓果	15
珍珠	18
義式	25
可可	30

次數分配就幫助了我們彙整這100筆購買資料在這五種鮮奶茶的分布情形。

從這個數據就能知道在100個客人中，可可鮮奶茶較受顧客喜愛，其次是義式，能夠從上表知道五種飲品的人氣度。

Introduce - 相對次數分配與百分比次數分配

相對次數與百分比次數

一個資料組的相對次數代表該組別的觀察值次數的比率，如果一個資料集有n筆觀察值，則相對次數定義如下：

$$\text{一個組別的相對次數} = \frac{\text{該組別的次數}}{n}$$

百分比次數則為相對次數乘上100。

相對次數分配與百分比次數分配

相對次數分配是資料的表格化彙整，顯示每組的相對次數，百分比次數分配亦同，顯示每組的百分比次數。

以上表的Uriah的鮮奶茶茶鋪為例，我們可以轉化成以下的相對次數與百分比次數分配。

鮮奶茶種類	相對次數	百分比次數
布丁	0.12	12%
莓果	0.15	15%
珍珠	0.18	18%
義式	0.25	25%
可可	0.3	30%
總和	1	100%

從上表也能看出來，前三名相對次數的鮮奶茶種類總和佔這100筆購買次數的73%。

Introduce - 長條圖與圓形圖

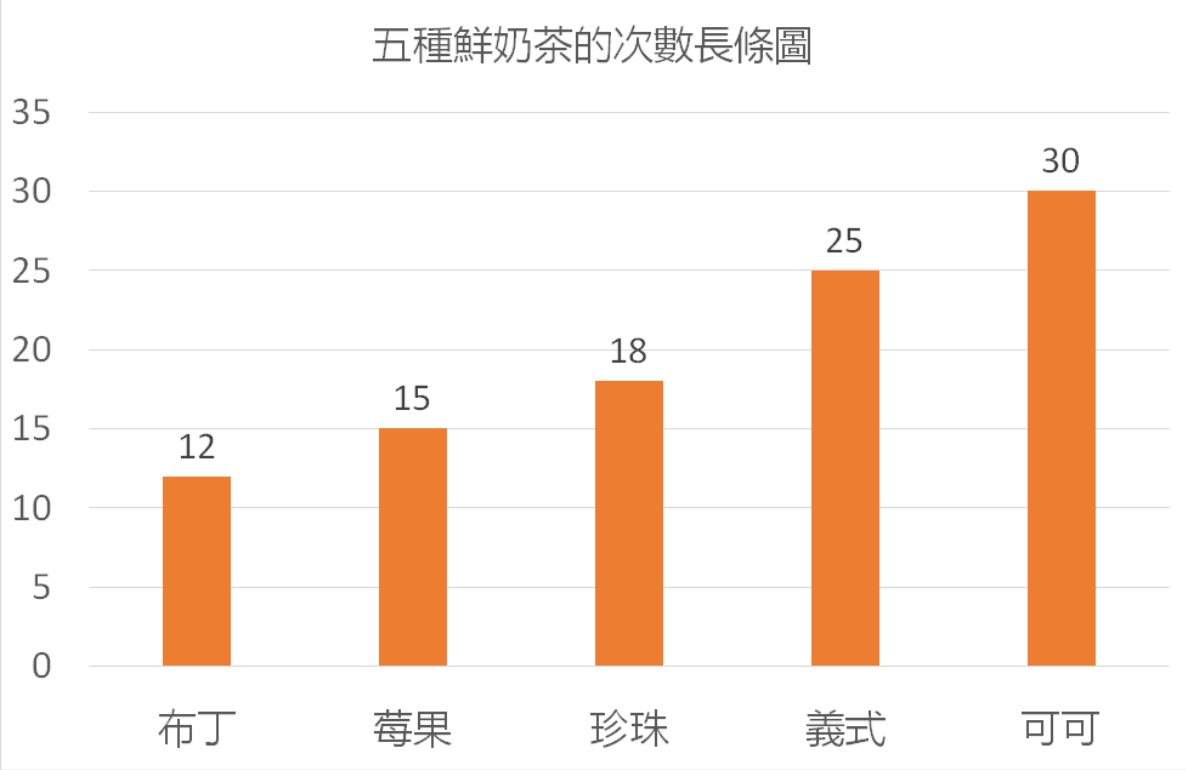
長條圖

長條圖是一種圖形，描述次數分配，相對次數分配或百分比次數分配進行彙總的資料。

圖形的一軸(通常來說，是橫軸)用來標記組別，另一軸則表示次數。

組別名稱上方有固定寬度的長條，高度表示次數。

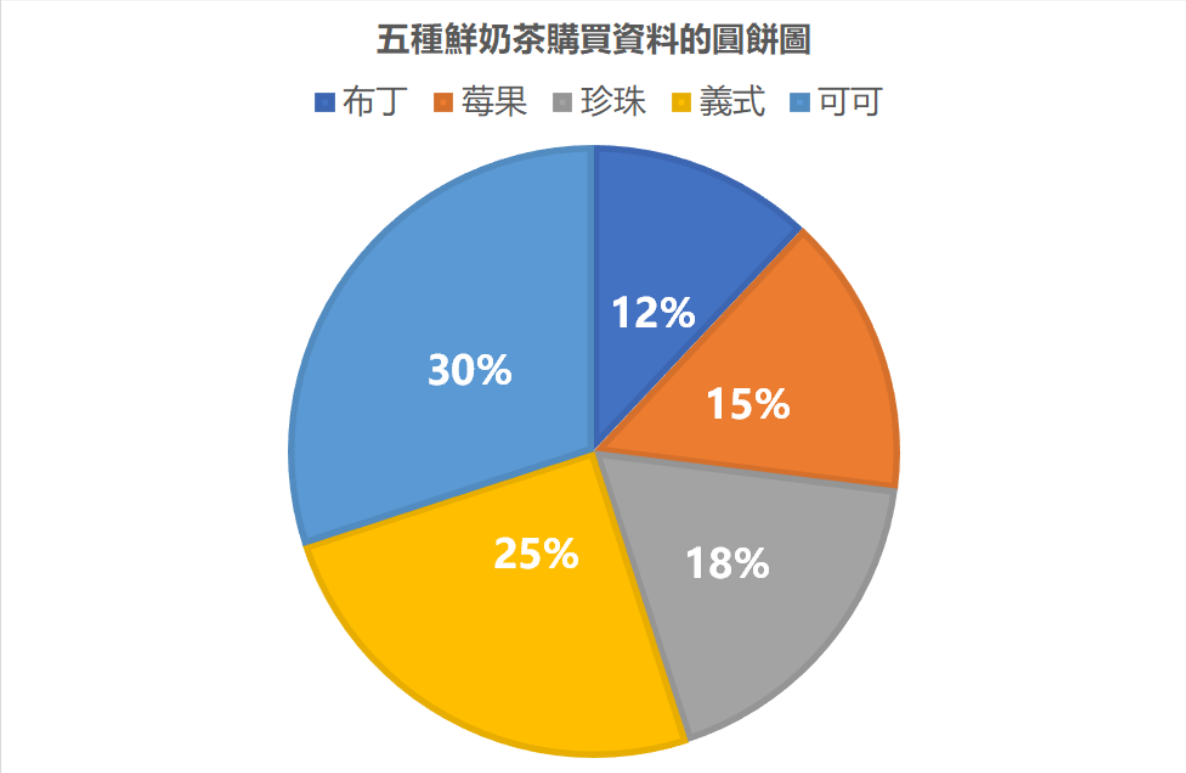
對類別資料而言，每個組別應分隔不相連，用來強調每個組別有所區隔，如下。



圓形圖

圓形圖是另一種表是類別資料的相對次數及百分比次數分配的圖形。

畫一個圓表示所有資料，再依各資料組的相對次數將圓形分為對應類別的扇形部分。



Section 2.2 定量資料的彙總

Introduce - 定量資料的次數分配

為定量資料建立次數分配有三個步驟：決定不相重疊的組數、決定每組的組距、決定每組的組限。

組數

由資料值的的範圍決定組數。

較多的資料需要較多的組數，分組時希望能夠以夠多的組數來顯示資料的變化。

但同時也不希望組數太多，以致於每個組別含有過少的資料數。

組距

組距通常會希望以相同的寬度，因此組距與組數相關，可以先根據資料的最大值與最小值，然後再決定組數。

我們可以使用以下的公式來計算組距：

$$\text{近似組距} = \frac{\text{最大資料值} - \text{最小資料值}}{\text{組數}}$$

求得的近似組距可以做微調，讓資料的歸納變得更加方便。

