給工程師的統計學與資料分析 123

第零單元：資料視覺化與摘要

孔令傑

國立臺灣大學資訊管理學系

2017 年 9 月 2 日

課程大綱

- 資料視覺化
- 資料摘要

公共腳踏車租借系統

- 在 2011 與 2012，我們記錄華盛頓特區公共腳踏車租借系統的每日租借次數。
  - 985、801、1349、1562、1600、...，以及 2729。
  - 最小和最大的數字分別為 22 及 8714。
- 要怎麼對這 731 個數字有感覺呢？

日期	租借次數
2011/1/1	985
2011/1/2	801
2011/1/3	1349
2011/1/4	1562
2011/1/5	1600
⋮	
2012/12/29	1341
2012/12/30	1796
2012/12/31	2729

次數分佈

- 原始的 731 個數字形成的是一組未分組資料 (ungrouped data)。
- 首先我們將這些資料分組成一個次數分佈 (frequency distribution)。
  - 對於每一組，我們呈現它的「組距」和「發生次數」。
- 讓我們來建立一個直觀的次數分佈吧！

次數分佈

- 一種分組方式:

編號	分組	代表意義
1	$[0, 1000)$	$0 \leq x < 1000$
2	$[1000, 2000)$	$1000 \leq x < 2000$
3	$[2000, 3000)$	$2000 \leq x < 3000$
	$\vdots$	
8	$[7000, 8000)$	$7000 \leq x < 8000$
9	$[8000, 9000)$	$8000 \leq x < 9000$

- 有無限多種分組方式；通常各組組距會等長。
- 各分組之間應該要沒有空隙：$[0, 999]$、$[1000, 1999]$、... 是錯的。
- 各分組織間應該要不重疊：$[0, 1000]$、$[1000, 1999]$、... 是錯的。

次數分佈

- 接著我們把 731 個數字一一丟進各分組中，來得到如右的次數分佈。
- 這是一組分組資料 (grouped data)。
- 可以看出在大部分日子中，租借次數分佈在 3000 到 6000 之間。
- 一般性原則：
  - 通常我們會設定 5 到 15 個分組，太多太少都不好。
  - 如果存在異常值，他們應該先被剔除。

分組	次數
$[0, 1000)$	18
$[1000, 2000)$	80
$[2000, 3000)$	74
$[3000, 4000)$	107
$[4000, 5000)$	166
$[5000, 6000)$	106
$[6000, 7000)$	86
$[7000, 8000)$	82
$[8000, 9000)$	12

更多資訊

- 我們可以增加組中點（class midpoint）、相對次數（relative frequency）以及 累計次數（cumulative frequency）：

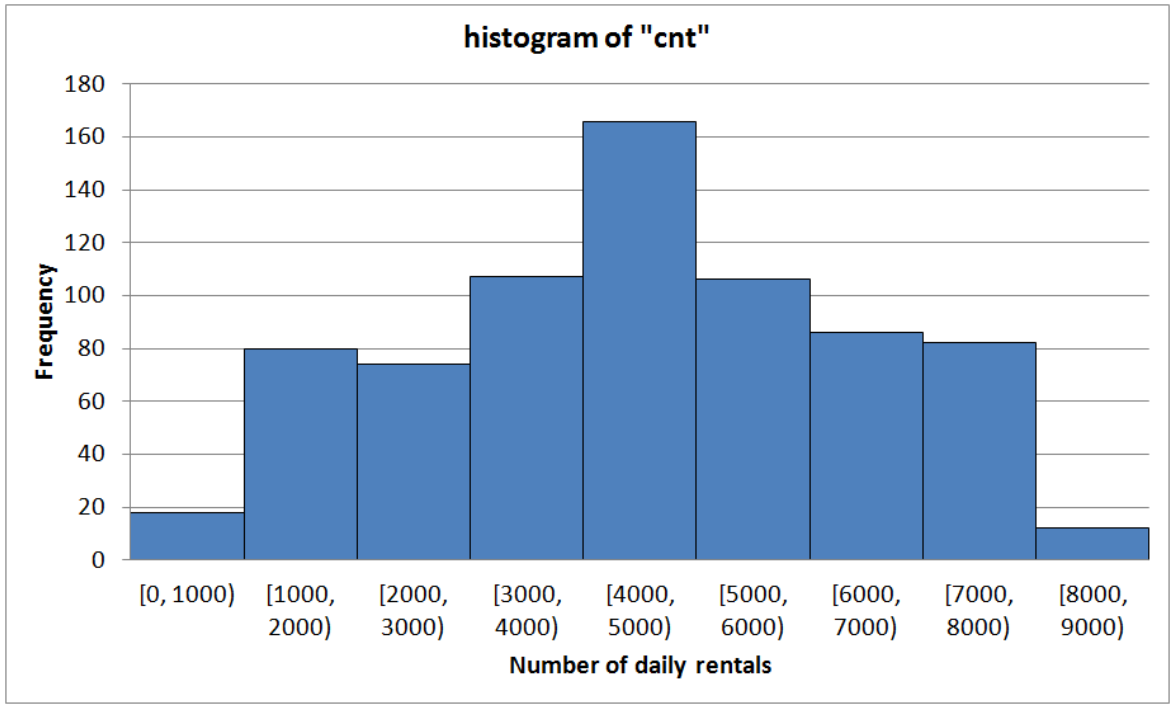
分組	次數	分組 組中點	相對 次數	累計 次數
$[0, 1000)$	18	500	2.46%	18
$[1000, 2000)$	80	1500	10.94%	98
$[2000, 3000)$	74	2500	10.12%	172
$[3000, 4000)$	107	3500	14.64%	279
		⋮		
$[8000, 9000)$	12	8500	1.64%	731

- 那如果是累計相對次數呢？

直方圖

- 我們經常用一個直方圖（histogram）來視覺化一個次數分佈。
  - 以一連串的連著的長方形組成，其高度代表一個分組的次數。

分組	次數
$[0, 1000)$	18
$[1000, 2000)$	80
$[2000, 3000)$	74
$[3000, 4000)$	107
$[4000, 5000)$	166
$[5000, 6000)$	106
$[6000, 7000)$	86
$[7000, 8000)$	82
$[8000, 9000)$	12

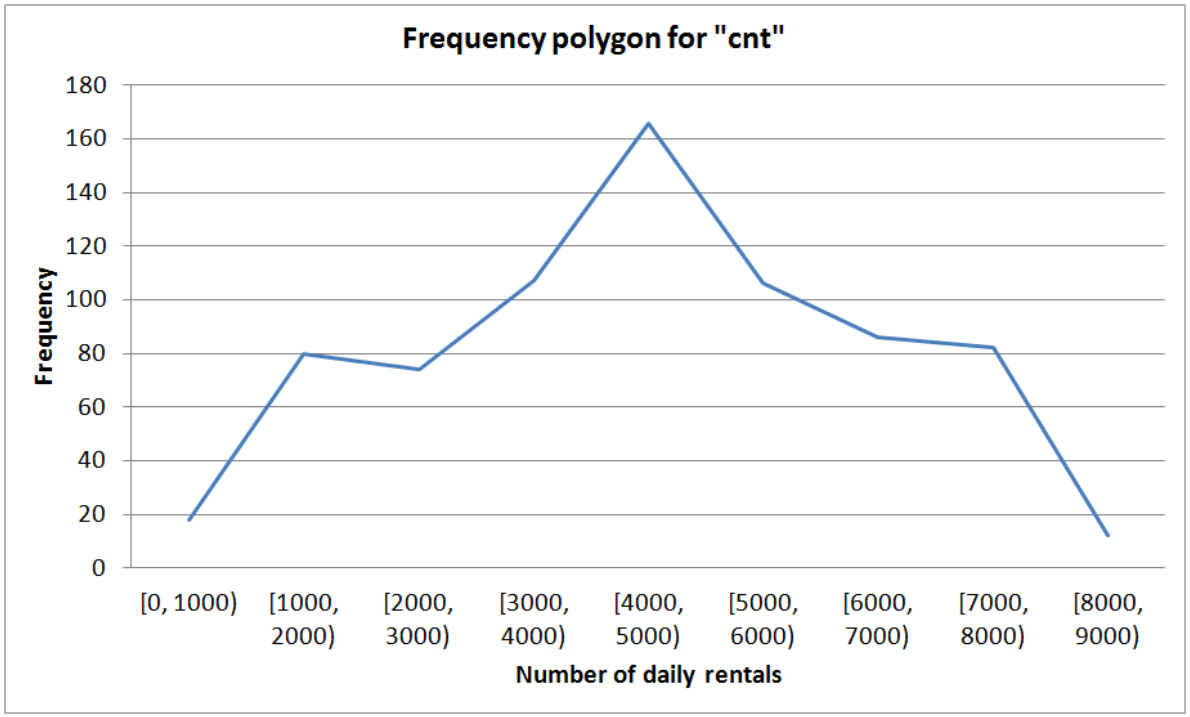


直方圖

- 直方圖或許是最重要的資料圖表類型。
- 繪製直方圖的一個主要的原因，是為了獲得一些資料分佈的概念。
  - 鐘型？M 型？偏態？
  - 異常值？

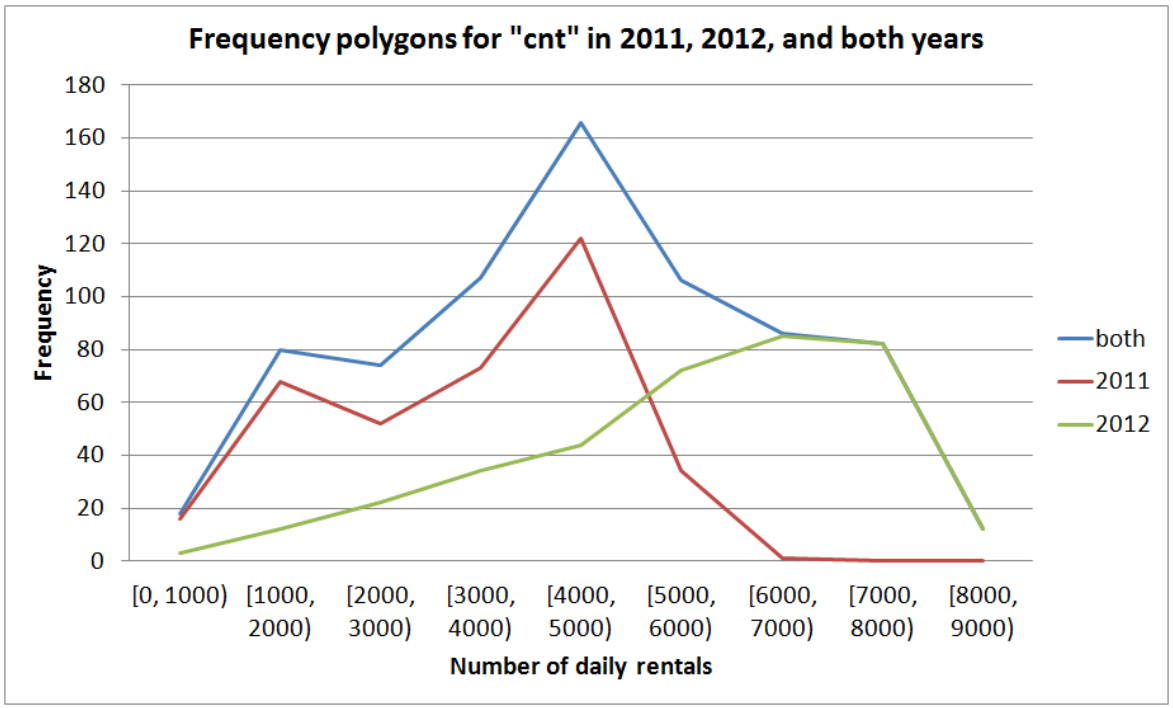
次數曲線圖

- 如果不想畫柱子，我們也可以連結各柱子頂端的組中點，組合這些線段來畫一個次數曲線圖（frequency polygon）。
  - 次數曲線圖所含的資訊與直方圖基本上相同。



次數曲線圖

- 使用次數曲線圖可以比較方便地比較多個次數分佈。

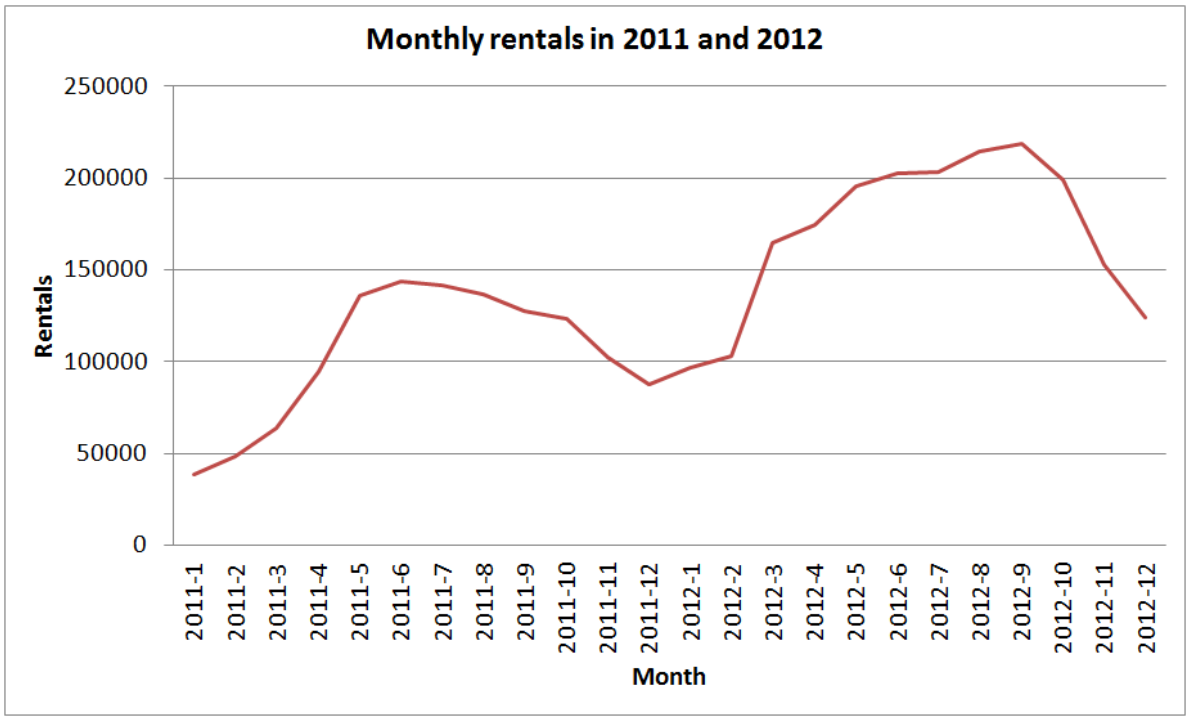


- 兩年合計：單峰型且對稱分佈。
- 2011: 雙峰型且右尾(長尾在右)。
- 2012: 單峰型且左尾(長尾在左)。

- 缺點：讀者可能會誤以為你畫的是折線圖。

折線圖

- 折線圖（line chart）被用於描繪時間序列的資料。
  - 圖的 x 軸標示的是時間。
  - 視覺化某個數量如何隨著時間變化。
- 我們每月的腳踏車租賃：

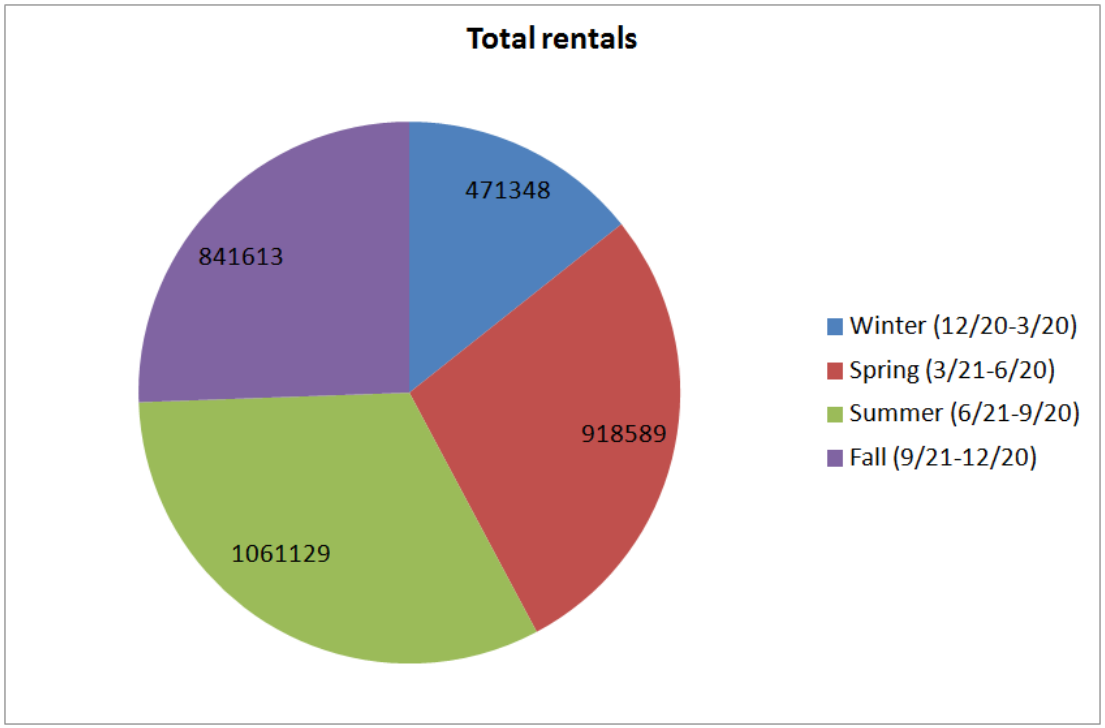


圓餅圖

- 圓餅圖（pie chart）是一個以圓形內每個區塊要來表示對應的品類所佔的百分比。
- 它很適合視覺化相對次數分佈（也就是各分組的比例）。
- 我們每月的腳踏車租賃：
  - 四個季節分別佔整個租賃的多少比例？
  - 星期一到星期日分別佔整個租賃的多少比例？

