# 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。

## 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

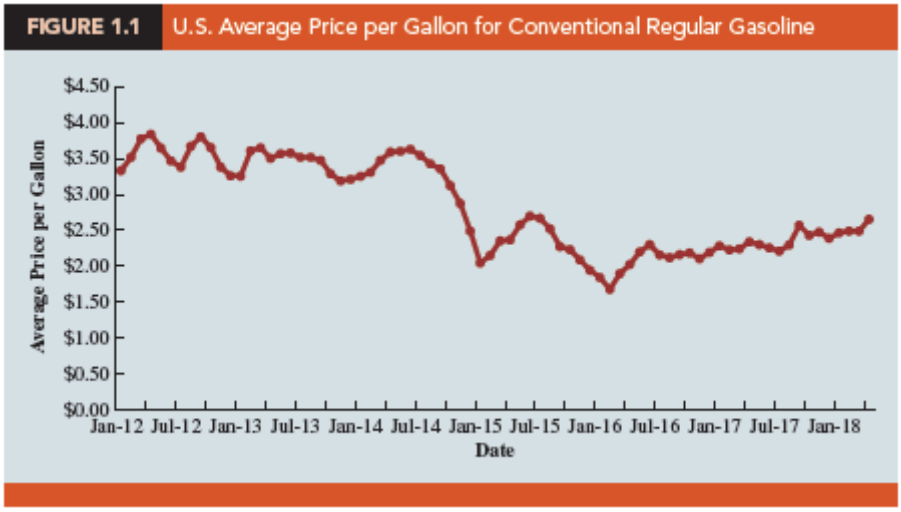
可以把資料分成類別資料與定量資料，類別資料使用名目尺度或順序尺度，而定量資料使用區間尺度與比例尺度。

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。

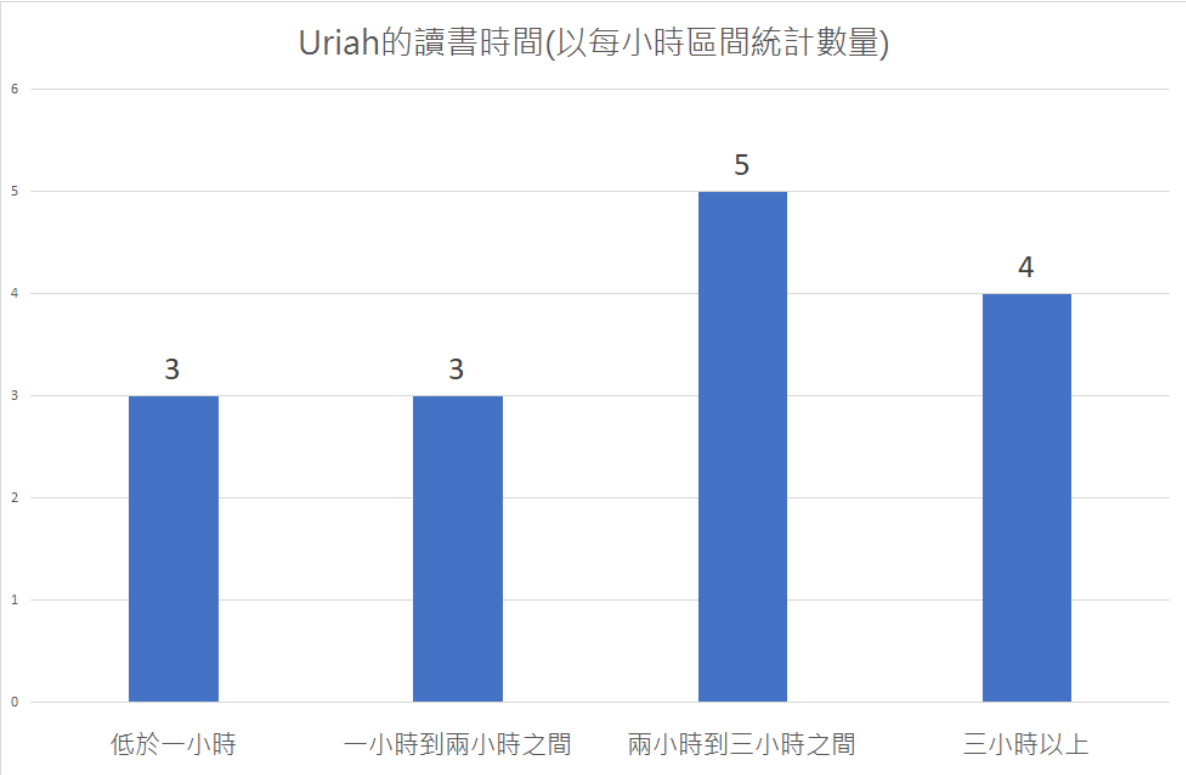
大部分刊登在媒體上，公司年報或其他出版品的統計資訊，是以讓人容易瞭解的資料形式來彙整公布，稱為敘述統計。

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。

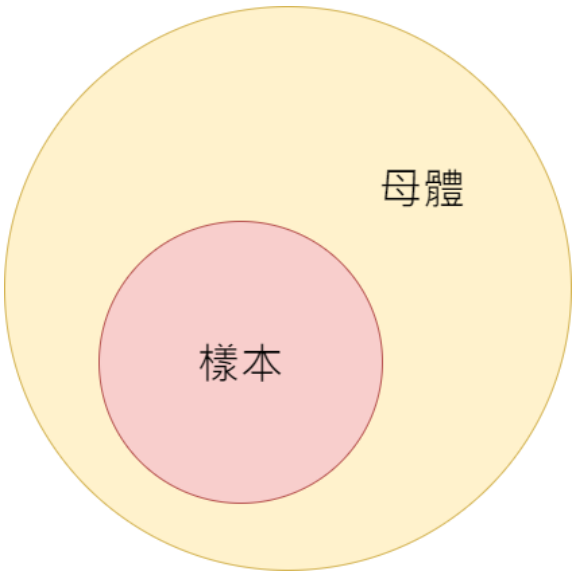
透過計算，可以得出Uriah每天讀書約2個小時24分鐘，平均數可以顯示資料集的中央趨勢或資料集的中央位置。