組限

選定組限之後，必須使每個資料只屬於唯一一組。

下組限是該組的最小可能值，而上組限則是最大可能值。

在類別資料的情況下，我們不用特別去規範上組限與下組限，因為類別資料必定能找到屬於自己的唯一一組。

但是在定量資料的情況下，我們必須規範上組限與下組限，以確定每個資料的歸屬。

組中點

組中點的定義即為上組限與下組限的中間值，如下。

$$\text{某組別的組中點} = \frac{\text{某組別的上組限} - \text{某組別的下組限}}{2}$$

Introduce - 相對次數分配與百分比次數分配

與類別資料相同，若資料有n個觀察值，則相對次數的定義如下。

$$\text{某組別的相對次數} = \frac{\text{該組別的次數}}{n}$$

而百分比次數則為相對次數乘上100。

Example - Uriah的作業排程

Uraih是一個就讀北科資工大二的學生，由於北科高中就是功課多到很著名，而Uriah又是超修怪，因此Uriah手邊堆滿許多作業

Uriah將每個功課仔細的規劃，並且列出了一個表，用來顯示每個作業所需要的時長。

(先不管這邊功課有40個，因為Uriah的肝近乎無敵，所以可以像中川龍一郎一樣工作72小時不睡覺)

19	15	20	03	18	18	01	04	09	12
27	06	06	24	11	26	25	26	27	24
07	12	11	01	09	20	18	01	26	29
26	19	13	03	02	21	10	13	07	21

由於有40筆資料，觀察到資料的最大觀測值是29，最小觀測值是1，因此Uriah打算把資料分成5組左右。

可以得到組數為5，組距為$\frac{29-1}{5} = 5.6$，因此Uriah打算以6當成組距。

接著Uriah想要先決定第一組的下組限，因此Uriah選定了1當成下組限，可以得到以下的列表

處理時間(日)	次數
01-07	11
08-14	9
15-21	6
22-28	7
29-35	7

由次數分配表可知：

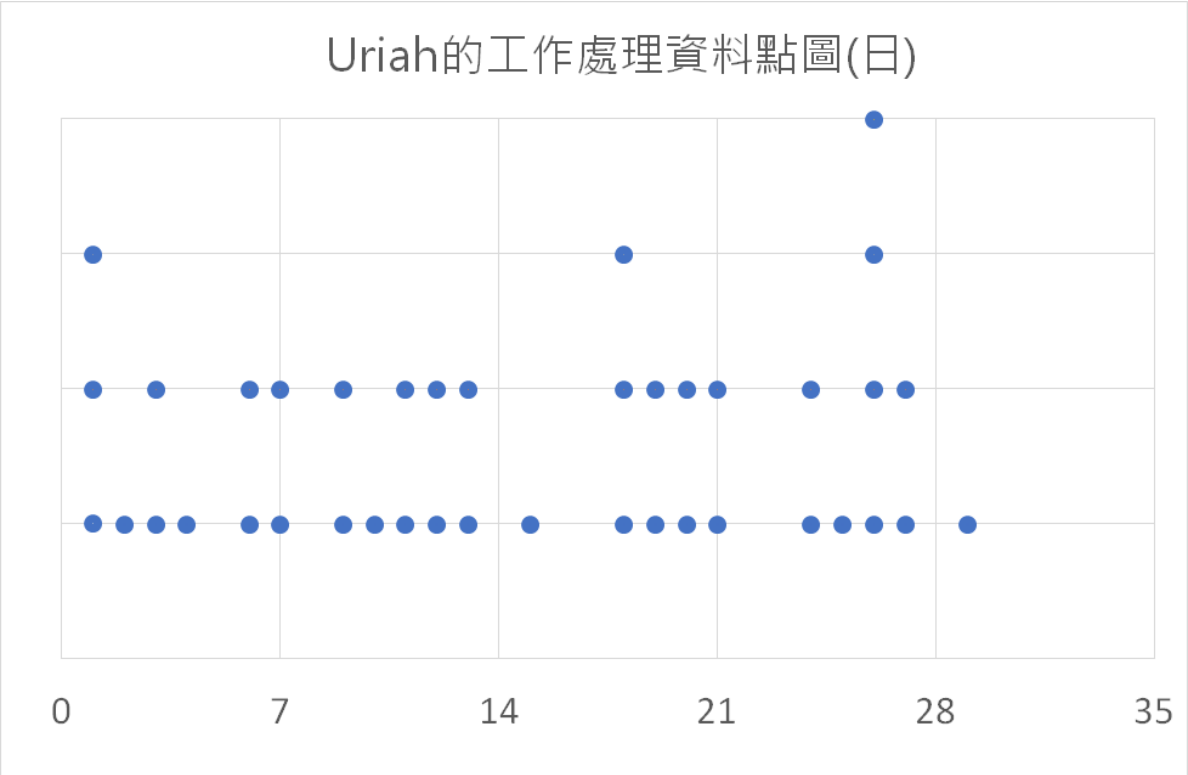
1. Uriah最常在7天內把功課解決，其次是13天內。
2. 五個分組的組中點分別是: 4、10.5、16.5、22.5、32

我們可以把表透過上述的公式轉成相對次數分配與百分比次數分配表，如下。

處理時間(日)	次數	相對次數分配	百分比次數分配
01-07	11	0.275	27.5%
08-13	9	0.225	22.5%
14-19	6	0.15	25%
20-25	7	0.175	22.5%
26-31	7	0.175	2.5%