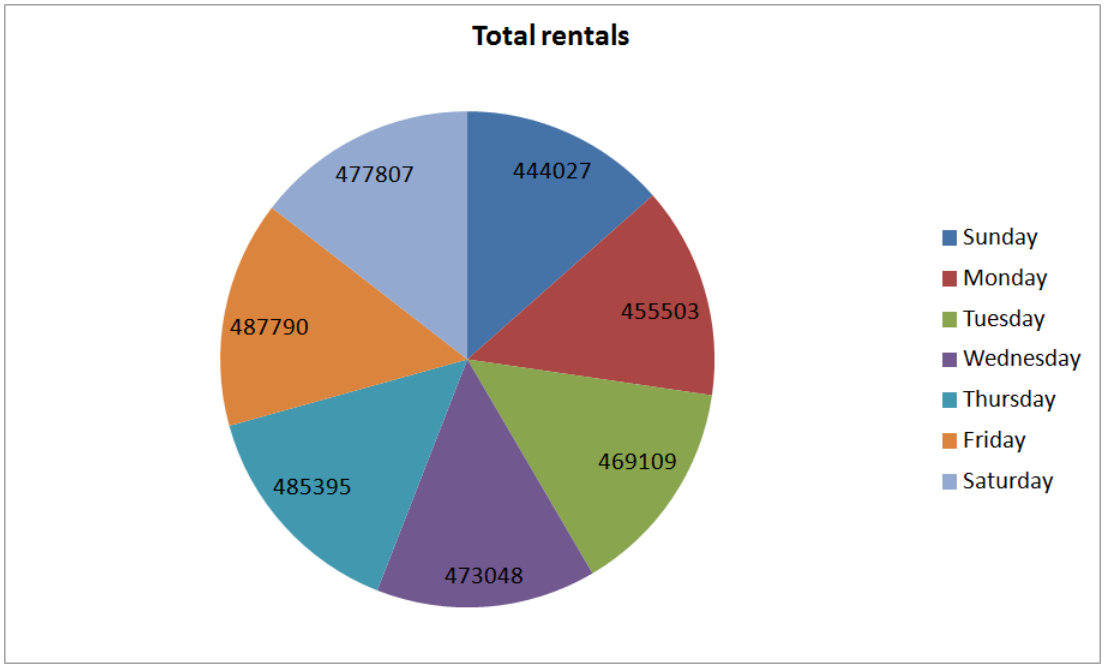
季節性的租賃圓餅圖

季節	總租賃數	佔比
冬天	471348	14.3%
春天	918589	27.9%
夏天	1061129	32.2%
秋天	841613	25.6%



星期一到星期日的租賃圓餅圖

日子	總租賃數
星期日	444027
星期一	455503
星期二	469109
星期三	473048
星期四	485395
星期五	487790
星期六	477807



不適合圓餅圖的資料

- 圓餅圖是用於視覺化各組的佔比，也就是各組佔整體的比例。
- 它不應該用於比較平均。
  - 男性與女性使用者的總租賃數適合呈現在圓餅圖。
  - 但是男性與女性的每人平均租賃次數不應該以圓餅圖呈現。

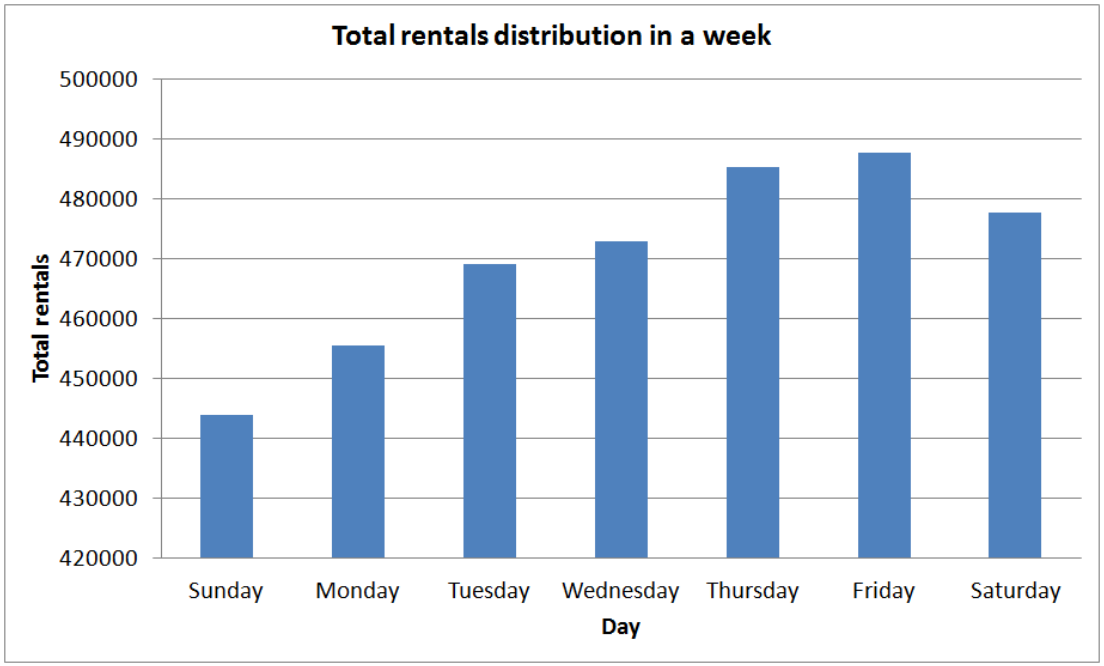
長條圖

- 圓餅圖適合視覺化不同類別的佔比。
- 而展示不同類別間的差異，長條圖（bar chart）是個更好的選擇。
  - 越大的類別，該長條就會越長。
  - 很多時間差異在圓餅圖上不明顯，此時用長條圖就可以清楚呈現。
  - 有些人將長條圖繪製成垂直的，有些則是水平的。

長條圖

- 讓我們把圓餅圖替代成長條圖吧！

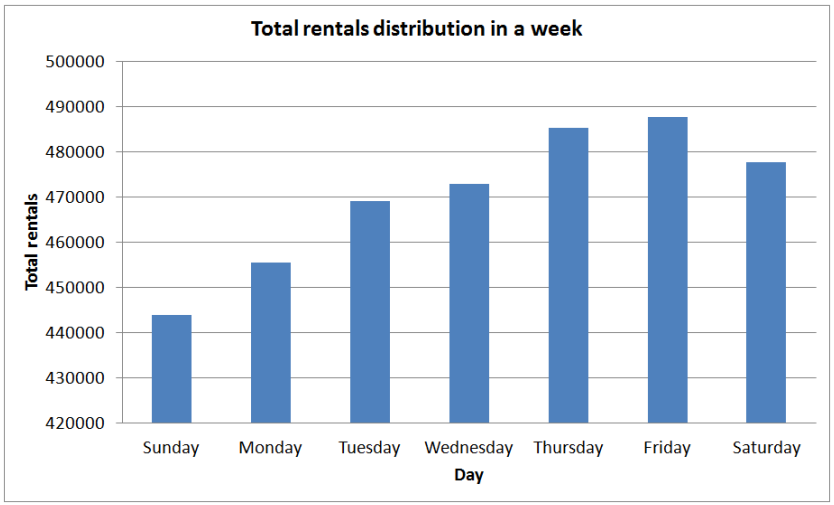
日子	總租賃數
星期日	444027
星期一	455503
星期二	469109
星期三	473048
星期四	485395
星期五	487790
星期六	477807

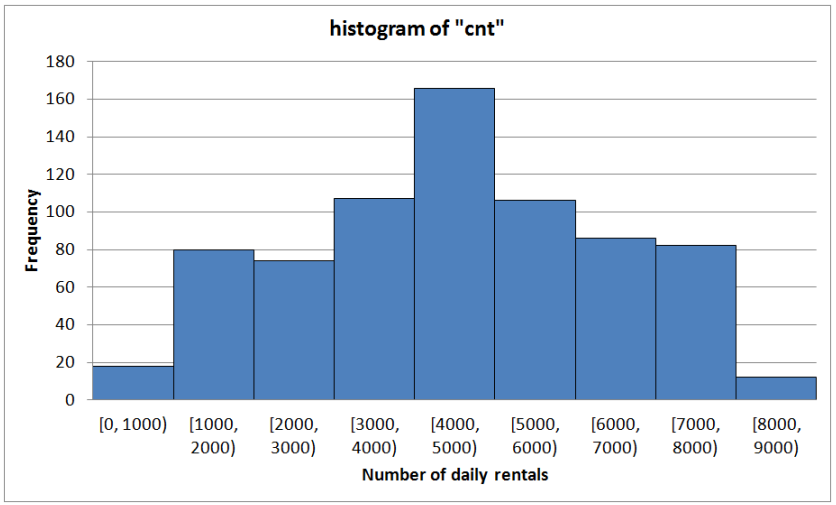


- 這張圖上 y 軸並非從 0 開始。
  - 當你要強調各組間的差異時，你可以這麼做。
  - 你應該明確地提醒讀者這件事。

長條圖 vs. 直方圖

- 長條圖和直方圖有何不同？





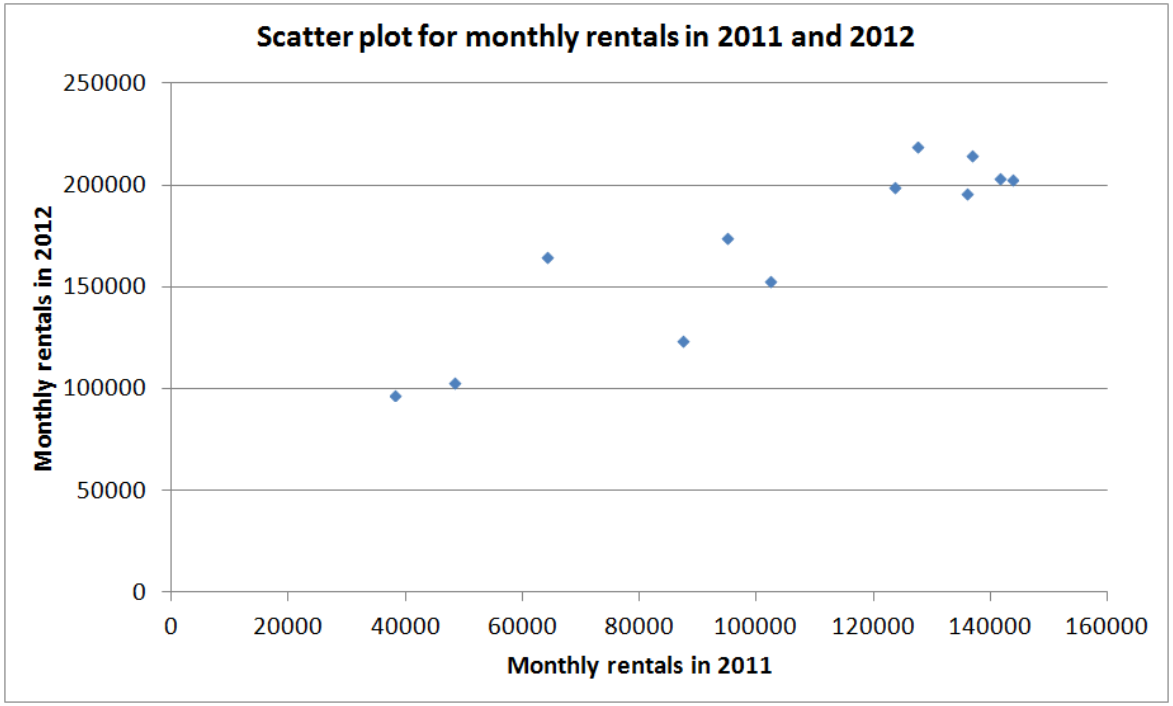
- 直條圖使用不連續的直條來視覺化類別型（categorical）資料。
- 直方圖使用連續的直條來視覺化數值型（numeric）資料。

視覺化兩個變數間的關係

- 當我們有兩個變數的資料，如何了解他們彼此之間有何關係？
- 若兩個變數都是數值資料，我們可以將每筆資料視作平面上的一個點，而畫出散佈圖（scatter plot）。
- 我們每月的腳踏車租賃例子:
  - 2011 和 2012 每月租借彼此間有什麼關係？

2011 和 2012 每月租賃

月份	2011	2012
1	38189	96744
2	48215	103137
3	64045	164875
4	94870	174224
5	135821	195865
6	143512	202830
	⋮	
11	102167	152664
12	87323	123713



- 大致分佈在一條斜率為正的直線上：高度正相關。

課程大綱

- 資料視覺化
- 資料摘要

將資料以數字取摘要

- 我們也可以用數字來做資料摘要。
  - 對於一組（很多個）數字，我們使用幾個數字來表現一些性質。
- 嚴謹地說，對於母體和對於樣本的摘要，具有不同意義：
  - 對於母體：這些數字是個參數。
  - 對於樣本：這些數字是統計量。
  - 這份教材只討論對母體的摘要。
- 我們會談三件事：
  - 測量集中趨勢（central tendency）來觀測中間段或中心資料。
  - 測量變異度（variability）來觀測資料的變異性。
  - 測量相關性（correlation）來了解兩個變數間的關係。

中位數

- 中位數（median）是位於一串已排序數字列的中間部份的量值。
  - 粗略而言，有一半的數字比中位數小，另一半則比較大。
- 假設有 N 個數字：
  - 如果 N 是奇數，那麼中位數就是第 $\frac{N+1}{2}$ 大的數字。
  - 如果 N 是偶數，那麼中位數就是第 $\frac{N}{2}$ 大和第 $(\frac{N}{2}+1)$ 大的數字的平均。
- 例如：
  - $\{1,2,4,5,6,8,9\}$ 的中位數就是 5。
  - $\{1,2,4,5,6,8\}$ 的中位數就是 $\frac{4+5}{2}=4.5$。

中位數

- 中位數不會受到極端值的影響：
  - $\{1, 2, 4, 5, 6, 8, 9\}$ 的中位數是 5。
  - $\{1, 2, 4, 5, 6, 8, 900\}$ 的中位數還是 5。
- 不幸地，中位數只使用了這些數字提供的部份資訊。
  - 只考慮次序，不考慮大小。

平均數

- 平均數（mean）是一組資料的平均。
  - $\{1, 2, 4, 5, 6, 8, 9\}$ 的平均數是

$$\frac{1+2+4+5+6+8+9}{7} = 5.$$

- 平均數使用所有涵蓋在這些數字裡的資訊。
- 但不幸地，平均數會受到極端值的影響。
  - $\{1, 2, 4, 5, 6, 8, 900\}$ 的平均數是 $\frac{1+2+4+5+6+8+900}{7} \approx 132.28$!
  - 同時呈現中位數和平均數是個比較好的做法
  - 在計算平均值（或是其他統計量）前，我們應該試著剔除異常值（那些看起來「奇怪」的極端值）。

四分位數與百分位數

- 中位數位於整個資料的中間。
- 第一四分位數 (first quartile) 位於前半部資料的中間。
- 第三四分位數 (third quartile) 位於後半部資料的中間。
- 第 p 個百分位數 (pth percentile):
  - 有 $\frac{p}{100}$ 的數比他小。
  - 有 $1 - \frac{p}{100}$ 的數比他大。
- 中位數、四分位數和百分位數：
  - 第 25 百分位數是第一四分位數。
  - 第 50 百分位數是中位數 (也是第二四分位數)。
  - 第 75 百分位數是第三四分位數