## Section 1.5 統計推論

### Introduce - 母體與樣本

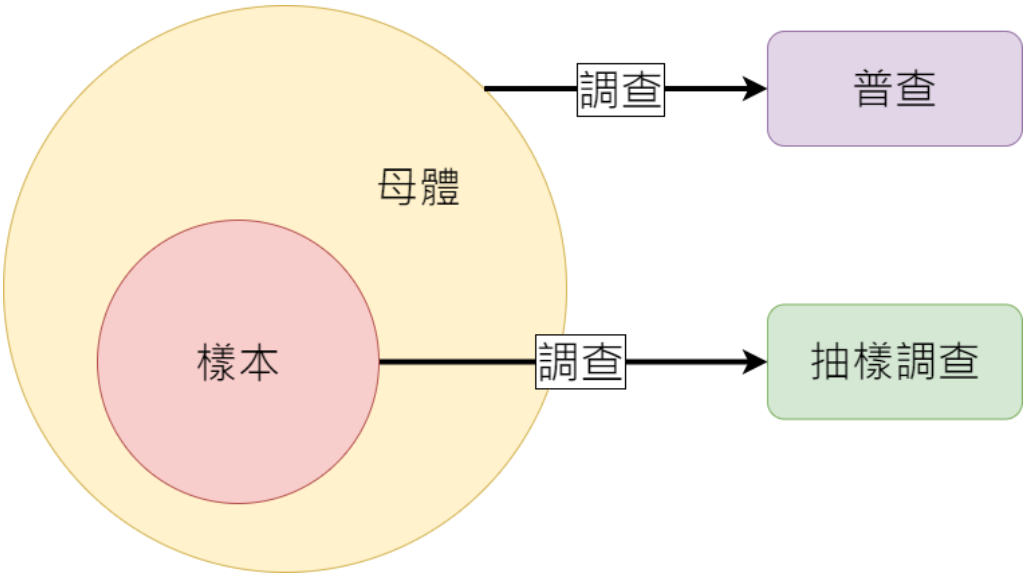
許多情況必須針對大量的元素來蒐集資料，但因為時間與成本的關係，僅僅只能蒐集到資料的一小部分。

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。



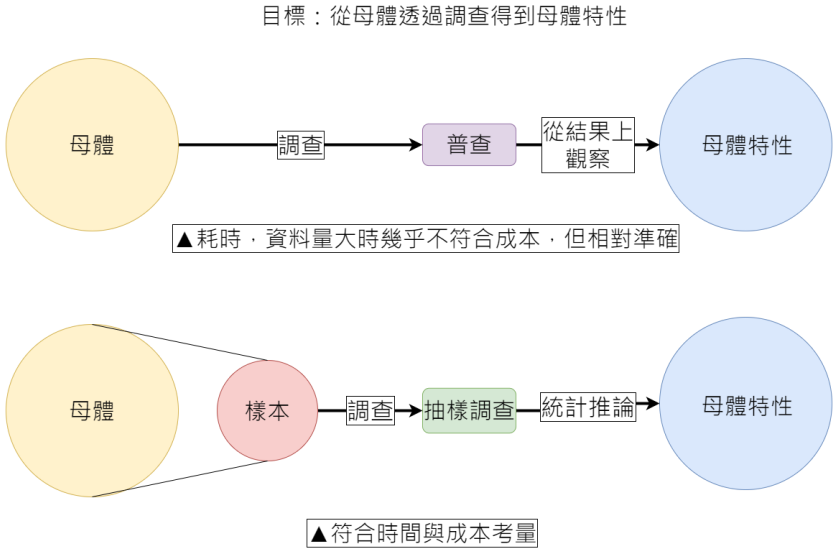

## Introduce - 普查與抽樣調查

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 電腦與統計方法

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。

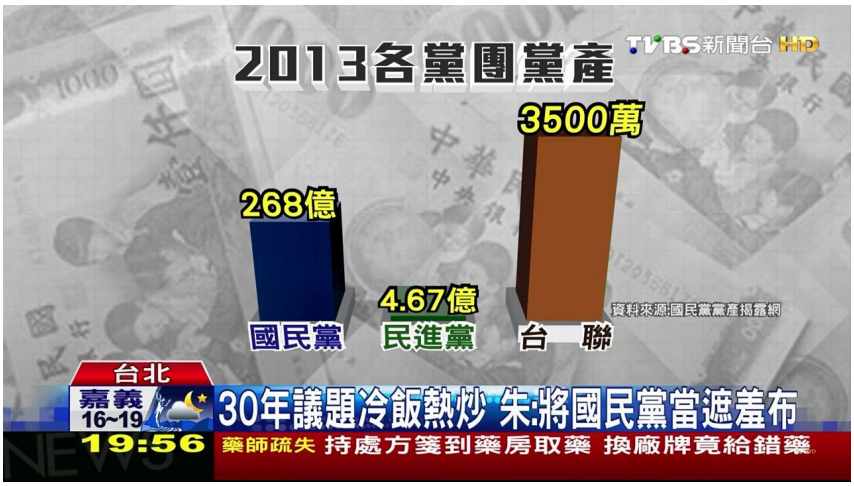
在統計學不適當的抽樣、不適當的資料分析、誤導的圖形、不適當的彙整統計資料及統計結果的偏差解釋，都是一種不道德行為。

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)