Introduce - 以點圖來做資料圖形彙總

我們可以利用點圖來進行資料上的匯總，以上述Example，可以把Uriah的工作處理時間表轉成以下的點圖。



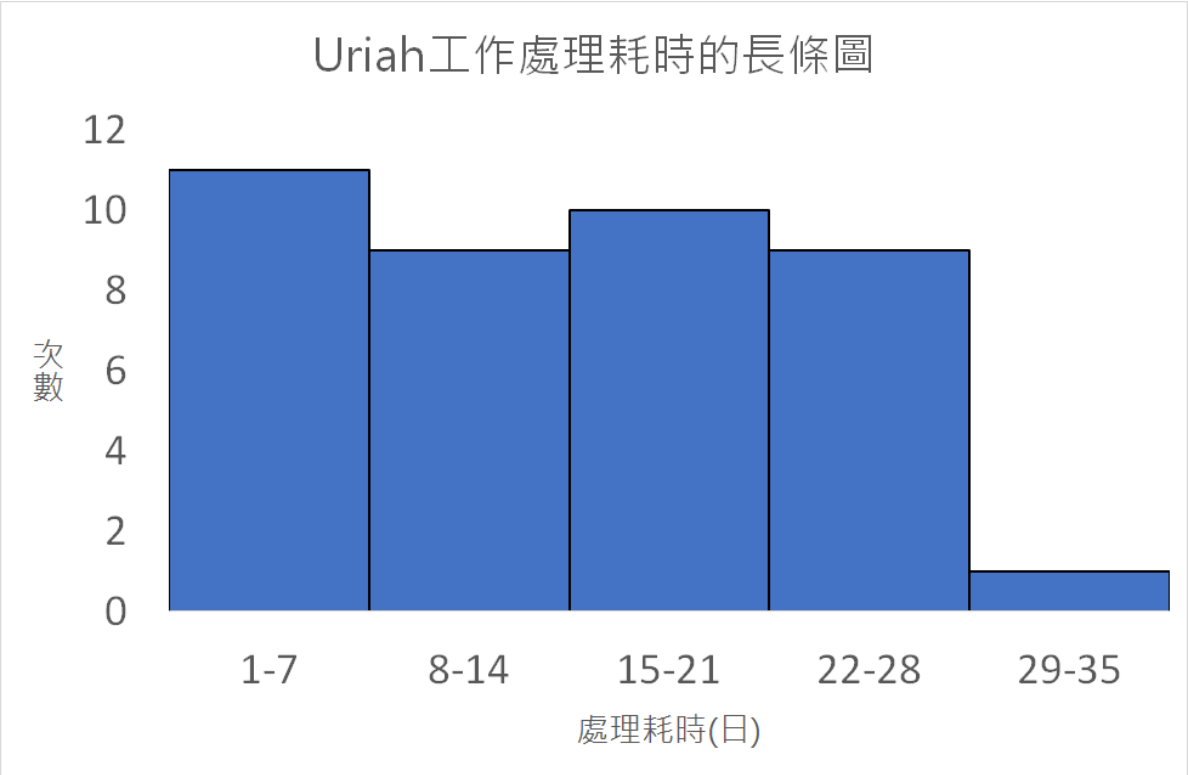
Introduce - 以直方圖來做資料圖形彙總

直方圖可以幫助我們更加瞭解資料，用來呈現以次數分配、相對次數分配或百分比次數分配彙整的資料。

直方圖最重要的用途之一，是讓人瞭解資料分布的形狀或形式。

如何偏態	定義與舉例
適度左偏	左端延伸的比較遠，例如統計比較簡單的考試的成績，大部分的成績都會高於70%，少數成績會極低。
適度右偏	右端延伸的比較遠，例如統計台北市的房價，少部分的豪宅會讓右端延伸。
對稱	左端是右端的鏡射，實務上不會完全對稱，例如統測成績，考極高與考極低都較少，較多人會在中間值之上或之下
高度右偏	右端與左端的落差極大，例如統計北科大學生的年齡，大多都會集中在18-24之間，少部分會集中在24之後。

例如下圖較接近高度右偏。



Introduce - 莖葉圖

莖葉圖可以用來同時呈現資料的順序與分配形狀，見以下範例。

Example - 微積分小考的測驗成績

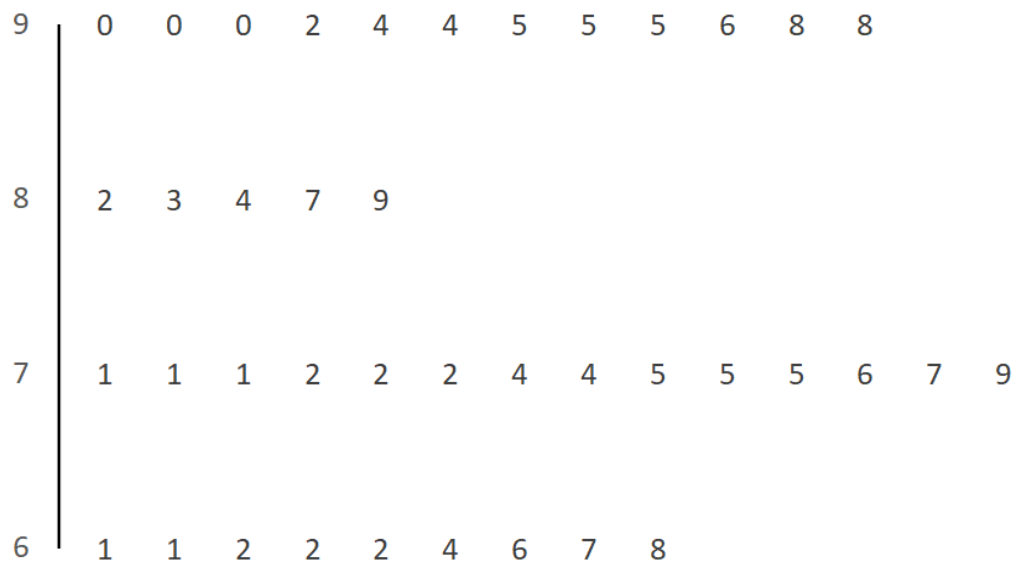
Uriah喜歡微積分，而他在某一年的時候擔任了微積分助教，並且負責幫微積分老師出小考與批改、計算成績。

而這一班共有50人的微積分，Uriah經過許多時間的批改之後，得到以下的成績表。

75	74	98	90	77	61	72	96	62	90
83	68	62	94	82	89	92	71	74	61
62	95	94	75	66	72	87	75	84	67
71	71	64	76	98	95	90	72	79	95

我們可以把它轉成莖葉圖，得到以下的圖形，並且可以觀察到70~79的次數較多，其次是90~100。

其中在8那一行的分數代表82, 83, 84, 87, 89。



Example - 大學生一個月的花費

Uriah統計了30個大學生一個月的花費，得到的結果如下。

```
11533 13553 13906 18376 18449 17422 10318 12395 18077  6674
15358 14361 11276 15471 14386 16261 12279 17706  6897 14945
11592 18377 15399  8992  8186 12753 14336 10553 11919  9015
```

如果這時候用前面百千萬來分組，就會顯得組別非常的雜亂，例如089只有一筆，而135也只有一筆。

所以我們考慮用千與萬來分組，但這時的莖葉圖也會顯得資料非常的雜亂。

我們其實可以用估算的方式，例如11533我們就把他歸類在11000裡面，而11276我們就把他歸類在11000裡面。

因此我們的葉單位就存在了，也就是每個葉都必須要乘上100，得到以下的莖葉圖

可以知道大學生大部分一個月的花費都集中在11000以上。

Section 2.3 運用表格彙總雙變數資料

Introduce - 交叉表格

交叉表格是雙變數資料的表格化彙總，較常見的交叉表格，變數之一是定量資料，另一變數則是類別資料。

Example - Christina的拉麵店評價

Christina是個喜歡吃拉麵的人，也因為很富有的關係，拉麵店幾乎是每天換一間，幾乎是嚐遍了馨築市的所有拉麵店。

Christina會把他吃過的所有拉麵店貼上評價與拉麵的價格來做對比，列成了以下的一張表格，包含了她吃過的100間拉麵店。

> Christina的拉麵店評價資料.csv

交叉表格

我們將這份資料做成交叉表格，得到以下的表格。

評價	100$ - 199$	200$ - 299$	300-399	400-500	總和
不錯	6	4	2	3	15
好吃	24	13	26	15	78
絕讚	2	2	0	3	7
總和	32	19	28	21	100

依照評價做出表格

我們可以根據以評價的次數做出相對次數與百分比次數表格，如下。

評價	相對次數	百分比次數
不錯	0.15	15%
好吃	0.78	78%
絕讚	0.07	7%
總和	1.00	100%

可以發現，Christina大部分吃的餐廳都是被Christina評為好吃(78%)，而極少部分被評為絕讚(7%)。

依照價格做出表格

我們可以根據以餐點價格的次數做出相對次數與百分比次數表格，如下。

餐點價格	相對次數	百分比次數
100-199	0.32	32%
200-299	0.19	19%
300-399	0.28	28%
400-500	0.21	21%
總和	1.00	100%

可以發現，Christina大部分吃的拉麵店價格都落在100-199之間，其次是300-399。

替換交叉表格的值(評價)

我們可以把交叉表格替換成百分比次數，以利於我們觀察。

根據評價，我們將每一欄除以該評價總數，來得到該價格在評價上的百分比。

評價	100$ - 199$	200$ - 299$	300-399	400-500	總和
不錯	40%	26.7%	13.3%	20%	100%
好吃	30.8%	16.7%	33.3%	19.2%	100%
絕讚	28.6%	28.6%	0%	42.9%	100%