眾數

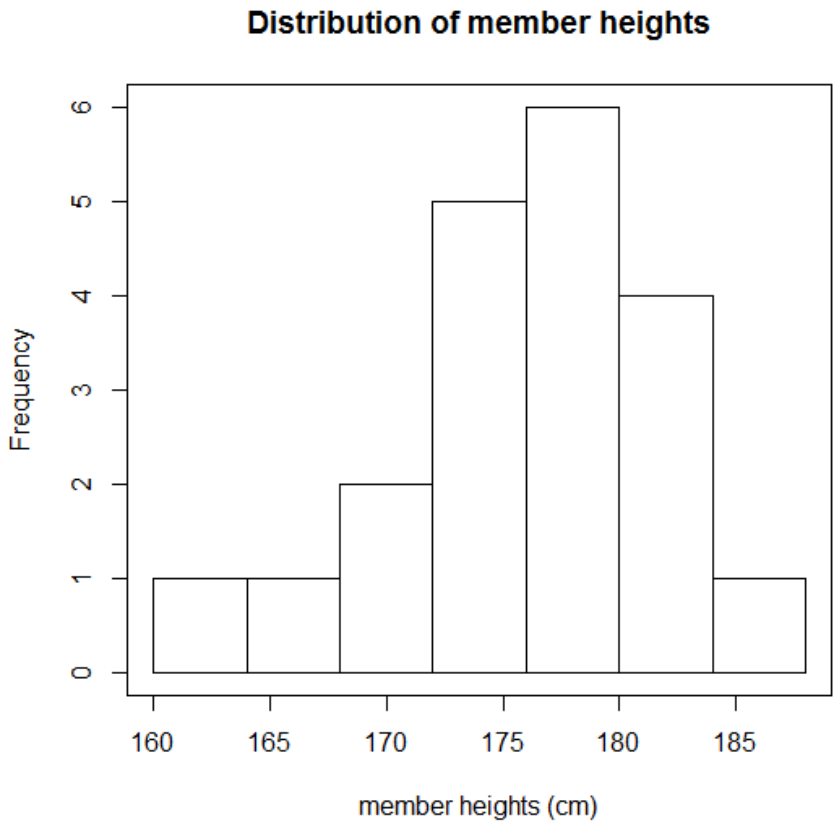
- 眾數（mode）是在一組資料中出現最多次的資料值。
  - 在 $\{A, A, A, B, B, C, D, E, F, F, F, G, H\}$ 之中，眾數是 A 與 F。這兩個眾數（A 與 F）出現的次數為 3。
  - 眾數是 A 與 F，不是 3。
  - 眾數可能有多個。
- 儘管以上的定義或許也適用於數值資料，但有時候會失效。
  - 在許多情況下，所有數值都是眾數！
- 對於數值資料，我們會更傾向於找尋眾數分組（可能有多個）。

眾數組

- 在一個棒球隊裡，球員的身高（公分）為：

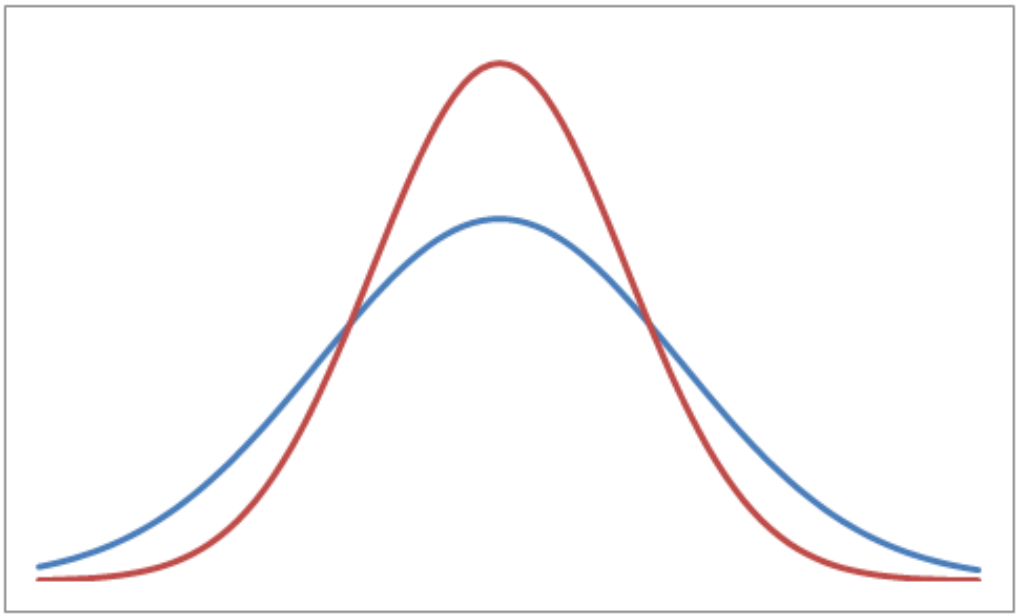
178	172	175	184
172	175	165	178
177	175	180	182
177	183	180	178
179	162	170	171

- 對於 $[160, 165)$、$[165, 170)$、... 等組別，眾數組為 $[175, 180)$。
- 我們有時候說這組資料的眾數是 177.5。
- 分組的方式會有影響！



變異性

- 我們經常也想描述一組資料的分散或離散程度。
- 在兩組資料有相同中心點時，描述變異性特別重要。



全距與四分位距

- 一組資料 $\{x_i\}_{i=1,\ldots,N}$ 的全距（range）是最大和最小數值間的差異，即

$$\max_{i=1,\ldots,N}\{x_i\} - \min_{i=1,\ldots,N}\{x_i\}.$$

- 一組資料的四分位距（inter-quartile range）是第一四分位數和第三四分位數間的差異。
  - 它是中間 50%資料的全距。
  - 它排除了極端值的影響。

與母體平均的差異

- 考慮一組母體資料 $\{x_i\}_{i=1,\ldots,N}$，其平均數為 $\mu = \frac{\sum_{i=1}^{N}}{N}$。
- 直覺上，一種測量離散程度的方式變是測試各個數字與母體平均的差異。
- 對於每個 x_i，與母體平均的差異被定做

$$x_i - \mu.$$

i	x_i	差異
1	1	$1-5=-4$
2	2	$2-5=-3$
3	4	$4-5=-1$
4	5	$1-5=0$
5	6	$6-5=1$
6	8	$8-5=3$
7	9	$9-5=4$
平均數	5	

平均差

- 我們可以總結 N 個差異於單一數字來概述這些差異嗎？
- 直覺上，我們會想把這些差異加總並計算平均差 (mean deviation)：

$$\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}.$$

- 是否永遠都等於 0？

i	x_i	差異
1	1	$1-5=-4$
2	2	$2-5=-3$
3	4	$4-5=-1$
4	5	$1-5=0$
5	6	$6-5=1$
6	8	$8-5=3$
7	9	$9-5=4$
平均數	5	0

調整平均差

- 有兩種常用的方式來調整平均差：
  - 平均絕對差異 (mean absolute deviation，MAD)：
$$\frac{\sum_{i=1}^{N} |x_i - \mu|}{N}.$$
  - 平均平方差異 (mean squared error，MSE)：
$$\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}.$$
  - MSE 比較常用，通常被稱為變異數 (variance)。
- 愈大的 MAD 或變異數表示資料愈離散。

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	1	-4	4	16
2	2	-3	3	9
3	4	-1	1	1
4	5	0	0	0
5	6	1	1	1
6	8	3	3	9
7	9	4	4	16
平均	5	0	2.29	7.43

MAD vs. 變異數

- MAD 將所有值都使用相同權重，變異數則會在極端值放上更多的權重。
- 它們可能給出不同排序的離散度：

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	0	-5	5	25
2	4	-1	1	1
3	5	0	0	0
4	6	1	1	1
5	10	5	5	25
平均	5	0	2.4	10.4

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	1	4	4	16
2	2	3	3	9
3	5	0	0	0
4	8	3	3	9
5	9	4	4	16
平均	5	0	2.8	10

- 一般而言，人們使用變異數多於 MAD。
  - 但是 MAD 還有有其受歡迎的領域，像是需求預測。
  - 分析師可以自己斟酌選擇較為合適者。

標準差

- 使用變異數的一個缺點：測量的單位是原始單位的平方。
- 對於我們的棒球隊，成員身高的變異數是 34.05 公分 ^{2}。那是什麼 ?!
- 人們將變異數開根號來得到標準差 (standard deviation)。
- 成員身高的標準差是

$$\sqrt{34.05} \approx 5.85 \text{ 公分}.$$

- 標準差通常比較有管理意涵。

178	172	175	184
172	175	165	178
177	175	180	182
177	183	180	178
179	162	170	171

變異係數

- 變異係數是標準差與平均數的比值：

$$\text{變異係數} = \frac{\sigma}{\mu}.$$

- 你何時會使用到變異係數呢？

z-score

- 對於一組資料 $\{x_i\}_{i=1,\ldots,N}$，若其平均數為 μ，標準差為 σ，則 x_i 的 z-score 為

$$z_i = \frac{x_i - \mu}{\sigma}.$$

- z-score 衡量一個值離平均數距離幾個標準差。

z-score vs. 異常值

- 欲找出異常值，一個常見的條件是看 x_i 是否滿足

$$|z_i| = \left|\frac{x_i - \mu}{\sigma}\right| > 3.$$

  - 不會有太多數值的 z-score 很大或很小。

- 有些人運用中位數和 MAD[1]：

$$\left|\frac{x_i - \text{中位數}}{\text{MAD}}\right| > 3.$$

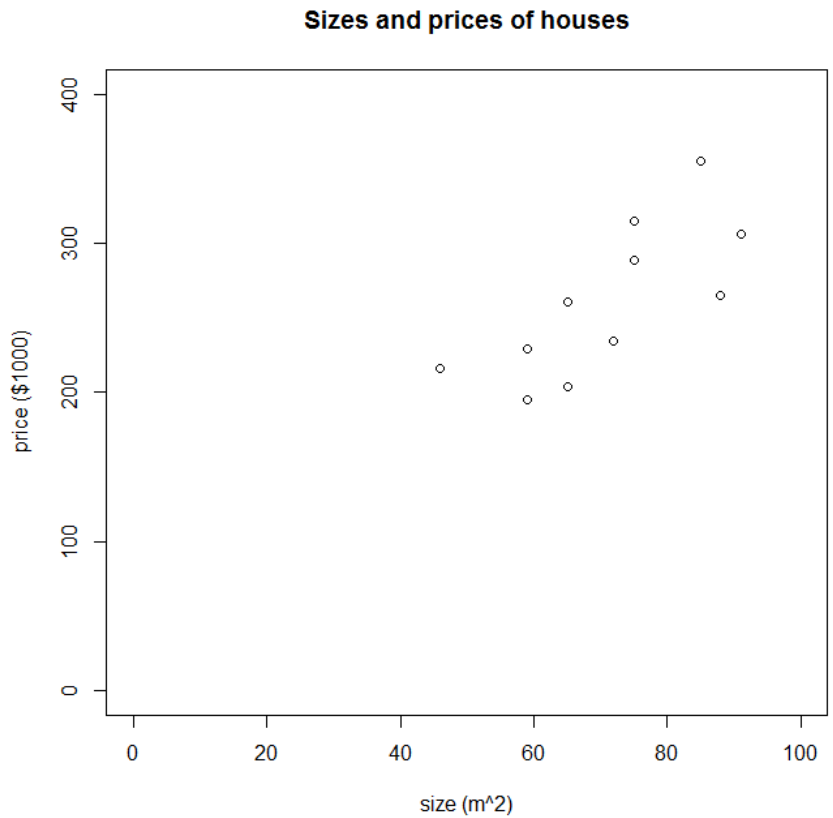
- 以上規則只能建議你去看看。它們對於異常值既不充分也不必要。

[1]「MAD」在這裡可以指相比平均的平均絕對離差、相比中位數的平均絕對離差，及相比中位數的絕對離差中位數等。

相關性

- 考慮房子的大小以及它在城市的價格：

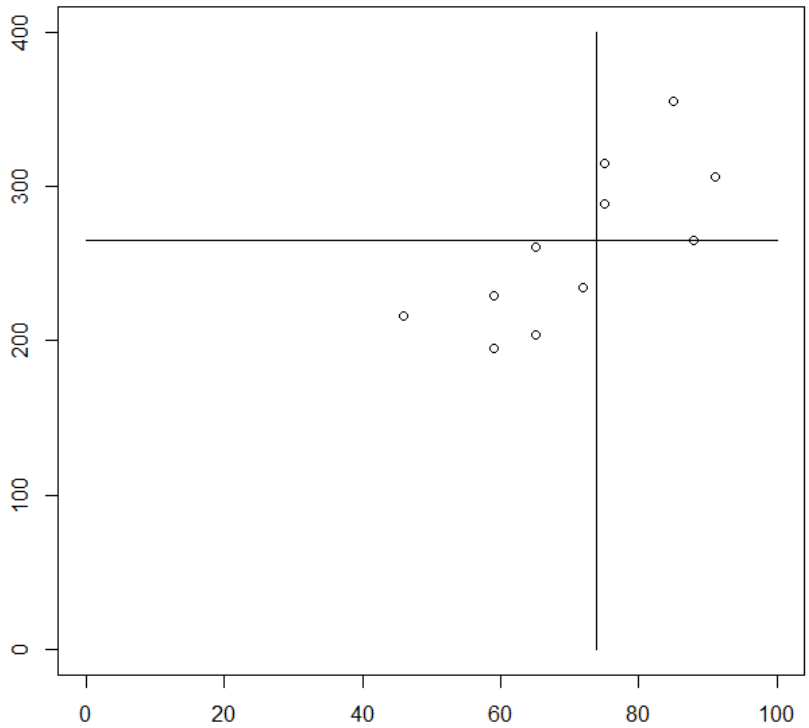
大小（平方公尺）	價格（$1000）
75	315
59	229
85	355
65	261
72	234
46	216
107	308
91	306
75	289
65	204
88	265
59	195



- 我們該如何測量/描述兩遍數間的相關性（線性關係）呢？

直觀

- 考慮成對資料 $\{(x_i, y_i)\}_{i=1,\ldots,N}$。
- 當其中一個變數上升時，另一個變數會傾向上升或是下降呢？
- 更精確地說，當 x_i 比 μ_x（所有 x_i 的平均）大時，比較有機會看到 $y_i > \mu_y$ 還是 $y_i < \mu_y$ 呢？
- 如果一個變數上升時另一個變數通常也上升，我們說兩個變數有正相關；反之則負相關。



共變異數

- 我們定義二維資料的共變異數（covariance）為

$$\sigma_{xy} \equiv \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

  - 如果大多數的資料點落在第一和第三象限，大多數的 $(x_i - \mu_x)(y - \mu_y)$ 會是正的，而且 σ_{xy} 會傾向為正的。
  - 否則，σ_{xy} 會傾向為負的。
- 所以房子大小和價格的共變異數為 617.16。
- 這樣算大還是小呢？
  - 這取決於這兩個變數的自身變異程度（auto-covariance）。

相關係數

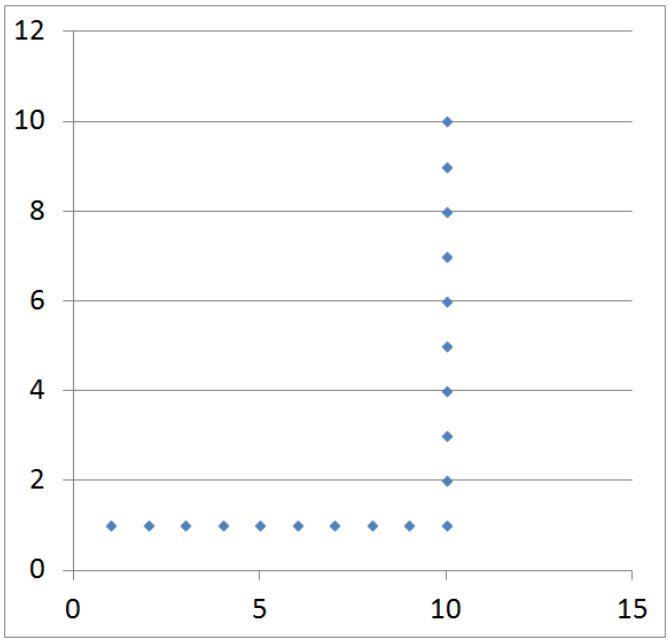
- 為了去除自身變異，我們定義相關係數（correlation coefficient）為

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

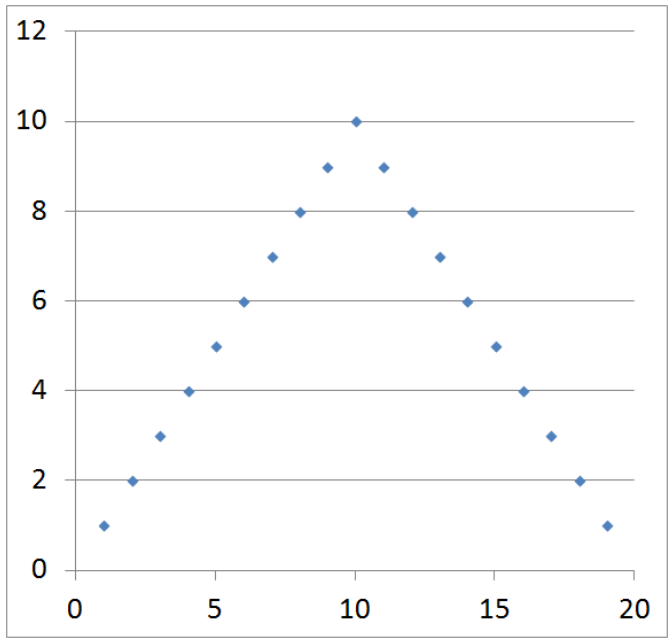
  - σ_x 和 σ_y 為 x_i 和 y_i 的標準差。
  - 在我們的例子裡，$\rho = \frac{617.16}{16.78 \times 50.45} \approx 0.729$。
- 可以發現，我們永遠都會得到 $-1 \leq \rho \leq 1$。
  - $\rho > 0$、$\rho = 0$ 和 $\rho < 0$ 分別表示正相關、無相關和負相關。
- 人們通常基於 $|\rho|$ 來決定相關性的程度：
  - $0 \leq |\rho| < 0.25$：弱相關。
  - $0.25 \leq |\rho| < 0.5$：中度弱相關。
  - $0.5 \leq |\rho| < 0.75$：中度強相關。
  - $0.75 \leq |\rho| \leq 1$：強相關。

相關性 vs. 獨立性

- 相關係數只能量測兩個變數間的線性關係。



$(\rho = 0.5973)$



$(\rho = 0)$

- 沒有線性相關不代表獨立（或無關）!