因此可以發現，在評價不錯的價位上大多發生在約100~199元的拉麵，而評價絕讚的價位上大多發生在400~500元的拉麵。

Introduce - 辛普森詭論

兩個或更多的交叉表格常會被整合成一個彙整的交叉表格，以顯示雙變數間的關聯。

在這些情況下***由彙整兩個或更多個別的交叉表格而得到的結論，可能與整合為單一交叉表格得到的結論恰恰相反***。

此種現象稱為辛普森詭論，見以下範例。

Example - 天下第一補習班老師

Uriah和Diana為兩個補習班老師，而這個補習班近期的員工小圈圈舉辦了一個活動，透過考卷解題正確率來選出天下第一補習班老師。

這個補習班有兩門學科，分別是國文與數學，Uriah和Diana解題了一定數量的題目，將題目解題的結果分為「答對」與「答錯」。

因此我們得到了以下的兩個表格。

Uriah的解題結果

解題結果	國文	數學
答對	97(97%)	186(93%)
答錯	3(3%)	14(7%)
總和	100(100%)	200(100%)

Diana的解題結果

解題結果	國文	數學
答對	94(94%)	194(97%)
答錯	6(6%)	6(6%)
總和	100(100%)	200(100%)

表格合併

我們可以根據Uriah的答對、答錯率，與Diana的答對、答錯率來獲得以下合併出來的雙變數表格。

解題結果	Uriah	Diana
答對	283(94.3%)	288(96%)
答錯	17(5.7%)	12(4%)
總和	100(100%)	200(100%)

所以Diana勝出了嗎？

迷思與謬誤

國文與數學是完全性質不同的學科，所以他們的題數相加並沒有實質上的意義。

所以我們***並沒有辦法將國文的答題正確數量與數學的答題正確數量進行合併***，這樣合併並沒有實質上的意義。

若後來的表格合併出的結果是Diana勝出，僅憑這個表格無法證明Diana比較會解題。

因為在國文的部分，Uriah是勝出於Diana的，也因此Diana的國文解題能力明顯輸給了Uriah。

由以上可以得知，初始的表格顯然比整合過的表格更具有解題能力的參考意義。

因此，資料經過整合的交叉表格，應該要檢查是否有可能會出現影響結果的關鍵變數，以致於產生辛普森詭論。

Section 2.4 運用圖形彙總雙變數資料

Introduce - 散布圖與趨勢線

散布圖示表示兩定量變數間關係的圖形，而趨勢線則是近似關係的直線，可以用來判別資料是正相關、負相關或不相關。

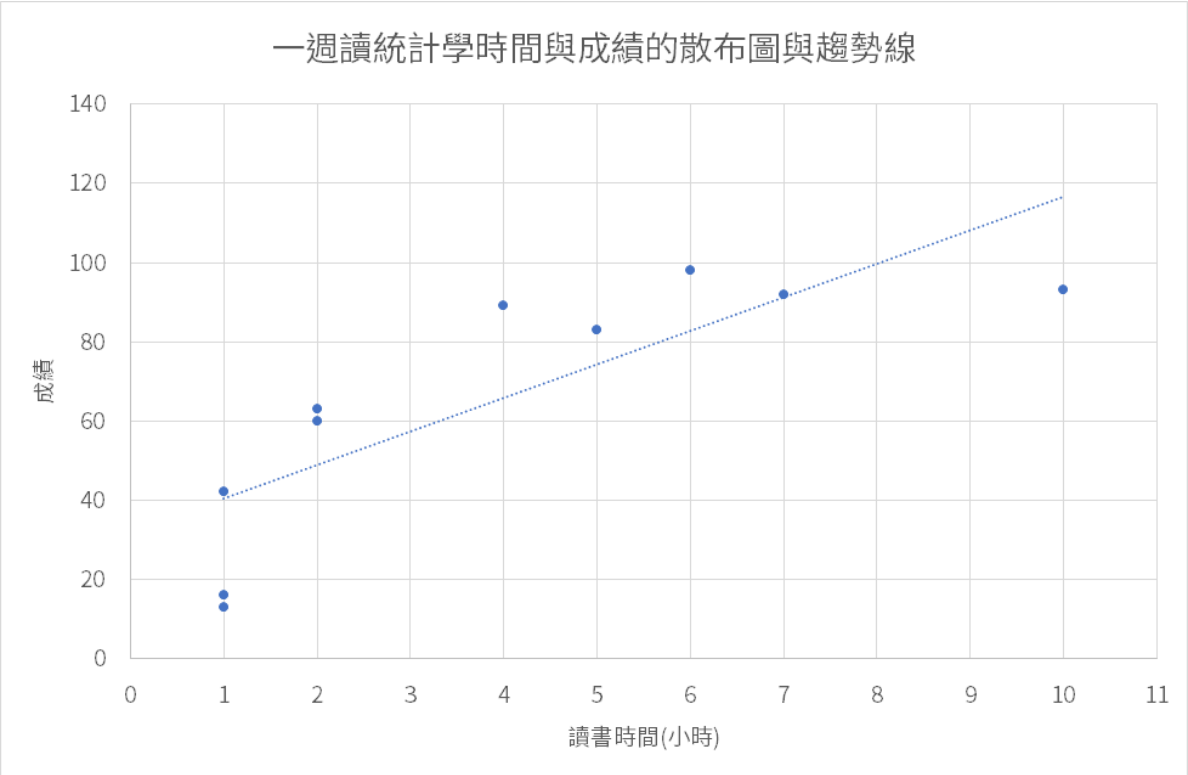
Example - 讀書時間與成績(散布圖)

就讀國立北海道科技大學的Uriah正在對修「統計學」的學生讀書時間與成績做研究。

這份研究資料有三個元素：學生編號、每週讀書時間(小時)、成績。

	每週讀書時間(小時)	成績
1	2	60
2	4	89
3	2	63
4	7	92
5	5	83
6	1	42
7	1	13
8	6	98
9	1	16
10	10	93

可以將上表轉成以下的散布圖與趨勢線



可以發現這個趨勢線是正相關型態，因此x與y的關係成正比，從這裡可以看出讀統計學的時間越多，分數會越高。

Introduce - 群組長條圖與堆疊長條圖

群組長條圖是在同張圖形上描繪多個長條圖。

堆疊長條圖的每個長條是標示為不同顏色的矩形堆疊而成，如同圓形圖一樣可以表示每組的相對次數。

Example - 讀書時間與成績(群組長條圖與堆疊長條圖)

就讀國立北海道科技大學的Uriah正在對修「統計學」的學生讀書時間與成績做研究。

這份研究資料有三個元素：

學生編號、每週讀書時間區間(小於1小時、1~3小時、4~6小時、6~9小時、9小時以上)、成績區間(不及格、低於80分、高於80分)。

樣本交叉表格

Uriah統計了100個學生的樣本，並且做成了以下的交叉表格。

	一小時以下	1到3小時	3到6小時	6到9小時	9小時以上	總計
不及格	14	10	0	0	0	24
60~80分	7	12	16	9	0	44
高於80分	0	1	2	11	18	32
總計	21	23	18	20	18	100