相關性 vs. 因果性

- 相關係數只能量測兩個變數是否相關。高度相關無法代表具因果性。

(http://www.tylervigen.com/spurious-correlations)

- A 導致 B，還是 B 導致 A？C 導致 A 和 B？還是純屬巧合？

給工程師的統計學與資料分析 123

第一單元：基本概念與抽樣分佈

孔令傑

國立臺灣大學資訊管理學系

2017 年 9 月 2 日

什麼是統計？

- 很多事情是未知的...
  - 顧客的喜好、產品的品質、股票明天的收盤價、新教學方法的有效性。
- 統計是一門收集、分析、闡釋及表達資料的科學。
  - (商業統計的) 最終目的：達到更好的決策。
- 統計學包含：
  - 敘述統計 (descriptive statistics)。
  - 機率。
  - 推論統計：估計 (estimation)。
  - 推論統計：假設檢定 (hypothesis testing)。
  - 推論統計：解釋變異 (variability explanation)。
- 總結：去估計、檢定這些未知，並且解釋變異。

今天的計畫

- 敘述統計：
  - 視覺化與摘要。
- 機率。
- 推論統計：
  - 抽樣分佈。
  - 假說檢定與 p-value。
  - 迴歸分析。

課程大綱

- 基本概念。
- 抽樣。
- 抽樣分佈：樣本平均數。
- 抽樣分佈：樣本比例。

母體 vs. 樣本

- 母體（population）是人、物件和物品的集合。
  - 普查（census）就是針對整個母體進行探查。
- 樣本（sample）是母體的一部分。
  - 我們以抽樣（sampling）探查母體的子集合。
  - 我們會用樣本包含的資訊去推論（猜測）母體。
- 以下幾個母體的樣本分別為何呢？
  - 全台大的學生。
  - 全商管學院的學生。
  - 在同一個工廠生產的全部晶片。
  - 所有購買 iPhone 6 的顧客。
- 兩個重要的問題：
  - 為什麼要抽樣？
  - 樣本是否具有代表性？

敘述統計 vs. 推論統計

- 敘述統計 (descriptive statistics)：
  - 描述 (視覺化或是摘要) 一組資料。
- 推論統計 (inferential statistics)：
  - 「以科學的方式」對未知的母體「進行猜測」。
- 哪個是敘述，哪個是推論呢？
  - 計算 1000 個隨機挑選的臺大學生的平均身高。
  - 使用這個數字去推估全臺大學生的平均身高。
- 另一個例子 (製藥研究)：
  - 母體：全部潛在病患。
  - 樣本：隨機挑選的一群病患。
  - 使用這個樣本的結果去推估整個母體。

參數 vs. 統計量

- 母體的數值摘要是個參數（parameter）。
  - 全部臺大學生的平均身高。
  - 當價格落在新台幣 50 元時，咖啡的預期需求。
- 樣本的數值摘要是統計量（statistic）。
  - 全部臺大男性學生的平均身高。
  - 過去 6 天當價格落在新台幣 50 元時，咖啡的平均預期需求。
- 人們幾乎總是用統計量來推論參數。
  - 有些統計量是「好的」，有些則是「壞的」。

參數 vs. 統計量：一個例子

- 全部臺大學生的平均身高是多少？
- 儘管普查是可能的，但總是挺貴的。
- 很自然的，我們會去：
  - 抽一些臺大學生。
  - 計算統計量。
  - 用這個統計量去推估平均身高（參數）。
- 一些（好的或壞的）樣本及統計量：
  - 全體管理學院學生的平均身高。
  - 從全部學生中隨機挑選 100 位的平均身高。
  - 從全部學生中隨機挑選 100 位裡最高的身高。
  - 從全部學生中隨機挑選 100 位的加總身高。
  - 從男性學生中隨機抽出 60 個、女性學生中抽出 40 個，取他們的平均身高。

資料型態

- 資料依照型態不同，可以被分成兩大類：
  - 類別資料 (qualitative or categorical data)。
  - 數值資料 (quantitiative or numeric data)。
- 類別資料又分為：
  - 名目資料 (nominal)。
  - 次序資料 (ordinal)。

名目資料

- 名目資料中的值是數個不具排序性的類別。
- 值可能看起來像數字，但不能拿來做加減乘除，也不具大小關係。
- 舉例：

類別變數	值（類別）
是否吃素	是、否
國籍	臺灣、日本...
國家代碼	886、86、1...

- 不同的值不能排序，也不能做算術運算。

次序資料

- 次序資料的值依然是類別，但是順序是有意義的。
- 舉例：

類別變數	值（類別）
產品滿意度	滿意、沒意見、不滿意
教授等級	正、副、助理
班排名	1、2、3、4.....

- 對次序資料進行算術運算仍然不具意義。
  - 助理教授 + 副教授 = 正教授 ?!
  - 第一名和第五名的差距有可能不等於第十一名和第十五名的差距。

數值資料

- 數值資料是真正的數量，可以排序，也可以做算術運算。
  - 身高、體重、收入、價格。
  - 華氏或攝氏溫度。
- 課本上常將數值資料分成間隔（interval）資料和比例（ratio）資料。
  - 不是很好分，也不是很重要（個人意見）。

小結

- 了解這些名詞：
  - 母體 vs. 樣本。
  - 參數 vs. 統計量。
  - 推論統計 vs. 敘述統計。
- 資料尺度：
  - 名目和次數資料被稱做類別資料或質性資料。
  - 間隔和比例資料被稱做數值資料或量化資料。
- 不同統計方法有不同適用範圍和應用方式。
  - 區分類別資料和數值資料非常重要。
  - 區分名目資料和次序資料有時也很重要。

課程大綱

- 基本概念。
- 抽樣。
- 抽樣分佈：樣本平均數。
- 抽樣分佈：樣本比例。

隨機 vs. 非隨機抽樣

- 抽樣是一個從整個母體挑選子集合的過程。
- 抽樣可以是隨機的或確定型的。
- 如果是隨機的，任一個個體是否會被抽到就是隨機的。
  - 今天抽跟明天抽（原則上）會得到不一樣的結果。
  - 從電話簿隨機挑選 1000 個電話號碼，並打給他們。
- 如果非隨機，那就是確定型的。
  - 詢問你所有一等親對於 iOS/Android 的偏好。
- 大部份統計方法只適用於隨機抽樣。
- 一些知名的隨機抽樣方法：
  - 簡單隨機抽樣。
  - 分層隨機抽樣。
  - 群集（或區域）隨機抽樣。

簡單隨機抽樣

- 在簡單隨機抽樣（simple random sampling），每個個體被挑選到的機率相同。
- 簡單隨機抽樣的好處就是簡單。
- 但是如果運氣不好，就可能會得到不具代表性的樣本。
  - 有機會出現太多樣本資料落在同一層，亦即有相同的屬性。
  - 比如說，可能所有隨機抽樣的投票者都小於 40 歲。
  - 那麼這個樣本便不具代表性。
- 要怎麼改善這個問題呢？

分層隨機抽樣

- 我可以運用分層隨機抽樣（stratified random sampling）。
- 首先，我們將整個母體分成數個層（stratum）。
  - 在同一層內的資料應該（相對）同質（homogeneous）
  - 在不同層內的資料則應該（相對）異質（heterogeneous）。
- 我們再在各層內進行簡單隨機抽樣。

分層隨機抽樣

- 假設我們想要從 1000 個畢業生中抽出 40 位來了解他們在學校取得多少學分。
- 假設有 100 個畢業生當年有雙主修，那我們可以將整個母體分成兩層：

分層	分層大小
雙主修	100
非雙主修	900

- 我們從雙主修學生中抽 $40 \times \frac{100}{1000} = 4$ 人，從非雙主修的抽 36 人。

分層隨機抽樣

- 我們可以將母體分成更多層。
  - 雙主修：是或否。
  - 畢業年份：1994-1998、1999-2003、2004-2008 或 2009-2012。
  - 在不同年代的學生是否傾向於修不同數量的學分？
- 分層隨機抽樣適合降低抽樣偏誤。
- 它也同時較為昂貴且費時，而且有時不容易找出一個合理的分層。

群集（或區域）隨機抽樣

- 想像你要到台灣全部的零售店推出新產品。
- 如果這個產品其實很不受歡迎，那麼大規模推出會產生很高昂的成本。
- 那要怎麼知道受歡迎的程度？
- 我們可以先在小區域介紹這個產品。我們僅將產品在特定的區域上架。
- 這就是群集（或區域）隨機抽樣（cluster sampling）的概念。
  - 樣本：在這些區域的客戶。

群集（或區域）隨機抽樣

- 在群集隨機抽樣，我們定義群集（cluster）。
- 我們只會選一個或少量的群集，然後收集在這些群集裡的所有資料。
  - 如果有一個群集過大，我們會將之再分成數個二階群集。
- 因此，我們想要在群集內的資料是異質的，而各群集都擁有同質資料。
- 例如，人們可以用群集隨機抽樣來了解一個新產品的受歡迎程度。那些被選擇的市場（城市、國家、州等）被稱作測試市場（、城市、國家、州等）。
  - 人們在這個情況下，使用群集隨機抽樣，是因為它的易用性和便利性。
  - 我們選擇的測試市場應該要與整個母體類似。

非隨機抽樣與小結

- 有的時候我們會做非隨機抽樣。
- 非隨機抽樣不能被接下來課程教的分析方法分析。
- 今天我們會假設所有抽樣都以隨機抽樣進行。
  - 也假設樣本具代表性。

課程大綱

- 基本概念。
- 抽樣。
- 抽樣分佈：樣本平均數。
- 抽樣分佈：樣本比例。

抽樣分佈

- 當我們沒有辦法探測整個母體時，我們便研究樣本。
  - 隨機樣本裡會包含什麼是無法預測的。
  - 我們需要知道樣本的機率分佈才能連結樣本與母體。
- 機率分佈：
  - 白話：可能的值，以及每個可能的值的可能性。
  - 數學上：樣本空間 (sample space)、機率密度函數 (probability density function，pdf)、累積分佈函數 (cumulative distribution function，cdf)。
- 樣本的機率分佈就是抽樣分佈 (sampling distribution)。

抽樣分佈

- 一個工廠生產糖果。
  - 理想上，糖果應該每包重 2 公斤。
  - 生產過程不可能完美，因此標準是每包糖果應該重 1.8 到 2.2 公斤之間。
- 令 X 為一包糖果的重量，μ 和 σ 為它的期望值和標準差。
  - $\mu = 2$ 嗎？
  - $1.8 < \mu < 2.2$ 嗎？
  - σ 有多大？
- 來抽樣吧：
  - 隨機抽一包，假設為 2.1 公斤，是否能說 $1.8 < \mu < 2.2$？
  - 如果在隨機樣本裡，五包的平均重量為 2.1 公斤呢？
  - 如果隨機樣本的大小是 10、50 或 100 呢？
  - 如果平均值是 2.3 公斤呢？
- 我們需要知道統計量（樣本平均數、樣本標準差等）的抽樣分佈。

樣本平均

- 樣本平均數（sample mean）是最重要的統計量之一。

定義 1

令 $\{X_i\}_{i=1,\ldots,n}$ 為從母體抽的一個樣本，那麼

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

就是樣本平均數。

- 有的時候我們用 $\bar{x}_n$ 來強調樣本大小是 n。
- 對於所有 $i \neq j$，我們假定 X_i 和 X_j 是獨立的。
  - 當 $n \ll N$，即我們從很大的母體抽出少量的項目，這樣假設就可以。
  - 實務上，我們需要 $n \leq 0.05N$。

樣本平均數的平均數和變異數

- 假設母體平均和變異數分別是 μ 和 σ^2。注意這兩個數字是固定的。
- 樣本平均 $\bar{x}$ 是個隨機變數。
  - 它有它的期望值 $\mu_{\bar{x}}$、變異數 $\sigma^2_{\bar{x}}$ 和標準差 $\sigma_{\bar{x}}$。這些數字都是固定的。
- 對於任何母體，我們有以下的定理：

定理 1 (樣本平均數的平均數和變異數)

令 $\{X_i\}_{i=1,\ldots,n}$ 為從母體抽出的樣本數為 n 的隨機樣本，而母體平均數為 μ、母體變異數為 σ^2，則我們有

$$\mu_{\bar{x}} = \mu \text{、} \quad \sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \text{ 且 } \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{。}$$

樣本平均的平均和變異數

- 這些名詞是否使你困惑？
  - 樣本平均數 vs. 樣本平均數的平均數。
  - 樣本變異數 vs. 樣本平均數的變異數。
- 就定義而言，它們：
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$；一個隨機變數。
  - $\mu_{\bar{x}} = \mathbb{E}[\bar{x}]$；一個常數項。
  - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$；一個隨機變數。
  - $\sigma_{\bar{x}}^2 = \mathrm{Var}(\bar{x})$；一個常數項。
- 樣本變異數也有它自己的平均和變異數。

例子：品質檢驗

- 每包糖果的重量服從常態分佈，平均數為 $\mu = 2$，標準差為 $\sigma = 0.2$。
- 假設品管長官決定要抽四包糖果並計算樣本平均 $\bar{x}$。如果 $\bar{x} \notin [1.8, 2.2]$，我就會受罰。
  - 我的生產流程其實是「好的」：$\mu = 2$。
  - 不幸地，它不是完美：$\sigma > 0$。
  - 我們可能還是會被懲罰（如果運氣不好），儘管 $\mu = 2$。
- 有多少的機率我會被懲罰呢？
  - 我們想要計算 $1 - \Pr(1.8 < \bar{x} < 2.2)$。
  - 我們知道 $\mu_{\bar{x}} = \mu = 2$ 且 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{4}} = 0.1$。
  - 但我們並不知道 $\bar{x}$ 的機率分佈！

從常態母體抽樣

- 如果母體是常態分佈，樣本平均數也會是常態分佈！

定理 2

令 $\{X_i\}_{i=1,\ldots,n}$ 為從常態母體抽出的樣本數為 n 的隨機樣本，母體平均數為 μ，標準差為 σ。則

$$\bar{x} \sim \mathrm{ND}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) 。$$

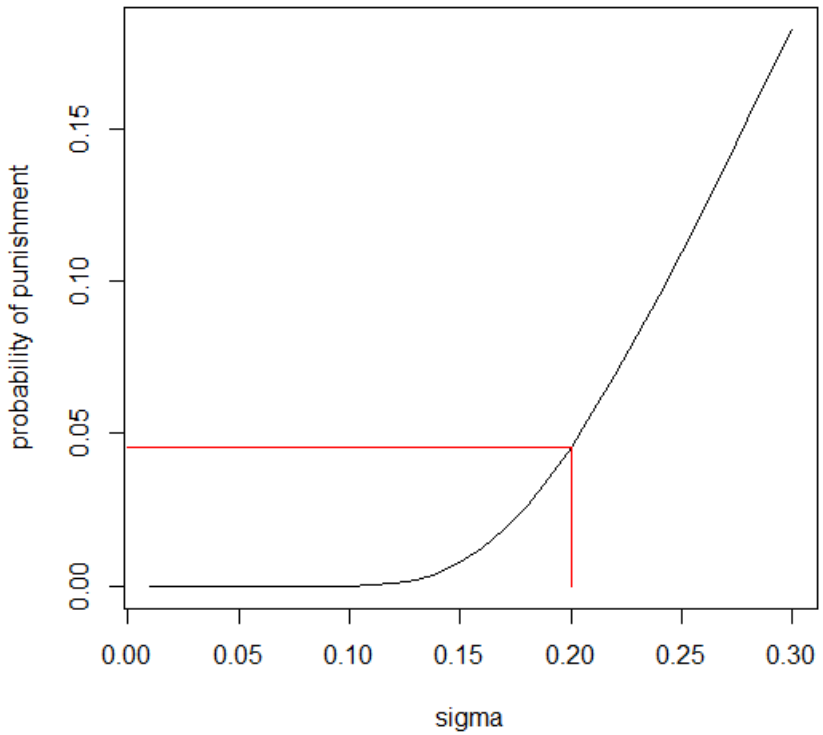
- 我們已知 $\mu_{\bar{x}} = \mu$ 且 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$。不論母體長怎樣，這都是對的。
- 當母體為常態分佈時，樣本平均數也會是常態分佈。

再回到這個例子：品質檢驗

- 每包糖果的重量服從常態分佈，平均數為 $\mu = 2$，標準差為 $\sigma = 0.2$。
- 假設品管長官決定要抽四包糖果並計算樣本平均 $\bar{x}$。如果 $\bar{x} \notin [1.8, 2.2]$，我就會受罰。
- 有多少的機率我會被懲罰呢？
  - 樣本平均數 $\bar{x}$ 的分佈為 $\mathrm{ND}(2, 0.1)$。
  - 受罰機率 $\Pr(\bar{x} < 1.8) + \Pr(\bar{x} > 2.2) \approx 0.045$。

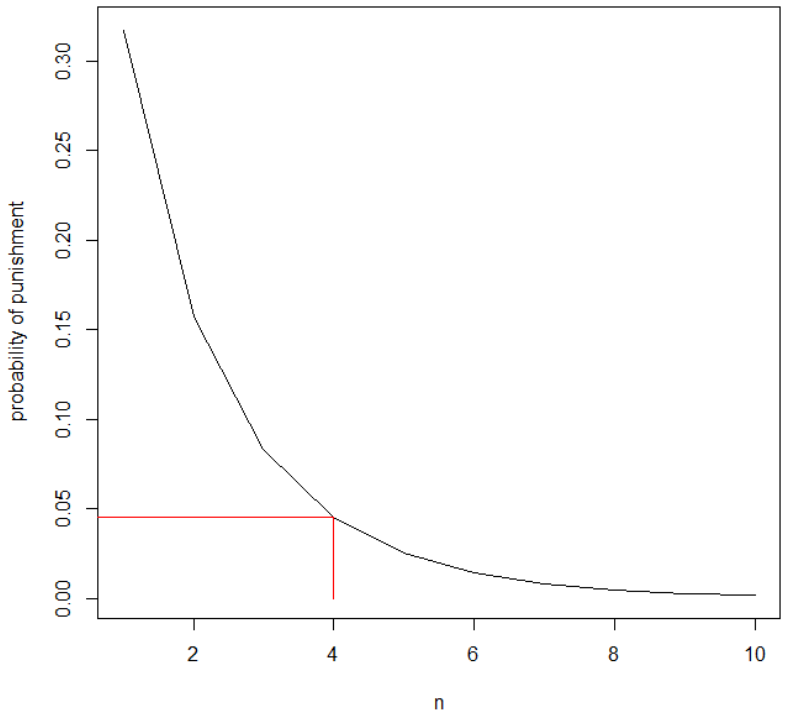
調整標準差

- 當母體為 $\mathrm{ND}(\mu = 2, \sigma = 0.2)$，而樣本大小為 $n = 4$，被懲罰的機率是 0.045。
- 如果調整標準差 σ（改進生產過程或變得更散漫），這個機率會改變。
- 降低 σ 會降低受罰機率。既然知道 $\bar{x}$ 的分佈，我們可以最佳化 σ。
  - 從 0.2 進步到 0.15 非常有幫助。
  - 從 0.15 進步到 0.1 則否。



調整樣本大小

- 當母體為 $\mathrm{ND}(2, 0.2)$，而樣本大小為 $n = 4$，被懲罰的機率為 0.045。
- 如果品管長官將樣本數量 n 增大，機率將會減少。
- $\mu = 2$ 其實是很符合品質要求的。較大的樣本數會降低受罰機率。



樣本平均數的分佈

- 我們現在知道，當我們從常態母體抽樣，樣本平均數也是常態。
  - 而且它的平均和標準差分別為 μ 及 $\frac{\sigma}{\sqrt{n}}$。
- 如果母體是非常態呢？
- 幸運地，我們有強大的中央極限定理（central limit theorem），可以被應用在任何母體。

中央極限定理

- 只要有足夠大的樣本數，樣本平均數會近似於常態分佈。

定理 3 (中央極限定理)

令 $\{X_i\}_{i=1,\ldots,n}$ 為從母體抽出的樣本數為 n 的隨機樣本，母體平均數為 μ，標準差為 σ。令 $\bar{x}_n$ 為樣本平均。只要 $\sigma < \infty$，則在 $n \to \infty$ 下，$\bar{x}_n$ 收斂至 $\mathrm{ND}(\mu, \frac{\sigma}{\sqrt{n}})$。

- 要多大才能算「足夠大」?
- 實務上，通常 $n \geq 30$ 被相信是足夠大。

課程大綱

- 基本概念。
- 抽樣。
- 抽樣分佈：樣本平均數。
- 抽樣分佈：樣本比例。

平均數 vs. 比例

- 對於數值資料，我們有樣本平均數。
  - 我們已經知道樣本平均數的抽樣分佈了。
- 對於類別資料，並沒有樣本平均數的概念。
  - 它們有樣本比例（sample proportion）的概念。

母體比例

- 如何知道臺大男生和女生的比例呢？
- 首先，我們先為學生們編碼，女生為 0、男生為 1。
- 對學生 i，$i = 1, ..., N$，令 $X_i \in \{0, 1\}$ 為學生的性別。
- 男生的母體比例 (population proportion) 被定義為

$$p = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- 女生的母體比例為 $1 - p$。

樣本比例

- 令 $\{X_i\}_{i=1,\ldots,N}$ 為母體。
- 令 $\{X_i\}_{i=1,\ldots,n}$ 為樣本數為 n 的樣本。
  - 假設對於所有 $i \neq j$，X_i 與 X_j 彼此獨立。
  - 即 n 個隨機挑選的學生。
- 接著樣本比例（sample proportion）被定義為

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- 母體比例 p 是確定的（儘管未知），而樣本比例 $\hat{p}$ 則是隨機的。
- 我們對於 $\hat{p}$ 的分佈感興趣。
  - 這就是樣本比例的抽樣分佈。

Bernoulli 隨機變數

- 假設隨機變數 X 的樣本空間為 $\{0, 1\}$，亦即它是個二元變數。
- 令 $p = \Pr(X = 1)$ 為「成功機率」。
- 我們說 X 服從一個 Bernoulli 分佈，其成功機率為 p。
  - 用 $X \sim \text{Ber}(p)$ 表示。
- 我們可以計算它的期望值：

$$\mu = p \times 1 + (1 - p) \times 0 = p$$

- 我們可以計算它的變異數和標準差：

$$\sigma^2 = p(1 - p)^2 + (1 - p)(0 - p)^2 = p(1 - p) \text{ 和}$$
$$\sigma = \sqrt{p(1 - p)}$$

樣本比例的分佈

- 什麼是樣本比例

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  的分佈？
- 樣本比例的母體（一次事件的結果）當然不可能是常態分佈。
- 然而樣本比例是一種特殊的樣本平均！
- 我們可以應用中央極限定理。
  - 如果 $n \geq 30$，樣本比例會近似常態分佈。
  - 它的平均和標準差為

$$\mu_{\hat{p}} = \mu = p \quad \text{and} \quad \sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

  - 注意雖然母體是類別資料，但是樣本比例是數值資料。

樣本比例：例子

- 2011 年時，臺大有 19756 個男生及 13324 個女生。
- 男生的母體比例為

$$p = \frac{19756}{33080} \approx 0.597$$

- 讓我們抽 100 位學生並找出它的樣本比例 $\hat{p}$。
  - $\hat{p}$ 的分佈是什麼呢？
  - 抽到男生少於女生的機率是多少呢？

樣本比例：例子

- $\hat{p}$ 的分佈是什麼呢？
  - 因為 $n \geq 30$，$\hat{p}$ 會服從常態分佈。
  - 它的平均為 $p \approx 0.597$。
  - 它的標準差為 $\sqrt{\frac{p(1-p)}{n}} \approx 0.049$。
- $\hat{p} < 0.5$ 的機率為

$$\Pr(\hat{p} < 0.5) \approx 0.024$$

- 小結：
  - 樣本比例「是」類別資料的樣本平均，是數值資料。
  - 其平均數和標準差可根據 Bernoulli 分佈計算而得。
  - 感謝中央極限定理，當樣本數足夠大時，它是常態的。

給工程師的統計學與資料分析 123

第二單元：假設檢定

孔令傑

國立臺灣大學資訊管理學系

2017 年 9 月 2 日

課程大綱

- 基本概念。
- 拒絕規則。
- p-value。
- 母體比例。
- t 檢定。

假說檢定

- 科學家（物理學家、化學家等）是怎麼做研究的呢？
  - 觀察現象。
  - 建立假說。
  - 利用實驗（或其他方式）測試假說。
  - 對於假說做結論。
- 社會科學家和商業研究者也同樣進行假設檢定（hypothesis testing）。
  - 最重要的技術之一就是統計推論：以統計的方式證明事情。
  - 根據抽樣分佈。

人們問的問題

- 在商業（或社會科學）界，人們會問問題：
  - 老員工是否對公司比較有忠誠？
  - 新聘的 CEO 是否將強我們的獲利能力？
  - 是否有某個候選人有超過 50% 選民的偏好支持？
  - 青少年是否比成年人較常吃速食？
  - 我們產品的品質是否足夠穩定？
- 我們該怎麼回答這些問題呢？
- 統計學家建議：
  - 首先先建立個假設。
  - 接著以隨機樣本和統計方法進行檢定。

統計假設

- 統計假設（statistical hypothesis）是一個正式的假設陳述，通常是個欲檢定參數的數學描述。
- 它包含兩個部分：
  - 虛無假設（null hypothesis，寫作 H_0）。
  - 對立假設（alternative hypothesis，寫作 H_a 或 H_1）。
- 對立假設是：
  - 我們想要（需要）證明的東西。
  - 唯有擁有很強的證據，我們才下結論說對立假設成立。
- 虛無假設則對應到一個預設立場（default position）。
  - 我們會先假設（假裝、想像、相信...）虛無假設是對的。
  - 接著我們收集（隨機）樣本資料。
  - 如果在虛無假設成立的前提下，我們極不可能看到我們實際從樣本觀察的結果，我們就說虛無假設是錯的（對立假設是對的）。

統計假設：例子一

- 在我們的工廠裡，我們生產糖果，每袋糖果的平均重量應為 1 公斤。
- 有一天，一個客人告訴我們，他那袋只重 900 公克。
- 我們需要知道那是否只是突發事件，還是我們的生產系統出了問題。
- 如果（我們相信）是系統出了問題，我們就需要將機器關機，並花兩天的時間進行檢查和維修。這至少會花我們 $100,000 元。
- 因此我們不應該只因為一個抱怨而相信我們的系統出了問題。我們該怎麼做呢？

統計假設：例子一

- 首先，我們先建立假設：「我們的生產系統一切正常」。
- 接著我們問：是否有足夠強的證據顯示這個假設是錯誤的？
  - 我們先假設我們的系統一切正常。
  - 然後我們進行問卷調查，看我們是否有足夠的證據。
  - 唯有我們可以「證明」系統確實出了問題，我們才會關閉機器。
- 令 μ 為平均重量，我們的統計假設是

$$\mathrm{H}_0 : \mu = 1$$

$$\mathrm{H}_\mathrm{a} : \mu \neq 1 \text{。}$$

統計假設：例子二

- 我們的社會採用「無罪推定原則」：被判定有罪前，每個人都無罪。
- 所以當有一個人可能偷了些錢，我們可能犯兩種錯誤：
  - 這人有罪，但我們認為他/她無罪。
  - 這人無罪，但我們認為他/她有罪。
- 哪一種比較嚴重？
  - 將一個無罪的判為有罪是不能被接受的。
  - 只有在有很強的證據支持下，我們才會說一個人有罪。
- 所以我們的統計假設是

$$\mathrm{H}_0: \text{這個人是無罪的}$$

$$\mathrm{H}_\mathrm{a}: \text{這個人是有罪的。}$$

統計假設：例子三

- 考慮以下假設：「這個候選人有超過 50% 選民的支持。」
- 我們需要一個預設立場，而我們在乎的百分比為 50%，因此我們選擇的虛無假設為

$$\mathrm{H}_0 : p = 0.5 。$$

  - p 是偏好支持該候選人的選民母體比例。
  - 更精確而言，令 $X_i = 1$ 如果該選民 i 偏好支持這個候選人，否則以 0 表示，$i = 1, ..., N$，那麼 $p = \frac{\sum_{i=1}^{N} X_i}{N}$。
- 那對立假設呢？是

$$\mathrm{H}_\mathrm{a} : p > 0.5 \quad 還是 \quad \mathrm{H}_\mathrm{a} : p < 0.5 ？$$

統計假設：例子三

- 對立假設的選擇取決於要進行的決策或行動。
- 假設一個人只有在相信自己會贏的時候（即 $p > 0.5$）才會參選，那麼對立假設為

$$\mathrm{H_a}: p > 0.5。$$

- 假設一個人傾向參選，並只有在獲勝機率低時才會退出，則對立假設為

$$\mathrm{H_a}: p < 0.5。$$

- 對立假設是「我們想要（需要）證明的事」。

兩種誤差

- 型一誤差（type-I error、false positive）：拒絕其實是事實的虛無假設。
  - 沒有任何東西，但我們卻說有。
- 型二誤差（type-II error、false negative）：沒有拒絕一個錯誤的虛無假設。
  - 有東西，但我們卻沒看到。

(http://9gag.com/gag/aRVbMvy/false-positive-false-negative-in-a-nutshell)

控制犯錯機率

- 我們想要控制犯那些錯誤的機會。
  - 不幸地，我們沒有辦法同時控制兩者。
  - 我們選擇控制型一錯誤的機率。
  - 除非有夠充分的理由，否則我們就相信我們的預設立場。
- 要建立一個統計假設：
  - 把我們的預設立場放在虛無假設。
  - 把我們想要證明的事情（需要強而有力證據的事情）放在對立假設。
- 以數學式型態呈現時：
  - 等於符號（$=$）永遠是放在虛無假設。[1]
  - 對立假設包含一個不等號或是嚴格不等式：$\neq$、$>$ 或 $<$。
  - 當對立假設是一個不等式時，其方向取決於後續的行動或決策。

[1] 有些學者喜歡用 $\geq$ 和 $\neq$。無論如何，概念和計算大同小異。

單尾檢定和雙尾檢定

- 如果對立假設是含有 $\neq$，它便是個雙尾檢定 (two-tailed test)。
- 如果對立假設是含有 $>$ 或 $<$，它便是個單尾檢定 (one-tailed test)。
- 假設我們想要對母體平均數做檢定。
  - 在雙尾檢定，我們檢定母體平均數是否和假設值有顯著差異，但我們不在乎是比較高還是比較低。
  - 在單尾檢定，我們有方向性地檢定母體平均和假設職是否有顯著差異。

課程大綱

- 基本概念。
- 拒絕規則。
- p-value。
- 母體比例。
- t 檢定。

第一個例子：雙尾檢定

- 讓我們來對我們商品的平均重量（公克）進行檢定吧。

$$\mathrm{H}_0 : \mu = 1000$$
$$\mathrm{H}_\mathrm{a} : \mu \neq 1000。$$

- 先假設我們知道產品重量的變異數為 $\sigma^2 = 40000\ \mathrm{g}^2$。
  - 未知 σ^2 的狀況會在之後被討論。
- 讓我們做一次隨機抽樣。
  - 假設樣本大小 $n = 100$。
  - 假設樣本平均 $\overline{X} = 963$。
- 該如何下結論呢？

控制誤差機率

- 我們所做的就是收集一個隨機樣本，並根據觀測到的樣本下結論。
- 很自然地，當我們宣稱 $\mu \neq 1000$，我們可能是錯的。
- 我們想要控制誤差機率。
  - 令 α 為我們犯這個錯的最大機率。
  - α 被稱為顯著水準 (significance level)。
  - $1 - \alpha$ 被稱為信心水準 (confidence level)。
  - 如果 $\mu = 1000$，則最多只有 α 的機率，我們的抽樣和檢定流程會使我們宣稱 $\mu \neq 1000$。