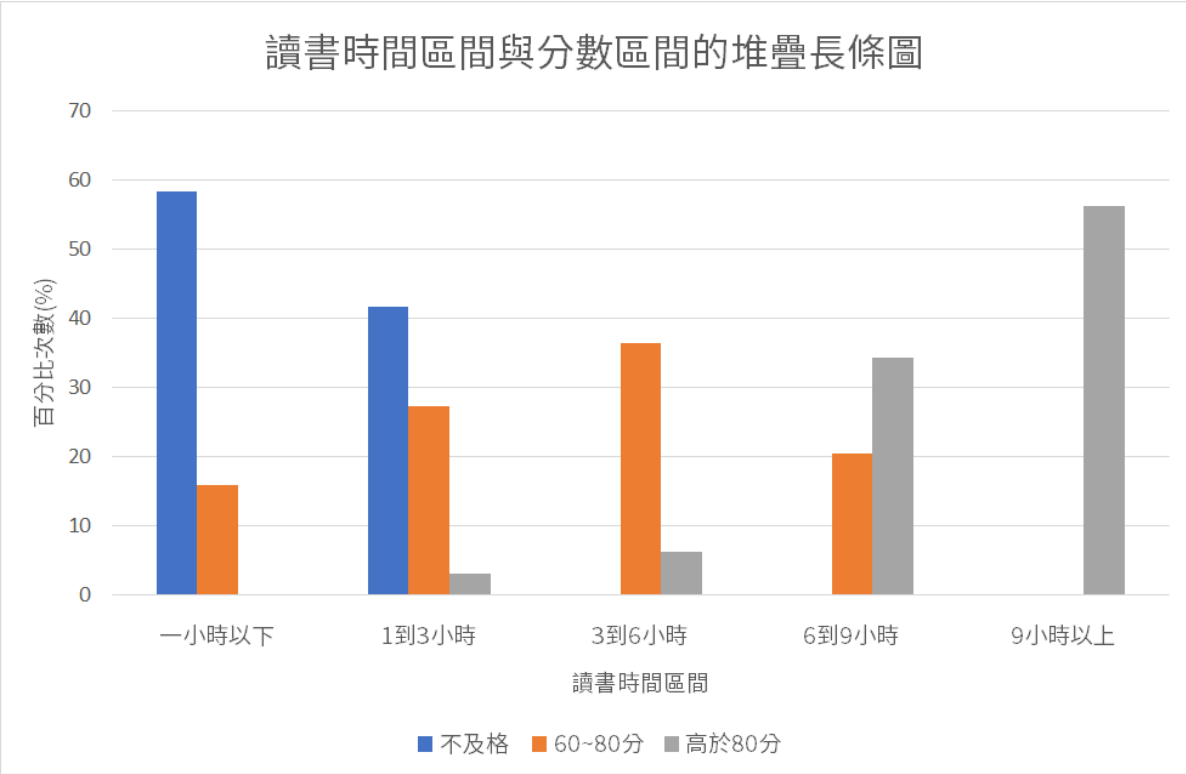
樣本百分比次數表格

再將其轉換成百分比次數表格。

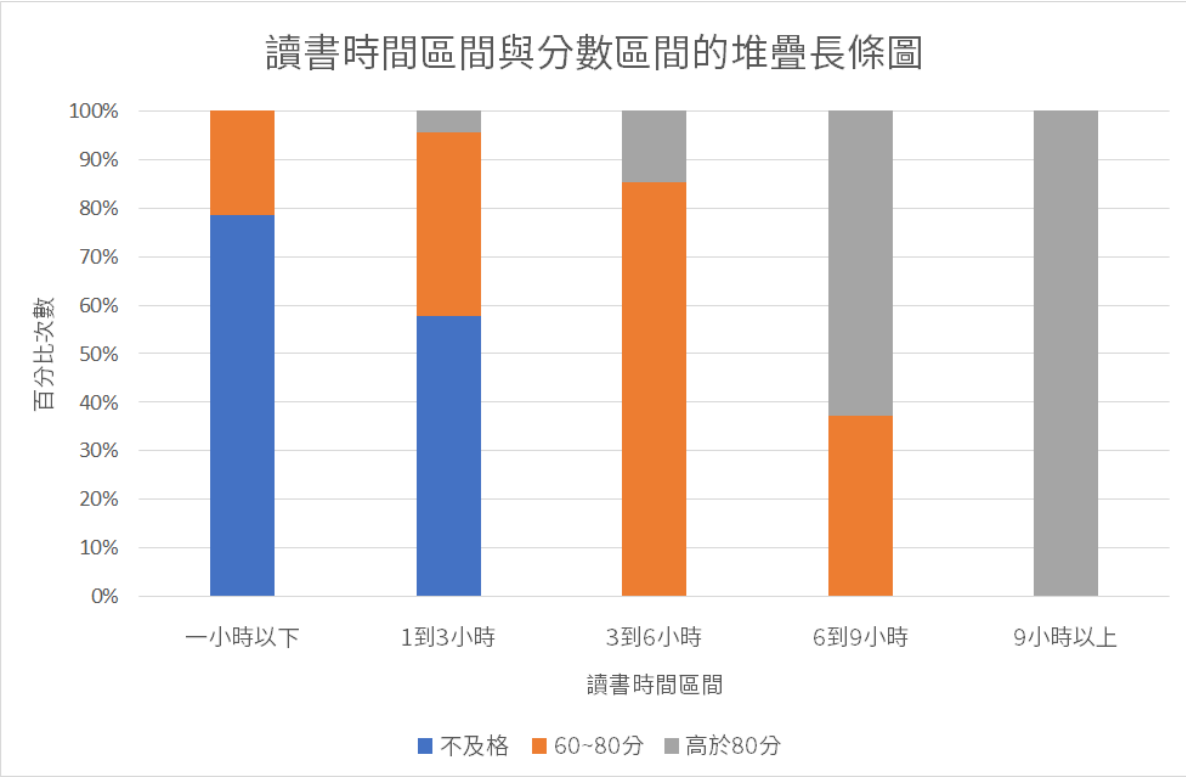
	一小時以下	1到3小時	3到6小時	6到9小時	9小時以上	總計
不及格	58.33%	41.67%	0.00%	0.00%	0.00%	100%
60~80分	15.91%	27.27%	36.36%	20.45%	0.00%	100%
高於80分	0.00%	3.13%	6.25%	34.38%	56.25%	100%

群組長條圖的描述統計

我們可以透過群組長條圖顯示資料，如下。



堆疊長條圖的描述統計



可以發現讀書時間越久，高於80分的人會越多，因此可以看出讀書時間越長，不及格的人數會越少，80分以上的人會越多。

Section 2.5 資料可視化：建立有效圖形表示的最佳作法

Introduce - 建立有效圖形表示

建立圖形的五大注意事項：

1. 清晰又簡潔的標題
2. 圖形盡可能簡單，能用二維就不要三維
3. 每個坐標軸要有清楚標示且要有刻度
4. 若要用顏色區別，確定顏色明顯不同
5. 如果使用多個顏色及線條樣式，使用圖例來定義顏色與線條的用法，放在圖形的旁邊。

Introduce - 選擇圖形表示的類別

圖形類別	用在資料分配上的圖形
長條圖	類別資料的次數分配與相對次數分配
圓形圖	類別資料的相對次數與百分比次數分配，偏好使用長條圖
點圖	在整個資料範圍中的分布
直方圖	資料在分組區間中的次數分配
莖葉圖	定量資料的順序及分布形狀