拒絕規則

- 直觀上，如果 $\overline{X}$ 與 1000 差距很大，我們應該拒絕虛無假設，並相信 $\mu \neq 1000$。
  - 因為如果 $\mu = 1000$，就很不可能觀測到那樣大的差距。
  - 所以那麼大的差距提供了很強的證據。
- 我們想要建構一個拒絕規則 (rejection rule)：找一個距離 d，如果 $|\overline{X} - 1000| > d$，我們就拒絕 H_0。
  - 顯然 d 的大小跟 α 有關：α 愈小則 d 愈大。
  - 讓我們把 α 設成 0.05。

拒絕規則

- 我們想要一個距離 d 使得若 H_0 為真，拒絕 H_0 的機率最多 5%，即

$$\Pr\left(|\overline{X} - 1000| \middle| \mu = 1000| > d\right) \leq 0.05。$$

- 滿足以上不等式的所有 d 之中最小的必須滿足

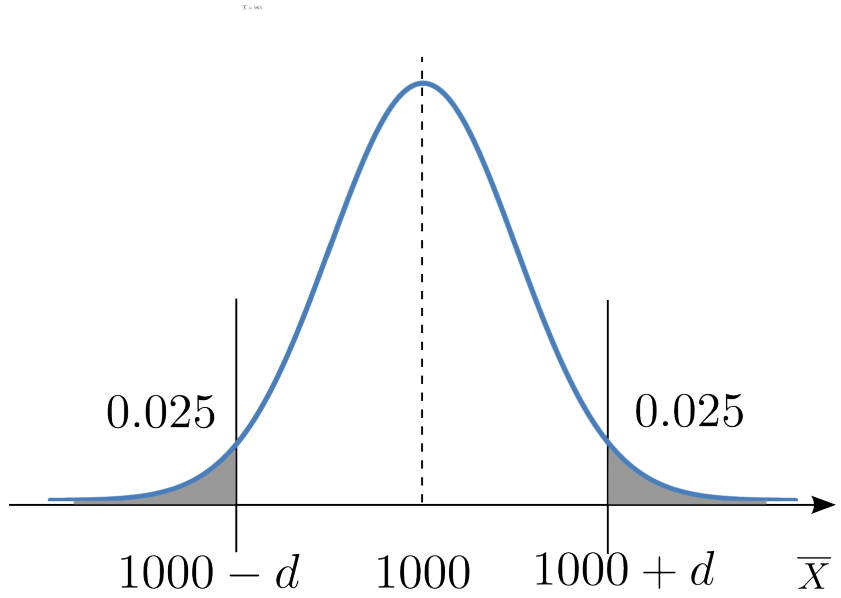
$$\Pr(|\overline{X} - 1000| > d \middle| \mu = 1000) = 0.05。$$

拒絕規則

- 考慮 $\overline{X}$：
  - 我們知道 $\sigma = 200$ 且 $n = 100$。
  - 我們假設 $\mu = 1000$。
  - 感謝中央極限定理，

$$\overline{X} \sim \text{ND}(1000, 20)。$$

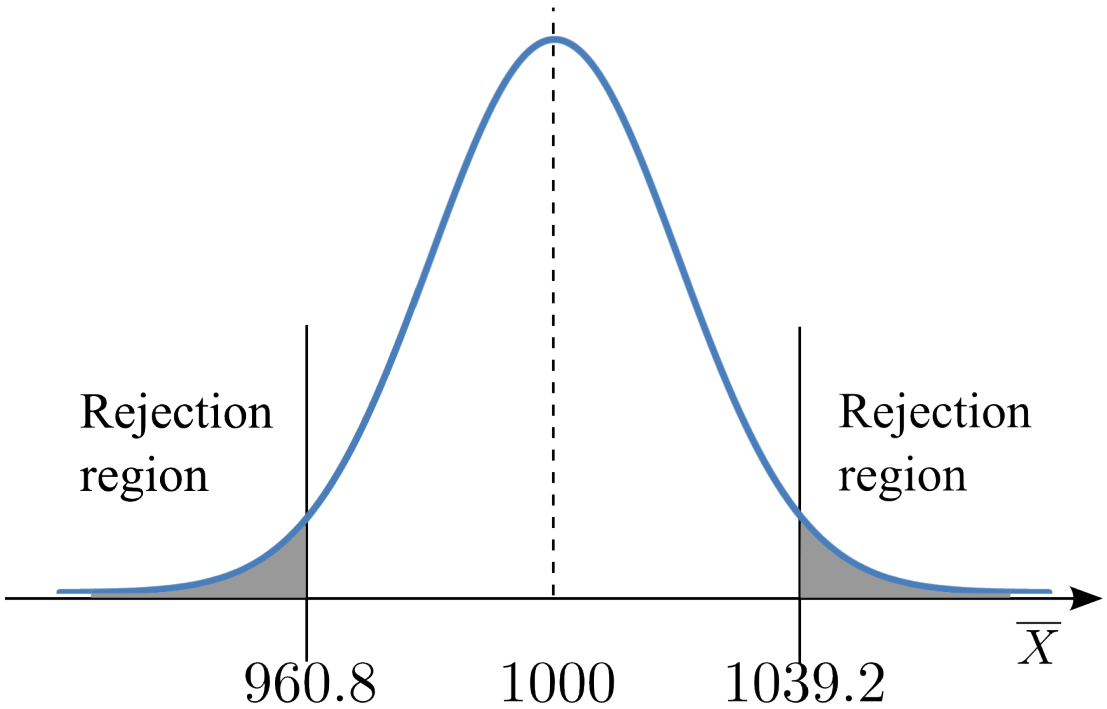
- 現在我們會找 d 去滿足 $\Pr(|\overline{X} - 1000| > d) = 0.05$ 了。



$$\Pr(|\overline{X} - 1000| > d) = 0.05$$

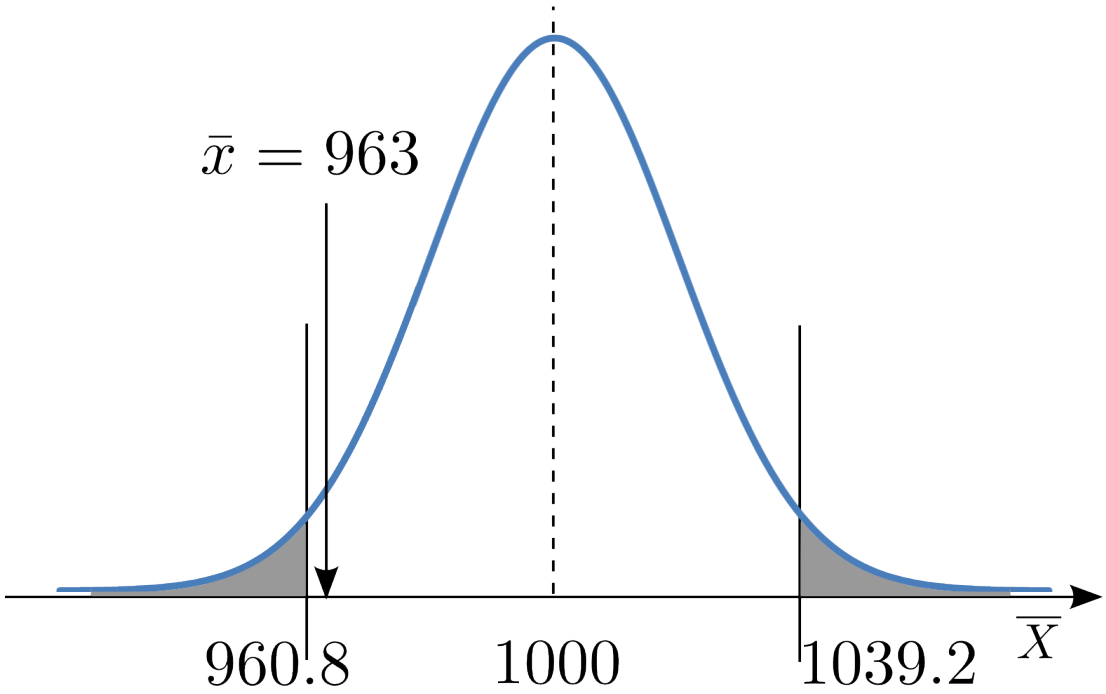
拒絕規則：臨界值

- 根據 $\overline{X} \sim \mathrm{ND}(1000, 20)$，$\Pr(|\overline{X} - 1000| > 39.2) = 0.05$。拒絕區域為 $R = (-\infty, 960.8) \cup (1039.2, \infty)$。
- 如果 $\overline{X}$ 落在拒絕區域，我們拒絕 H_0。



拒絕規則：臨界值

- 因為 $\bar{x} = 963 \notin R$，我們無法拒絕 H_0。
  - 與 1000 的差距不夠大。
  - 這個證據不夠強而有力。



拒絕規則：臨界值

- 在這個例子裡，960.8 和 1039.2 這兩個值是拒絕區的臨界值。
  - 如果樣本平均數超過任一臨界值，我們便拒絕 H_0。
  - 否則，我們不會拒絕 H_0。
- $\bar{x} = 963$ 不夠強來支持 H_a: $\mu \neq 1000$。
- 結論：
  - 因為樣本平均數沒有落在拒絕區，我們不拒絕 H_0。
  - 在 95% 信心水準下，沒有足夠強的證據顯示平均重量不是1000 公克。
  - 因此，我們不應該關閉機器來進行檢查。

小結

- 我們想要知道機器是否出了問題。
  - 如果機器是好的，我們不想要得到一個會使我們得進行檢查和維修的結論。
  - 只有當我們有足夠強的證據顯示 $\mu \neq 1000$，我們才會檢查。
- 我們想知道 H_0 是否是假的，即 $\mu \neq 1000$。
- 我們控制下錯誤結論的機率。
  - 我們控制型一錯誤：在 H_0 為真時不應該拒絕它。
  - 我們限制型一錯誤的機率在 $\alpha = 5\%$ 以下。
- 如果 $\overline{X}$ 落在拒絕區，我們會宣稱 H_0 是錯的。
  - 臨界值的計算是基於常態分佈，它可以被轉換成標準常態分佈 (z 分佈)。
  - 上述方法稱為 z 檢定。

不拒絕 vs. 接受

- 我們應該小心地寫我們的結論：
  - 錯誤寫法：因為樣本平均不在拒絕區域，我們接受 H_0。在 95% 信心水準下，有足夠強的證據顯示平均重量是1000 公克。
  - 正確: 因為樣本平均不落在拒絕區域，我們無法拒絕 H_0。在 95% 信心水準下，沒有足夠強的證據顯示平均重量不是1000 公克。
- 沒有辦法證明一件事是錯的，不代表它就是真的！

第一個例子（第二部分）

- 假設我們修正假設為有向的：

$$\mathrm{H}_0 : \mu = 1000$$
$$\mathrm{H}_\mathrm{a} : \mu < 1000。$$

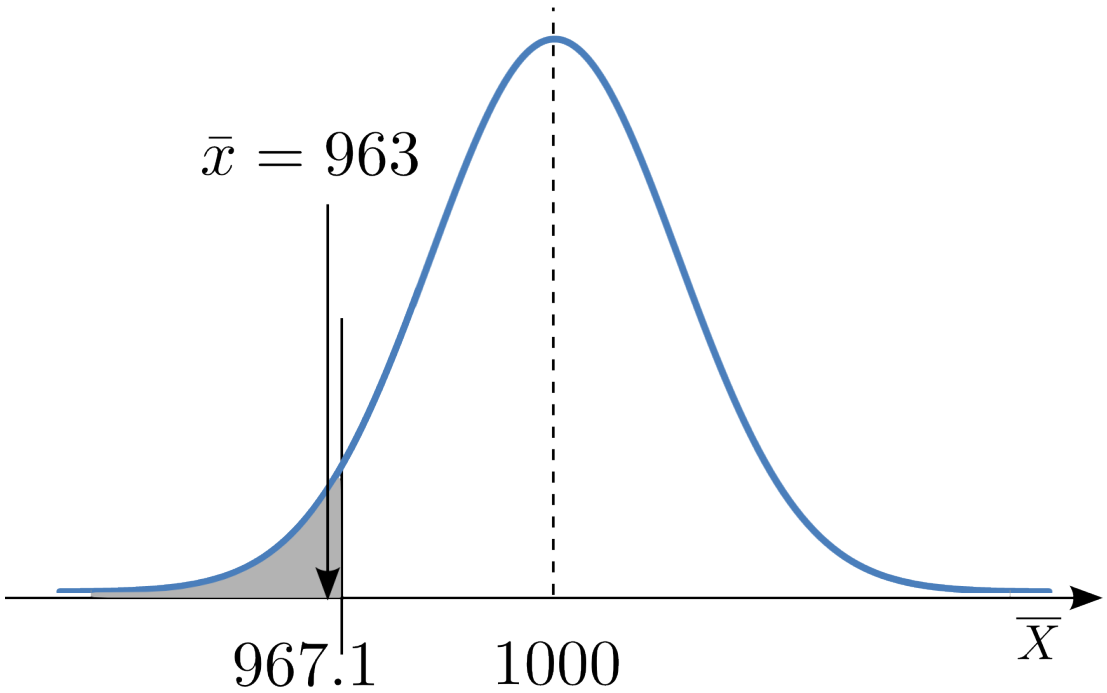
我們仍有 $\sigma^2 = 40000$，$n = 100$ 及 $\alpha = 0.05$。

- 這是一個單尾檢定。
- 當我們有很強的證據支持 H_a，我們會下結論說 $\mu < 1000$。
- 我們需要找一個距離 d 使得

$$\Pr\left(1000 - \overline{X} > d \middle| \mu = 1000\right) = 0.05。$$

拒絕規則：臨界值

- $d = 32.9$ 滿足 $0.05 = \Pr(1000 - \overline{X} > d)$。
- 當觀測樣本平均 $\bar{x} = 963 \in (-\infty, 967.1)$，我們拒絕 H_0。
  - 與 1000 的差距足夠大；這個證據夠強。

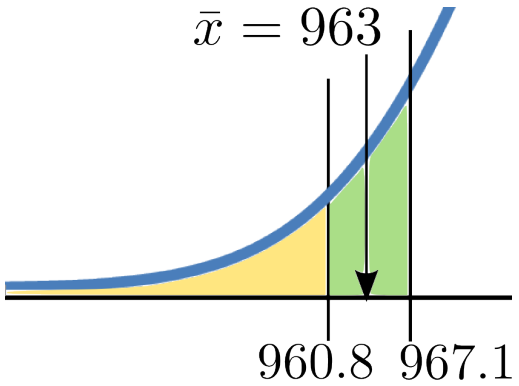


拒絕規則：臨界值

- 在這個例子，967.1 是拒絕的臨界值。
  - 如果樣本平均數 (在這個例子裡) 低於臨界值，我們便拒絕 H_0。
  - 否則，我們不拒絕 H_0。
- 有很強的證據支持 H_a: $\mu < 1000$。
- 結論：
  - 因為樣本平均數落在拒絕區，我們拒絕 H_0。在 95% 信心水準下，有足夠強的證據顯示平均重量少於1000 公克。

單尾檢定 vs. 雙尾檢定

- 什麼時候我們使用雙尾檢定呢？
  - 當我們沒有方向性資訊時，我們使用雙尾檢定。
  - 例：我們懷疑母體平均數改變了，但我們不曉得它到底變小或是變大。
- 如果我們知道或相信這個改變在某個方向，我們可以使用單尾檢定。
- 擁有更多資訊（知道改變的方向）使拒絕變得「更簡單」，即更容易找到足夠強的證據。



小結

- 區別以下各個成對的概念：
  - 單尾檢定 vs. 雙尾檢定。
  - 沒有證據顯示 H_0 是錯的 vs. 有證據顯示 H_0 是對的。
  - 不拒絕 H_0 vs. 接受 H_0。
  - 在虛無假設中使用 $=$ vs. 在虛無假設中使用 $\geq$ 或 $\leq$。

課程大綱

- 基本概念。
- 拒絕規則。
- p-value。
- 母體比例。
- t 檢定。

p-value

- p-value 是假設檢定裡一個重要的、富有意義的且被廣泛使用的工具。

定義 1

在統計檢定裡，對於一個觀測到的統計量，p-value 是在虛無假設成立的情況下，觀測到比此觀測值更極端的結果的機率。

- 計算是基於觀測到的統計量。
- 是觀測值的尾端機率 (tail probability)。
- 假設虛無假設為真。

p-value

- 數學上的意思：
  - 考慮對母體平均數 μ 進行單尾檢定

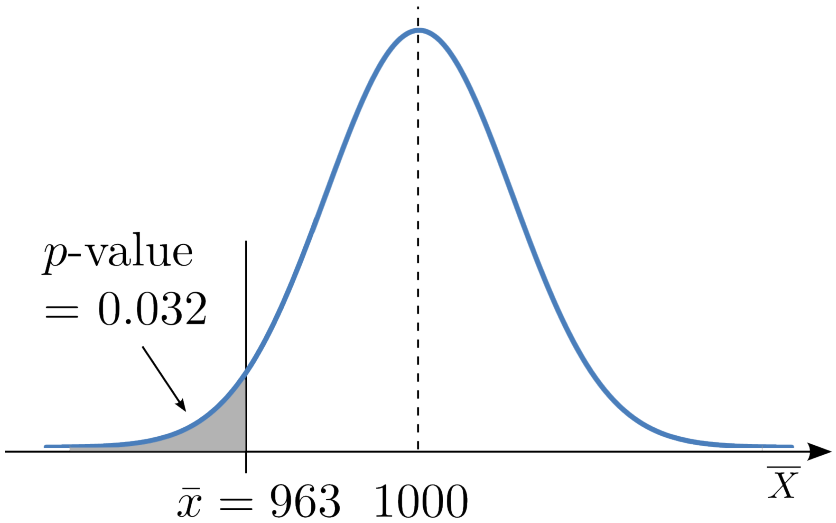
$$\mathrm{H}_0 : \mu = 1000$$

$$\mathrm{H}_\mathrm{a} : \mu < 1000。$$

  - 給定觀測到的 $\bar{x}$，p-value 照定義是

$$\Pr(\overline{X} \le \bar{x})。$$

- 在之前的例子，$\sigma = 200$，$n = 100$，$\alpha = 0.05$ 及 $\bar{x} = 963$。
  - 如果 H_0 為真，即 $\mu = 1000$，我們得到 $\Pr(\overline{X} \le 963) = 0.032$。
  - $\bar{x}$ 的 p-value 為 0.032。



如何使用 p-value 呢？

- p-value 可以用來建構拒絕規則。
- 對於單尾檢定：
  - 如果 p-value 小於 α，我們便拒絕 H_0。
  - 如果 p-value 大於 α，我們就不拒絕 H_0。
- 在我們的例子裡，統計假設是

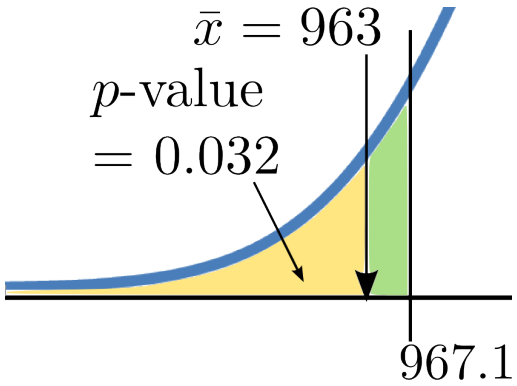
$$\mathrm{H}_0 : \mu = 1000$$

$$\mathrm{H}_\mathrm{a} : \mu < 1000。$$

  - $\alpha = 0.05$。
  - 因為 p-value $0.032 < 0.05$，我們拒絕 H_0。

p-value vs. 臨界值

- 使用 p-value 等同於使用臨界值。
  - 兩個方法在拒絕與否會得到一樣的結論。



使用 p-value 的好處

- 在很多的研究中，研究者在進行檢定之前，不會決定顯著水準 α。
- 他們計算 p-value，然後以星號標記結果的顯著性 。
- 一個典型給予星號的方式：

p-value	顯著	標記
$(0, 0.01]$	高度顯著	***
$(0.01, 0.05]$	中等顯著	**
$(0.05, 0.1]$	輕微顯著	*
$(0.1, 1)$	不顯著	(Empty)

p-value 的大小

- 假設我們想討論不同年齡層的人是否平均每天至少睡八小時。
  - 年齡層：$[10, 15)$、$[15, 20)$、$[20, 35)$ 與其他。
  - 對於小組 i，實行單尾檢定。$\mathrm{H_a}: \mu_i > 8$。結果可以被以表格呈現：

小組	年齡組	p-value
1	[10,15)	0.0002***
2	[15,20)	0.2
3	[20,25)	0.06*
4	[25,30)	0.04**
5	[30,35)	0.03**

- 小的 p-value 不代表較大的差距！
  - 我們沒有辦法做出 $\mu_5 > \mu_4$，$\mu_1 > \mu_3$ 這些結論。
  - 要瞭解兩個母體平均間的差異，應使用其他的檢定。

p-value 和雙尾檢定

- 如何建構出雙尾檢定的拒絕規則呢？
  - 如果 p-value 小於 $\frac{\alpha}{2}$，我們拒絕 H_0。
  - 如果 p-value 大於 $\frac{\alpha}{2}$，我們不拒絕 H_0。
- 考慮雙尾檢定

$$\mathrm{H}_0 : \mu = 1000$$

$$\mathrm{H}_\mathrm{a} : \mu \neq 1000。$$

  - 我們有 $\alpha = 0.05$。
  - 因為 p-value $= 0.032 > \frac{\alpha}{2} = 0.025$，我們不拒絕 H_0。
- 有些研究者/書/軟體使用其他定義：
  - 雙尾檢定的 p-value 是其對應的單尾檢定 p-value 的兩倍。
  - 然後再將這個 p-value 與 α 比較。

小結

- p-value 是在虛無假設成立的狀況下，基於統計量觀測值的尾端機率。
- p-value 方法是一個建構拒絕規則的方法。
  - 它等同於臨界值方法。
- 有統計顯著性，不表示有實務顯著性。
  - p-value 很小，只表示有顯著差異，不表示有很大的顯著差異。
  - p-value 並不衡量差距的大小。

課程大綱

- 基本概念。
- 拒絕規則。
- p-value。
- 母體比例。
- t 檢定。

檢定母體比例

- 在很多情況下，我們需要檢定母體比例。
  - 生產系統的缺陷率和收益率。
  - 支持一個候選人或政策的人民比例。
  - 瀏覽產品頁面後真的購買的比例（轉化率）。
- 如何檢定母體比例呢？
- 假設我們想要檢定男性使用者的比例：
  - 讓我們先標記男性使用者為 1，非男性使用者為 0。
  - 母體比例 $p = \frac{\sum_{i=1}^{N} X_i}{N}$ 就是個母體平均數。
  - 一個樣本比例 $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ 是樣本平均數。
  - 因為母體顯然不常態，因此不能用 t 檢定。
  - 因為可以由 p 計算 σ 為 $\sqrt{p(1-p)}$，我們用 z 檢定來檢定母體比例。
  - 限制：$n \geq 30$，$n\hat{p} \geq 5$ 及 $n(1-\hat{p}) \geq 5$。

假設

- 母體比例是 p。
- 若想知道母體比例是否為 p_0，雙尾檢定是

$$H_0 : p = p_0$$
$$H_a : p \neq p_0 \text{。}$$

- 在一個單尾檢定中，對立假設可以是

$$H_a : p > p_0 \quad \text{或} \quad H_a : p < p_0 \text{。}$$

例子

- 在一座工廠裡，我們產品的缺陷率似乎太高了。理想上，它應該少於 1%，但是有些工人認為是高過 1% 的。
- 如果缺陷率高過 1%，我們就應該修理機器，反之就不要[2]。
- 令 p 為缺陷率，假設為

$$\mathrm{H}_0 : p = 0.01$$

$$\mathrm{H}_\mathrm{a} : p > 0.01。$$

[2]什麼時候使用 $\mathrm{H}_\mathrm{a} : p < 0.01$ 呢？

例子

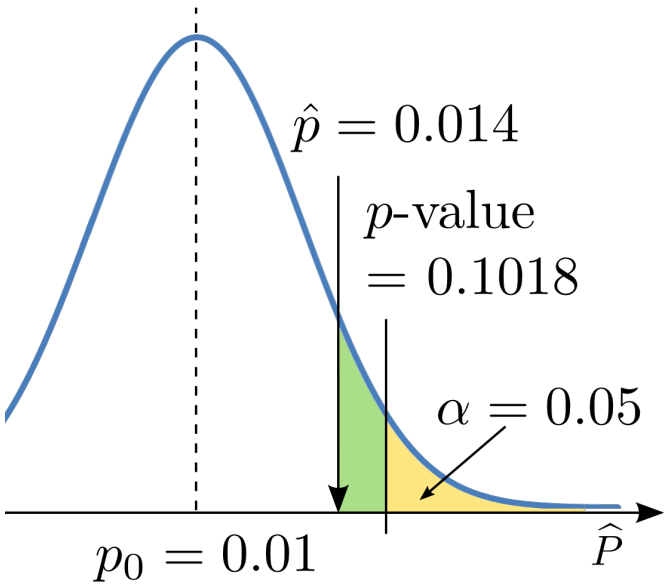
- 在幾批隨機生產後，我們發現 1000 個生產出來的東西，有 14 個是缺陷品。
  - 觀測樣本比例 $\hat{p} = 0.014$。
  - 全部的限制都滿足；$n = 1000$，$n\hat{p} = 14$ 及 $n(1 - \hat{p}) = 986$。
- 假設顯著水準設在 $\alpha = 0.05$，我們的結論是什麼呢？

例子：計算與解讀

- 計算與結論：
  - 對於這個單尾檢定，因為

$$p\text{-value} = \Pr(\hat{p} > 0.014 | p = 0.01)$$
$$= 0.1018 > 0.05 = \alpha$$

    我們不拒絕 H_0。
  - 沒有足夠強的證據證明損壞率高於 1%。
- 決策：
  - 我們不應該試著修理機器。



課程大綱

- 基本概念。
- 拒絕規則。
- p-value。
- 母體比例。
- t 檢定。

z 檢定

- 在例子一，基本上我們是用 $\overline{X} \sim \mathrm{ND}(\mu, \frac{\sigma}{\sqrt{n}})$ 這件事實。
  - 這隱含了 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{ND}(0,1)$，也就是所謂的標準常態分佈，或是 z 分佈。
  - 因此，這個檢定被稱為 z 檢定。
- 這需要知道 σ。

當變異數未知

- 當母體變異數 σ^2 為未知，$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 的大小也就未知。
- 如果我們用樣本變異數 S^2 作為替代呢？

定理 1

對於一個常態的母體，統計量

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

服從 t 分佈，且自由度為 $n-1$。

- 什麼是 t 分佈？

t 分佈

- t 分佈被定義為以下：

定義 2

若一個隨機變數 X 服從自由度為 n 的 t 分佈，則其 pdf 為

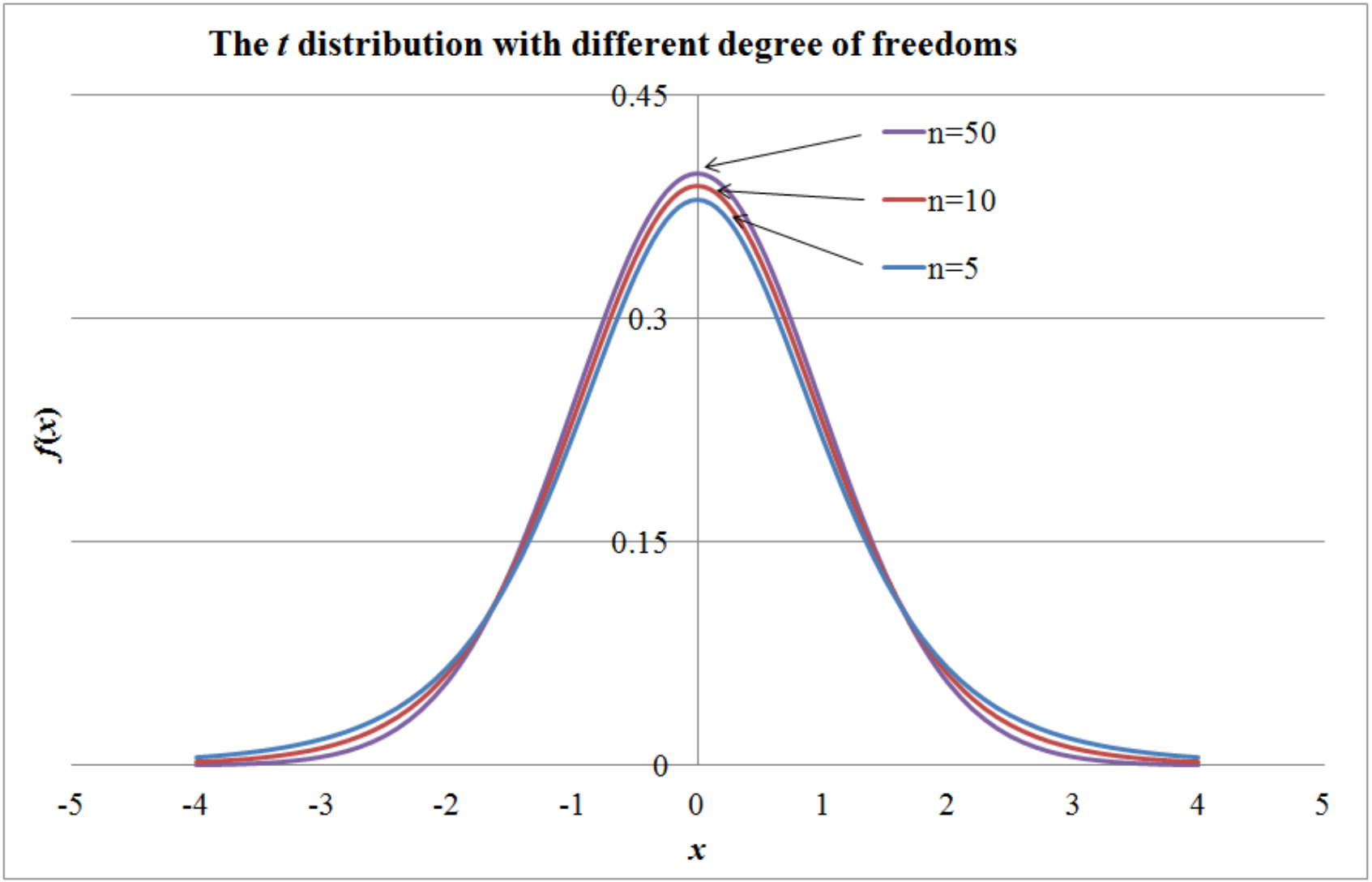
$$f(x|n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

對於所有的 $x \in (-\infty, \infty)$。我們用 $X \sim t(n)$ 表示。

- $\Gamma(x) = \int_0^\infty z^{x-1}e^{-z}dz$ 是個 gamma 函數。

z 和 t 分佈

- 讓我們來比較 $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 和 $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$。
  - 因為我們不知道 σ，我們用 S 來替代。
  - $Z \sim \mathrm{ND}(0,1)$ 且 $T \sim t(n-1)$。
  - 因為 t 是 z 分佈的替代品，它也被設計為以 0 為中心：$\mathbb{E}[T] = \mathbb{E}[Z] = 0$。
  - 但是，因為我們多加了一個隨機變數入算式 (σ 是個已知的常數)，T 會變得比 Z「更隨機」，即 $\mathrm{Var}(T) > \mathrm{Var}(Z)$。
  - 圖形上，t 曲線會比 z 曲線更平。
  - 當 $n \to \infty$，$t(n) \to \mathrm{ND}(0,1)$。


The t distribution with different degree of freedoms
0.45
0.3
0.15
0
f(x)
n=50
n=10
n=5
-5
-4
-3
-2
-1
0
1
2
3
4
5
x

t 檢定

- 針對母體變異數未知的常態母體，我們通常使用 t 檢定 去檢定母體平均數。
  - 如果樣本數很大，也可以使用 z 分佈，並以 s 替代 σ。

例子

- 某個 MBA 很少錄取工作經驗不長於兩年的申請者。
- 為了去檢定是否被錄取者的平均工作年限高於兩年，我們隨機挑選了 20 個被錄取的申請者。
- 我們記錄他們在進入 MBA 之前的工作經驗。
  - 在進入 MBA 前，他們平均工作經驗為 2.5 年。這是個樣本平均。
  - 樣本標準差為 1.3765 年。
- 母體為是常態分佈。
- 信心水準被設在 95%。

例子：假設

- 假設問這個問題的人是個有一年工作經驗的申請者。他是個悲觀主義者：只有在平均工作經驗被證實少於兩年才會申請 MBA。
- 假設是

$$\mathrm{H}_0 : \mu = 2$$

$$\mathrm{H}_\mathrm{a} : \mu < 2 \text{。}$$

  - μ 是全部錄取的申請者在進入 MBA 之前的平均工作經驗（年）。
  - 為了鼓勵他，我們想找一個足夠強的證據顯示機會是高的（$\mu < 2$）。

例子：假設與檢定

- 假設他是個樂觀主義者：只有在被證實平均工作經驗高於兩年時才不會申請 MBA。
- 假設變為

$$\mathrm{H}_0 : \mu = 2$$

$$\mathrm{H}_\mathrm{a} : \mu > 2 \text{。}$$

  - 為了勸退他，我們想找一個很強的證據顯示機會不高 ($\mu > 2$)。
- 讓我們考慮樂觀的申請者 (及 $\mathrm{H}_\mathrm{a} : \mu > 2$)) 先。
- 因為母體變異數未知且母體為常態，我們可以使用 t 檢定。