圖形類別	用於進行比較上的圖形
群組長條圖	比較兩個定量變數
堆疊長條圖	比較兩個定量變數的相對次數或百分比次數

圖形類別	用於表示關係上的圖形
散布圖	表示兩個定量變數的關係
趨勢線	近似散布圖中的關係

Introduce - 資料儀表板

資料儀表板用來呈現各式各樣的資訊，以易於判讀、瞭解及解釋的方法予以整編和呈現的視覺化表現。

資料儀表板應該用來呈現對於使用者重要的「關鍵績效指標」(也就是說，對於使用者重要的各式指標)，而非轟炸使用者。

Chapter 3. 敘述統計：數值方法

Section 3.1 位置量數

Introduce - 樣本平均數

一個有n個觀察值的樣本，樣本平均數的公式如下。

$$\overline{x} = \frac{\sum x_i}{n}$$

其中x_i代表該樣本的第i個觀察值。

Example - Uriah的睡眠時間平均

Uriah是資工系的學生，下表蒐集了一個禮拜Uriah的睡眠時間資料

	一	二	三	四	五	六	日
睡眠時間(小時)	8	8	7	7	7	5	7

可以知道七天睡眠時間的平均值$\overline{x} = \frac{8+8+7+7+7+5+7}{7} = 7$小時

Introduce - 母體平均數

與樣本平均數的公式相同，以μ來表示母體平均數，並以N來表示母體包含的所有元素個數。

$$\overline{\mu} = \frac{\sum x_i}{n}$$

Introduce - 加權平均數

為了反映個別觀察值的重要性，我們可以將觀察值賦予加權，算出來的平均數就是加權平均數，以以下的式子表示。

$$\overline{x} = \frac{\sum w_i x_i}{\sum w_i}$$

其中w_i為第i個觀察值的權重。

Example - Uriah的學期平均成績

Uriah是資工系的學生，下表蒐集了Uriah某個學期的學期成績資料。

我們可以用學分當作科目的權重，用來計算Uriah的學期成績平均。

科目	學分	分數
英文	2	94
國防	0	82
勞作教育	0	86
國文	2	80
微積分	3	74
離散數學	3	93
程式	3	100
數邏實習	1	82
計算機組織	3	98
體育	0	93
藝術概論	2	90
婚姻與家庭	2	84

可以根據加權平均的計算方法，得到

$$\frac{2 \cdot 94 + 0 \cdot 82 + 0 \cdot 86 + 2 \cdot 80 + 3 \cdot 74 + 3 \cdot 93 + 3 \cdot 100 + 1 \cdot 82 + 3 \cdot 98 + 0 \cdot 93 + 2 \cdot 90 + 2 \cdot 84}{21} = 89.2$$

因此可以算出Uriah的學期平均：89.2分。

Introduce - 中位數

中位數即為將資料「非嚴格遞增排序」之後，若資料項數為奇數項時，中位數為此資料之中間值，否則為中間兩個數值的平均數。

可以把計算中位數的部份視為不停刪去極大極小值後，剩餘的數值若為1項或2項時，計算這幾項的平均數。

若資料存在極端值時，中位數可以較好的反應中央位置的量數。

Example - Uriah的程式小考

Uriah是某堂程式課的助教，這次的課程要小考，Uriah統計了班上9個人的成績，排序之後希望可以找到中位數，以下是資料：

$$20, 20, 20, 40, 40, 60, 60, 60, 80$$

如果我們不停刪去極大極小值後，可以發現只剩下一個數值40，此數值即為中位數。

若有一位同學補考，把這個同學的分數加入成績，可以得到以下的資料：

$$20, 20, 20, 40, 40, 60, 60, 60, 80, 150$$

如果我們不停刪去極大極小值後，可以發現只剩下兩個數值$40, 60$，將兩個數值取平均之後可以得到50，此數值即為中位數。

Introduce - 幾何平均數

幾何平均數與樣本/母體平均數不同，利用值的「乘積」來指示一組數字的集中趨勢或典型值，可以用以下的式子定義：

$$\overline{x_g} = \sqrt[n]{(x_1)(x_2)(x_3)\dots(x_n)} = [x_1 x_2 x_3 \dots x_n]^{1/n}$$

幾何平均數的一些特點是，幾何平均數受到極值的影響會更小，且幾何平均數適用於計算在等比資料上，用以下的式子來解釋一些情況下，樣本平均數比幾何平均數還要來得不適用。

Example - Uriah的果園

為了推廣Uriah的鮮奶茶鋪的新品項(當然不是芋頭)，Uriah嘗試將果汁與鮮奶融合，合出一種神奇的風味。

因此Uriah剛好注意到了橘子跟鮮奶似乎很搭，所以打算開始種植橘子然後做成果汁。

Uriah買了一顆橘子樹，原先他在買了這顆橘子樹後有100顆

Uriah紀錄了每年橘子樹的產量，得到了以下的這張表。

	種植後第一年	**種植後第二年**	**種植後第三年**
橘子樹產量(顆)	180	210	300
成長比例	80%	16.6666%	42.8571%

樣本平均數計算平均成長率

我們可以用樣本平均數求得平均，可以得到每年平均成長率為$\frac{80+16.6666+42.8571}{2}=46.5079\%$。

如果我們回推回去，從第零年去預估第一年、第二年、第三年的成長率。

	第一年	第二年	第三年
橘子樹產量(顆)	146	213	312

會發現有312顆橘子而不是300顆，所以隨之超越幅度會越來越大。

幾何平均數計算平均成長率

如果我們利用幾何平均數來計算平均成長率，會得到每年平均成長率為$\sqrt[3]{1.80\times 1.16666\times 1.428571}\approx 1.44224$

如果我們回推回去，從第零年去預估第一年、第二年、第三年的成長率。

	第一年	第二年	第三年
橘子樹產量(顆)	144	208	300

這樣的預測就會較為準確。

Introduce - 眾數

眾數是資料集中出現次數最多的數值就是眾數。

Introduce - 百分位數

百分位數可以瞭解資料在最小值與最大值間的分布狀況，以pth百分位數可以把資料分成兩個部份。

大約pth百分比的觀察值會小於pth百分位數，而大約有(100-p)百分比的觀察值會大於pth百分位數。

計算時需要先「非嚴格遞增排序」，計算公式如下：

$$L_p=\frac{p}{100}(n+1)$$

Example - 模擬考成績

Uriah待的補習班舉辦了統測模擬考，針對100位學生拿到了以下的模擬考數據並進行排序。

302	306	311	314	317	317	318	321	323	327
329	331	332	341	342	349	350	355	361	364
370	372	375	377	377	382	382	388	389	389
391	412	412	412	433	442	444	454	455	456
464	464	465	468	470	470	475	477	477	479
485	498	498	501	501	503	511	529	530	534
534	534	543	549	549	554	555	559	562	562
562	566	582	599	600	605	606	606	608	622
624	625	627	627	633	640	649	651	657	659
660	669	670	672	673	673	677	683	694	695

取得80th百分位數

若我們要取得$80th$百分位數的位置，可以得到$L_{80} = (\frac{80}{100})(100+1) = 80.8$

因此可以知道$80th$百分位數落在第80個人(622)到第81個人(624)之間，且離位置80的距離是位置81與位置80之差的80%

故我們可以得到

$$80\text{th percentile} = 622 + 0.8(624 - 622) = 623.6$$

取得50th百分位數

若我們要取得$50th$百分位數的位置，可以得到$L_{50} = (\frac{50}{100})(100+1) = 50.5$

因此可以知道$50th$百分位數落在第50個人(479)到第51個人(485)之間，且離位置50的距離是位置51與位置50之差的50%

故我們可以得到

$$50\text{th percentile} = 479 + 0.5(485 - 479) = 482$$

我們考慮取得這個資料的中位數，即為$\frac{479+485}{2} = 482$，因此50th percentile也是中位數。

Introduce - 四分位數

四分位數將整筆資料分成四個部份，每個部份大概含有25%的資料個數，定義如下：

1. Q_1為第一四分位數或25th percentile
2. Q_2為第二四分位數或50th percentile，也就是中位數
3. Q_3為第三四分位數或75th percentile

Example - 模擬考成績

Uriah待的補習班舉辦了統測模擬考，針對100位學生拿到了以下的模擬考數據並進行排序。

302	306	311	314	317	317	318	321	323	327
329	331	332	341	342	349	350	355	361	364
370	372	375	377	377	382	382	388	389	389
391	412	412	412	433	442	444	454	455	456
464	464	465	468	470	470	475	477	477	479
485	498	498	501	501	503	511	529	530	534
534	534	543	549	549	554	555	559	562	562
562	566	582	599	600	605	606	606	608	622
624	625	627	627	633	640	649	651	657	659
660	669	670	672	673	673	677	683	694	695

我們可以對這些資料利用四分位數分割成四個部份，透過先前公式，我們可以找到

$$L_{25} = (\frac{25}{100})(100+1) = 25.25$$

因此$25\text{th percentile} = 377 + 0.25(382-377) = 378.25 = Q_1$

$$L_{50} = (\frac{50}{100})(100+1) = 50.5$$

因此$50\text{th percentile} = 479 + 0.5(485-479) = 482 = Q_2$

$$L_{75} = (\frac{75}{100})(100+1) = 75.75$$

因此$75\text{th percentile} = 600 + 0.75(605-600) = 603.75 = Q_3$

我們根據這三個數值，將資料做分割。

302	306	311	314	317	317	318	321	323	327
329	331	332	341	342	349	350	355	361	364
370	372	375	377	377	382	382	388	389	389
391	412	412	412	433	442	444	454	455	456
464	464	465	468	470	470	475	477	477	479
485	498	498	501	501	503	511	529	530	534
534	534	543	549	549	554	555	559	562	562
562	566	582	599	600	605	606	606	608	622
624	625	627	627	633	640	649	651	657	659
660	669	670	672	673	673	677	683	694	695

會發現資料很均等的被分割成了四個部份。

Section 3.2 離散量數

Introduce - 全距(Range)

全距的定義為一筆資料的最大值與最小值之差，寫成以下的形式：

$$\text{Range} = \text{Largest value - Smallest value}$$

一般來說不會單看全距，因為全距容易受極端值影響。

Introduce - 四分位距(IQR)

四分位距較能克服極端值的影響，定義如下

$$\mathrm{IQR} = Q_3 - Q_1$$

也就是中間資料50%的全距。

Introduce - 變異數(variance)

變異數是利用**全部資料**所得到的離散量數，較大的變異數時資料會擁有較大的離散程度。

變異數是根據每一個觀察值x_i與平均數μ之差求得，此差稱為離差。

計算變異數時，取離差的平方值，方法如下：

若資料集為母體時，這些離差平方的平均稱為母體變異數σ^2，定義如下：

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

若資料集為樣本時，這些離差平方的平均稱為樣本變異數s^2，定義如下：

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

此樣本變異數為母體變異數的「不偏估計量」，而使用n時為「偏估計量」。

課本並未提及這個部份，有興趣可以看以下的補充教材。

Introduce - 為什麼是n-1? (補充教材)

假設我們有一筆資料：1 2 3 4 5 6 7 8 9 10

我們可以知道這筆資料的母體平均數為6.5，大概屆於中間左右

我們今天用同樣的方式抽取樣本，假設我們可以抽{1, 5, 10}，依然會得到大概在中間的平均值

不過一些特例的情況下，我們可能會抽到{1, 2, 3}，這樣得到的平均值就會距離中間非常遠

因此，以n來當成變異數的分母，就是此資料的「偏估計」量

Introduce - 標準差(standard deviation)

標準差即為變異數的正平方根，標準差能將原始資料的單位一致，更好的去比較原始資料。

樣本標準差s與母體標準差σ的定義如下：

$$s = \sqrt{s^2}$$
$$\sigma = \sqrt{\sigma^2}$$

Introduce - 變異係數

我們可能會想要知道標準差相對於平均數的比例，這種量數稱為變異係數，通常表示為百分比的形式，定義如下。

$$\text{coefficient of variation} = (\frac{\text{standard deviation}}{\text{mean}} \times 100)\%$$

可以用來比較單位不同，或者單位相同但資料相差甚大的資料分散情形。

Example - Uriah與Roger的案子處理效率(變異)

Uriah與Roger是兩個接案工程師，由於上級想要評估兩個人的案子處理穩定性，因此記錄了兩個人抽樣了10件案子所需的耗時天數，如下

案子編號	1	2	3	4	5	6	7	8	9	10
Uriah的處理天數(日)	5	6	7	7	6	7	6	5	5	6
Roger的處理天數(日)	5	7	9	11	9	7	5	6	12	13

我們可以計算出Uriah的處理天數平均為6天，且Roger的處理天數平均為8.4天

我們可以計算這兩個人的樣本變異數，得到$s^2_{\text{Uriah}} = 0.66667(\text{天}^2)$，且$s^2_{\text{Roger}} = 8.26667(\text{天}^2)$

故Uriah的案子處理穩定性較高。

我們可以得到$s_{\text{Uriah}} = 0.8165(\text{天})$，且$s_{\text{Roger}} = 2.8752(\text{天})$

因此Uraih的變異係數為$(\frac{0.8165}{6} \times 100)\% = 13\%$，且Roger的變異係數為$(\frac{2.8752}{8.4} \times 100)\% = 34\%$

Introduce - 平均絕對誤差(MAE)

平均絕對誤差是所有離差的絕對值的平均。

平均絕對誤差由於離差被絕對值化，因此不會出現正負抵銷的問題

若一筆資料有n個元素，則MAE的定義如下：

$$\text{MAE} = \frac{\sum |x_i - \overline{x}|}{n}$$

Example - Uriah與Roger的案子處理效率(MAE)

Uriah與Roger是兩個接案工程師，由於上級想要評估兩個人的案子處理穩定性，因此記錄了兩個人抽樣了10件案子所需的耗時天數，如下

案子編號	1	2	3	4	5	6	7	8	9	10
Uriah的處理天數(日)	5	6	7	7	6	7	6	5	5	6
Roger的處理天數(日)	5	7	9	11	9	7	5	6	12	13

我們可以計算出Uriah的處理天數平均為6天，且Roger的處理天數平均為8.4天

我們可以計算這兩個人的MAE，得到$\mathrm{MAE}_{\mathrm{Uriah}} = 0.6$，且$\mathrm{MAE}_{\mathrm{Roger}} = 2.4$，因此Roger有較大的$\mathrm{MAE}$。

Section 3.3 分配形狀的量數、相對位置及離群值的偵測

Introduce - 偏度

衡量形狀的重要數值。

Introduce - 分配的形狀

當資料左偏時偏度為負值，資料右偏時偏度為正值，對稱時偏度為0

> 補充教材: 偏度公式
>
> $$\frac{n}{(n-1)(n-2)}\sum(\frac{x_i-\overline{x}}{s})^3$$

Introduce - z分數

z分數稱為標準化值，用來解釋觀察值在資料集中的相對位置。

因此，不同資料集的觀察值具有相同z分數者，可解釋兩觀察值距其資料之平均值有相同個標準差之遠，也就是有同樣的相對位置。

Example - 班級人數資料的z分數

有以下五個班級的人數，針對於這五個班級算出離差與z分數。

班級學生人數	離差	z分數
46	2	0.25
54	10	1.25
42	-2	-0.25
46	2	0.25
32	-12	-1.5

可以看出-1.5是離平均值最遠的資料值，比平均數少1.5個標準差。

Introduce - 柴比雪夫定理

根據柴比雪夫定理可知，在資料集內，至少有$(1-\frac{1}{z^2})$百分比的觀察值與平均數的差距必須在z個標準差之內，z為任何大於1的值。

柴比雪夫定理可以讓我們知道與平均數的差距在特定標準差之內的觀察值佔整個資料集的百分比。

帶入$z=2$得到至少有0.75或75%的觀察值，與平均數的差距在$z=2$個標準差內。

帶入$z=3$得到至少有0.89或89%的觀察值，與平均數的差距在$z=3$個標準差內。

帶入$z=4$得到至少有0.94或94%的觀察值，與平均數的差距在$z=4$個標準差內。

Example - 期末考

假設某科系的Python程式設計有100位同學修課，期中考成績的平均值μ是70，標準差σ是5。

我們想要知道有多少的學生的分數介於60到80之間，又有多少的學生的分數介於58到82之間。

介於60到80之間即為兩個標準差，透過柴比雪夫定理可以知道$1-\frac{1}{2^2}=75\%$

介於58到82之間可以知道$\frac{58-70}{5}=-2.4$，故小於平均數2.4個標準差，且$\frac{82-70}{5}=2.4$，故大於平均數2.4個標準差。

所以可以得到$1-\frac{1}{2.4^2}=0.826=82.6\%$，至少有$82.6\%$的學生的分數必須介於$58$到$82$之間。

Introduce - 經驗法則

針對於鐘形分配的資料集而言，大約有68%的觀察值與平均數的差距在一個標準差內，95%的觀察值與平均數的差距在兩個標準差內，

99.7%的觀察值與平均數的差距在三個標準差內。

> 簡單來說，是高職的常態分佈

Example - Uriah的茶鋪

Uriah有一個神奇的鮮奶茶封裝產業鏈，平均每一杯鮮奶茶是600公克，標準差是3公克。

我們可以利用經驗法則得到：

大概68%的鮮奶茶重量介於597到603公克之間。

大概95%的鮮奶茶重量介於594到606公克之間。

幾乎所有鮮奶茶介於591到609公克之間。

Introduce - 離群偵測

有時在一個資料集中會有極大與極小的數值，這些數值稱為離群。

我們可以利用z分數去偵測離群值，一般來說，約有99.7%的資料會落在標準差±3內，我們會希望資料的標準差與中心的距離不超過3。

另一種方式是使用第一分位、第三分位與四分位距來做偵測，能夠給定一個區間來要求值必須要在這個區間內，定義如下

$$\text{Lower Limit}=Q_1-1.5(\text{IQR})$$
$$\text{Upper Limit}=Q_3+1.5(\text{IQR})$$

Example - 北科生活費調查

Uriah對於北科學生的生活費非常感興趣，所以他對隨機10位北科學生的生活費進行訪查，得到以下的資料。

1	2	3	4	5	6	7	8	9	10
10000	10000	10000	10000	12000	12000	12000	12000	14000	40000

透過計算可以知道$Q_1 = 10000$，$Q_3 = 12500$，故$\text{IQR} = 12500 - 10000 = 2500$

因此$\text{Lower Limit} = 10000 - 1.5(2500) = 6250$，且$\text{Upper Limit} = 12500 + 1.5(2500) = 16250$

因此40000的資料顯然是離異的。

Section 3.4 五數彙總與箱型圖

Introduce - 五數彙總

五數彙總是指蒐集以下五個數來彙總資料。

1. 最小值
2. 第一四分位數
3. 中位數
4. 第三四分位數
5. 最大值

透過這個五數彙總，可以用來繪製箱型圖。

Introduce - 箱形圖

箱型圖是透過五數彙總所畫出的圖，以第一、第三四分位數為前後邊，以垂直線為中位數。

而箱型圖的邊界取決於資料的界線(limit)，用水平線來表示界線，而用一點來表示離群值。

例如以下的資料：

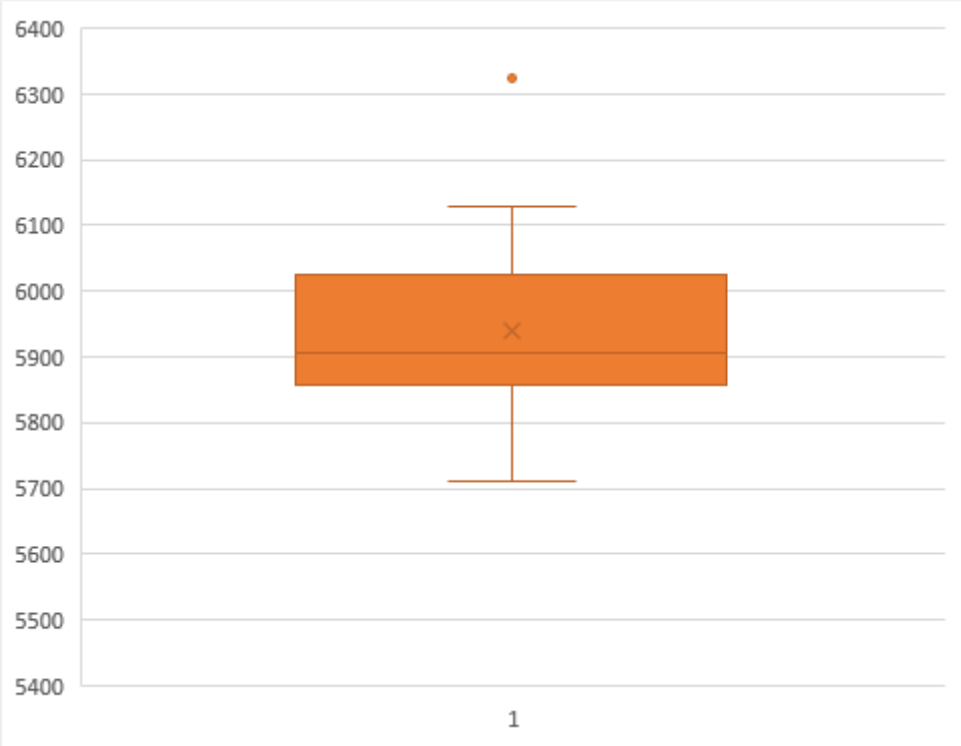
5710	5755	5850	5880	5880	5890	5920	5940	5950	6050	6130	6325

可以得到$Q_1 = 5857.5$，$Q_3 = 6025$，且$\text{IQR} = 6025 - 5857.5 = 167.5$。

因此上邊界為$5857.5 - 1.5(167.5) = 5606.25$，下邊界為$6025 + 1.5(167.5) = 6276.25$。

所以6325是離群值，故以橙色點點表示

移除離群值後，可以發現最大值為6130，且最小值為5710。



Section 3.5 兩變數的相關性量數

Introduce - 共變異數

我們用共變異數做為兩變數間線性相關的敘述量數。

對於樣本大小為n的配對觀察值如(x_1, y_1)、(x_2, y_2)等，樣本共變異數定義如下：

$$s_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

而對於母體大小N的共變異數定義如下：

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Introduce - 共變異數的解釋

我們可以用$x = \overline{x}$軸與$y = \overline{y}$軸分割成四個平面，接著我們考慮

- 如果點落在第一象限與第三象限，則$(x_i - \overline{x})(y_i - \overline{y}) > 0$

  可以推導出x與y有正的線性相關，也就是當x值增加，y值也增加。

- 如果點落在第二象限與第四象限，則$(x_i - \overline{x})(y_i - \overline{y}) < 0$

  可以推導出x與y有負的線性相關，也就是當x值增加，y值則減少。

- 如果點落的非常的散，則s_{xy}的值會趨近於0，表示x與y沒有線性關係

若共變異數取絕對值後越大，則正的線性相關與負的線性相關越強烈，反之則越微弱。

Introduce - 相關係數

考慮：如果我們要比較一個人的身高與體重的關係

如果身高的單位用公分與用公尺，則用公分的$(x - \overline{x_i})$可能會比公尺還大，造成共變異數變的很大。

但是用公尺與用公分是沒關係的。

因此有了不受衡量單位的衡量量數，稱作相關係數。

樣本相關係數的定義如下：

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

其中s_{xy}是樣本的共變異數，且s_x是x的樣本標準差，s_y是y的樣本標準差。

母體相關係數的定義如下：

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

其中σ_{xy}是樣本的共變異數，且σ_x是x的樣本標準差，σ_y是y的樣本標準差。

Introduce - 相關係數的解釋

考慮一個強烈正相關的散布圖，我們可以知道所有點都會聚集在$y = x$上，此時可以發現$r_{xy} = 1$

而負相關的散布圖，我們可以知道所有點都會聚集在$y = -x$上，此時可以發現$r_{xy} = -1$

因此我們可以說$r_{xy} = 1$時為完全正相關，且$r_{xy} = -1$時為完全負相關。

如果是毫無相關的話，則$\sigma_{xy} \approx 0$，因此相關係數也會趨近於0，因此$r_{xy} \approx 0$時代表線性關係微弱