例子（樂觀）：計算與解讀

- 計算：
  - p-value 是 $\Pr(\overline{X} > 2.5|\mu = 2) = 0.0604$。
- 結論：
  - 對於這個單尾檢定，因為 p-value $> 0.05 = \alpha$，我們不拒絕 H_0。
  - 沒有足夠強的證據顯示平均工作經驗高於兩年。
  - 結果沒有強到可以阻擋這個只有一年工作經驗的申請者。
- 決定：
  - 你這麼樂觀，你就申請吧！

例子（悲觀）

- 假設這個申請者是悲觀的：

$$H_0 : \mu = 2$$
$$H_a : \mu < 2。$$

  - p-value 是 $\Pr(\overline{X} < 2.5 | \mu = 2) = 1 - 0.0604 = 0.9396$。
  - 這是基於 t 分佈的計算結果。.
  - 我們不拒絕 H_0，不能下結論說 $\mu < 2$。沒有足夠強的證據來鼓勵他。
  - 他這麼悲觀，那就別申請。
- 因為我們使用了不同的對立假設，最終決策也因此不相同！
  - 這只會發生在我們都不拒絕 H_0 的時候。

小結

- 為檢定母體平均數 μ：

σ^2	樣本數	母體分佈	
		常態	非常態
已知	$n \geq 30$	z	z
	$n < 30$	z	無母數
未知	$n \geq 30$	t 或 z	z
	$n < 30$	t	無母數

- 更多可以被檢定的母體參數：
  - 母體比例（z 檢定）、母體變異數（χ^2 檢定）。
  - 兩母體平均數的差異（t 檢定）、兩母體變異數的比例（F 檢定）。

給工程師的統計學與資料分析 123

第三單元：迴歸分析 (1)

孔令傑

國立臺灣大學資訊管理學系

2017 年 1 月 14 日

相關性與預測

- 我們經常想要找出變數間的相關性。
- 比如說，如果給定下列 12 間房子的價錢和大小：

房子編號	1	2	3	4	5	6
大小（平方公尺）	75	59	85	65	72	46
價錢（千元）	315	229	355	261	234	216
房子編號	7	8	9	10	11	12
大小（平方公尺）	107	91	75	65	88	59
價錢（千元）	308	306	289	204	265	195

- 我們可以計算其相關係數為 $r = 0.729$.
- 如果有一間房子大小為 100 平方公尺，我們能預測（估計）它的價錢嗎？
  - 價錢感覺跟大小有關，不過該怎麼做？

超過兩個變數間的相關性

- 有時我們有超過兩個變數。
- 比如說，我們可能也知道每間房子有幾個臥房：

房子編號	1	2	3	4	5	6
大小（平方公尺）	75	59	85	65	72	46
價錢（千元）	315	229	355	261	234	216
臥房數	1	1	2	2	2	1
房子編號	7	8	9	10	11	12
大小（平方公尺）	107	91	75	65	88	59
價錢（千元）	308	306	289	204	265	195
臥房數	3	3	2	1	3	1

- 怎麼描述三個變數之間的相關性？
- 給定大小和臥房數，如何預測（估計）價錢？

迴歸分析

- 迴歸分析（regression）是個好工具！
- 做為最被廣為使用的統計方法，迴歸分析可以討論：
  - 哪個變數對某個目標變數有影響：影響房價的是大小、房間數，還是都有？
  - 那個變數如何產生影響：大房子比較貴還是便宜？大一坪貴（便宜）多少？
- 我們將會根據一至多個自變數來解釋、預測或估計一個應變數。
  - 應變數（dependent variable）：我們所關心的目標變數。
  - 自變數（independent variable）：我們所關心的目標變數的潛在影響因子。
  - 自變數又被稱為解釋變數（explanatory variable），而應變數又被稱為回應變數（response variable）.
- 如果我們想要預測明天的來店顧客人數：
  - 應變數：明天的來店顧客人數。
  - 自變數：天氣、是否是假日、有無促銷活動...

迴歸分析的種類

- 根據自變數的個數：
  - 單迴歸 (simple regression)：只有一個自變數。
  - 複迴歸 (multiple regression)：超過一個自變數。
- 根據應變數的資料型態：
  - 在普通迴歸 (ordinary regression) 中，應變數是數值資料。
  - 在羅吉斯迴歸 (logistic regression) 中，應變數是分類資料.
- 還有其他種迴歸模型。

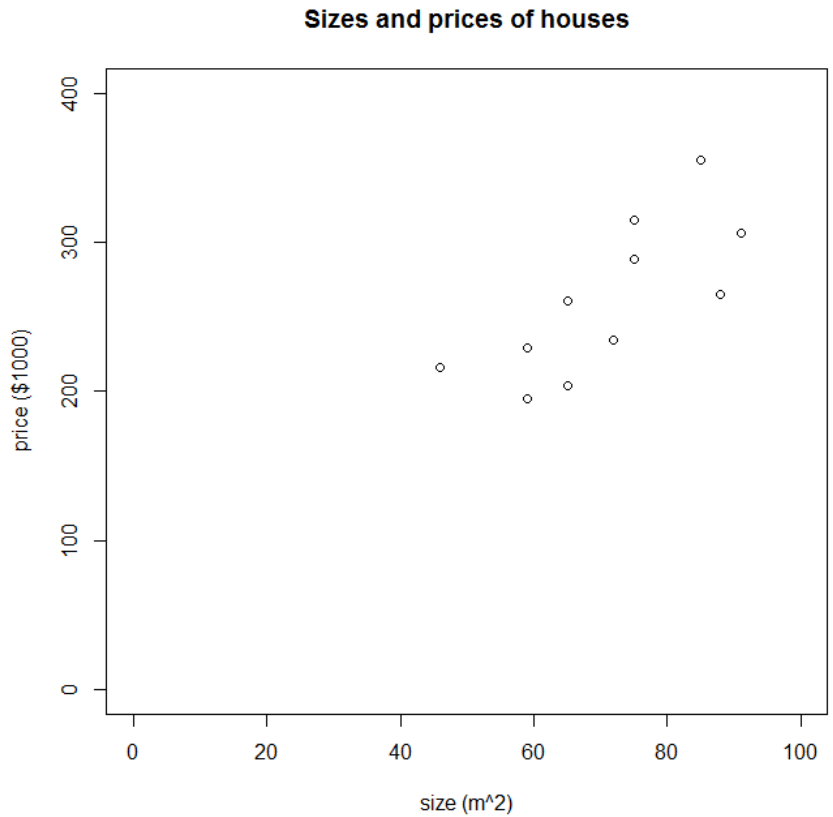
課程大綱

- 基本原理。
- 變數轉換與選擇。
- 一個案例。
- 類別型態自變數。

基本原理

- 令 x_i 跟 y_i 分別是房子 i 的大小跟價格，$i = 1, ..., 12$。

大小（平方公尺）	價錢（千元）
46	216
59	229
59	195
65	261
65	204
72	234
75	315
75	289
85	355
88	265
91	306
107	308



- 如何以找出大小和價格間的關係？

線性估計

- 如果對於所有房子，這兩個變數間的關係是線性的，就表示

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{。}$$

  - β_0 是這個方程式的截距 (intercept)。
  - β_1 是這個方程式的斜率 (slope)。
  - ϵ_i 是用大小估計房價時的常態隨機誤差 (normal random noise)。
- 冥冥之中這個方程式存在，但我們不知道 β_0 跟 β_1 的值。
  - β_0 跟 β_1 是所有房子這個母體的參數.
  - 我們想要用手上有的樣本資料 (也就是那 12 間房子) 去估計 β_0 和 β_1。
  - 我們想要計算出兩個統計量 $\hat{\beta}_0$ 跟 $\hat{\beta}_1$ 去做為我們對 β_0 跟 β_1 的估計值。

線性估計

- 給定我們用樣本資料算出的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$，我們就會用 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 來做為我們對 y_i 的估計值。
- 我們希望我們的估計誤差 (estimation error) $\epsilon_i = y_i - \hat{y}_i$ 愈小愈好。
- 把所有誤差 ϵ_i 集合起來，我們希望總平方誤差 (sum of squared errors，SSE) 愈小愈好：

$$\sum_{i=1}^{n} \epsilon_i^2 = (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big]^2 。$$

- 我們求解 (給定樣本資料後的)

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \Big[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big]^2$$

最小平方估計 (least square approximation) 問題。

最小平方估計

- 最小平方估計問題

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \Big[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big]^2$$

的最佳 $(\hat{\beta}_0, \hat{\beta}_1)$ 是有公式解的：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{和} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}。$$

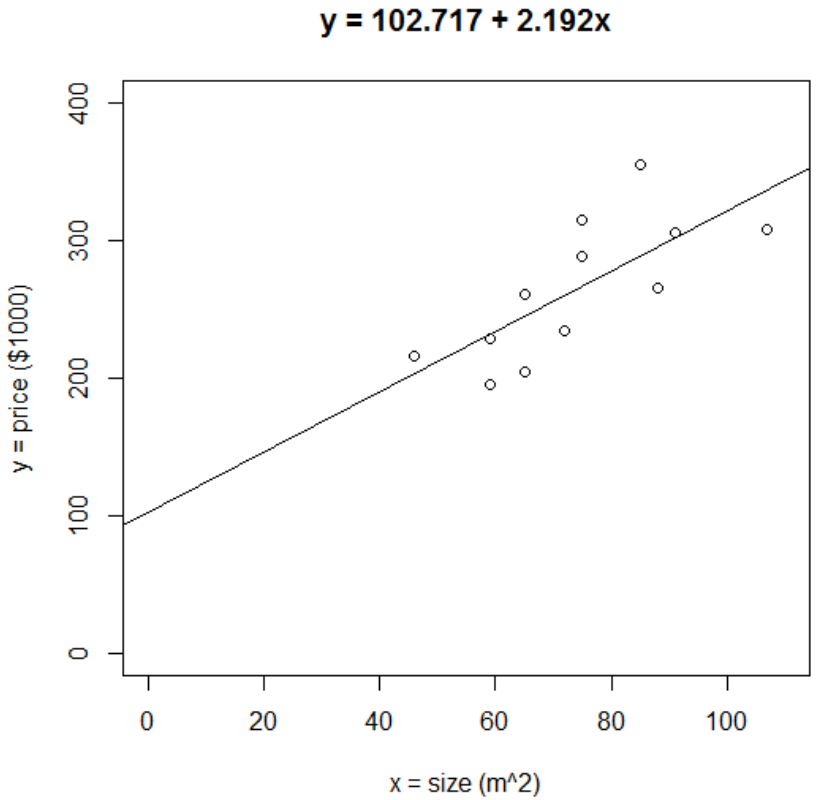
- 根據我們的 12 間房子，我們會得到 $(\hat{\beta}_0, \hat{\beta}_1) = (102.717, 2.192)$.
  - 這組樣本的 SSE 是 13118.63.
  - 我們永遠不知道真正的 β_0 和 β_1。不過，根據我們的樣本資料，我們「最佳的」猜想是 $\beta_0 = 102.717$ 和 $\beta_1 = 2.192$。

模型意涵

- 我們的迴歸模型是

$$y = 102.717 + 2.192x。$$

- 模型意涵：
  - 當房子大小增加 1 平方公尺時，我們預期房價會上升 $2,192。
  - 模型意涵：大小為 70 平方公尺的房子的預期房價為 $256,197。
- （不太好的）模型意涵：大小為 0 平方公尺的房子，我們預期其房價為 $102,717.



複迴歸

- 絕大部分的時候，使用超過一個自變數可以更好地解釋或估計應變數。
- 讓我們來同時用大小和房間數做複迴歸 (multiple regression) :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i \text{ 。}$$

  - y_i 是價格 (千元)。
  - $x_{1,i}$ 是大小 (平方公尺)。
  - $x_{2,i}$ 是房間數。
  - ϵ_i 是隨機誤差。
- 我們的 (最小平方) 估計是 $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (82.737, 2.854, -15.789)$。

價錢 (千元)	大小 (平方公尺)	房間數
315	75	1
229	59	1
355	85	2
261	65	2
234	72	2
216	46	1
308	107	3
306	91	3
289	75	2
204	65	1
265	88	3
195	59	1

模型意涵

- 我們的迴歸模型是

$$y = 82.737 + 2.854x_1 - 15.789x_2 \text{。}$$

  - 當房子變大 1 m^2（而且其他自變數都固定）時，房價預期上升 $2,854。
  - 當房間數加 1（而且其他自變數都固定）時，房價預期下降 $15,789.
- 研究者必須判讀這些意涵是否合理（或對他是否有用）。
  - 房間數可能不是解釋房價的好因子（至少不是以線性的方式）。
- 我們不能光只是計算出係數：
  - 我們需要衡量一個迴歸模型的整體品質。
  - 我們需要比較不同迴歸模型的相對品質。
  - 我們需要檢定迴歸模型中每個係數的顯著性。

模型檢驗：整體品質

- 如何衡量一個迴歸模型 $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_k x_k$ 的品質？
- 如果完全不使用任何自變數，我們會用 $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ 估計 y_i。此時最大平方誤差 (sum of squared total errors，SST) 是 $SST = \sum_{i=1}^n (y_i - \bar{y})^2$。
- 根據我們的迴歸模型，我們把誤差降到

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left[(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 。$$

- 自變數的變異中，能被我們的迴歸模型解釋的比例是

$$0 \leq R^2 = 1 - \frac{SSE}{SST} \leq 1。$$

R^2 愈大，迴歸模型愈好。

計算 R^2

- 每當我們計算出一個迴歸模型的各系數時，我們就能同時算出 R^2。
- 統計軟體都會在報表中呈現 R^2。
- 對於 $y = 102.717 + 2.192x$，我們的 $R^2 = 0.5315$：
  - 大約 53% 的房價變異可以被房子大小解釋。
- 若 (且唯若) 只有一個自變數，則 $R^2 = r^2$，而 r 就是自變數跟應變數的相關係數。
  - $-1 \leq r \leq 1$。
  - $0 \leq r^2 = R^2 \leq 1$。

比較迴歸模型

- 現在我們可以用 R^2 來比較迴歸模型了。
- 以剛剛的例子來說：

自變數	房子大小	房間數	房子大小和房間數
R^2	0.5315	0.29	0.5513

  - 只用房子大小比只用房間數好。
  - 同時用兩個自變數有比較好嗎？
- 事實上，增加自變數一定會提高 R^2！
  - 加了自變數了不起是係數被設為 0，不會讓 R^2 變小。
  - 即使加入毫不相干的自變數，R^2 也會變大。
- 若要進行「公平」的比較並且找出有意義的影響因子，我們必須根據自變數的數量調整 R^2。

調整後的 R^2

- 標準的把 R^2 調整成調整後的 R^2（adjusted R^2）是

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)。$$

  - n 是樣本數，k 是模型中的自變數個數。

- 以剛剛的例子來說：

自變數	房子大小	房間數	房子大小和房間數
R^2	0.5315	0.290	0.5513
R^2_{adj}	0.4846	0.219	0.4516

- 其實只使用自變數是三個模型中最好的！

檢定係數顯著性

- 另一個重要的工作是檢定係數顯著性（significance）。
- 比如說剛剛的雙自變數模型

$$y = 82.737 + 2.854x_1 - 15.789x_2 \text{。}$$

- 2.854 和 -15.789 是完全根據樣本而算出來的。我們永遠不會知道 β_1 和 β_2 是否真的是這兩個值！
- 我們甚至不確定 β_1 和 β_2 是否不是 0. 我們必須檢定它們：

$$\mathrm{H}_0 : \beta_i = 0$$
$$\mathrm{H}_\mathrm{a} : \beta_i \neq 0.$$

- 我們希望有足夠的證據令我們相信 $\beta_i \neq 0$.

檢定係數顯著性

- 檢定的結果在報表中都有。統計軟體 (比如說 R) 告訴我們：

	Coefficients	Standard Error	t Stat	p-value	
Intercept	82.737	59.873	1.382	0.200	
Size	2.854	1.247	2.289	0.048	**
Bedroom	−15.789	25.056	−0.630	0.544	

  - 因為不知道母體變異數，我們使用 t 檢定.
  - 「Coefficients」記錄的是樣本平均數 $\bar{x}$；「Standard Error」記錄的是 $\frac{s}{\sqrt{n}}$；「t Stat」記錄的是 $t = \frac{\bar{x}-0}{s/\sqrt{n}}$。
  - 「p-value」是 t 統計量的雙尾機率 (在大部分統計軟體中)，用來跟 α 比較。

- 別忘了我們假設 ϵ_i 是常態的。

檢定係數顯著性

- 根據統計軟體：

	Coefficients	Standard Error	t Stat	p-value	
Intercept	82.737	59.873	1.382	0.200	
Size	2.854	1.247	2.289	0.048	**
Bedroom	−15.789	25.056	−0.630	0.544	

- 在 95% 的信心水準下：
  - 我們相信 $\beta_1 \neq 0$，亦即房子大小對房價確實有影響。
  - 我們不相信 $\beta_2 \neq 0$，亦即沒有證據顯示房間數對房間有影響。
- 如果只用房子大小當自變數，它的 p-value 會是 0.00714。我們同樣會相信房子大小對房價有影響。

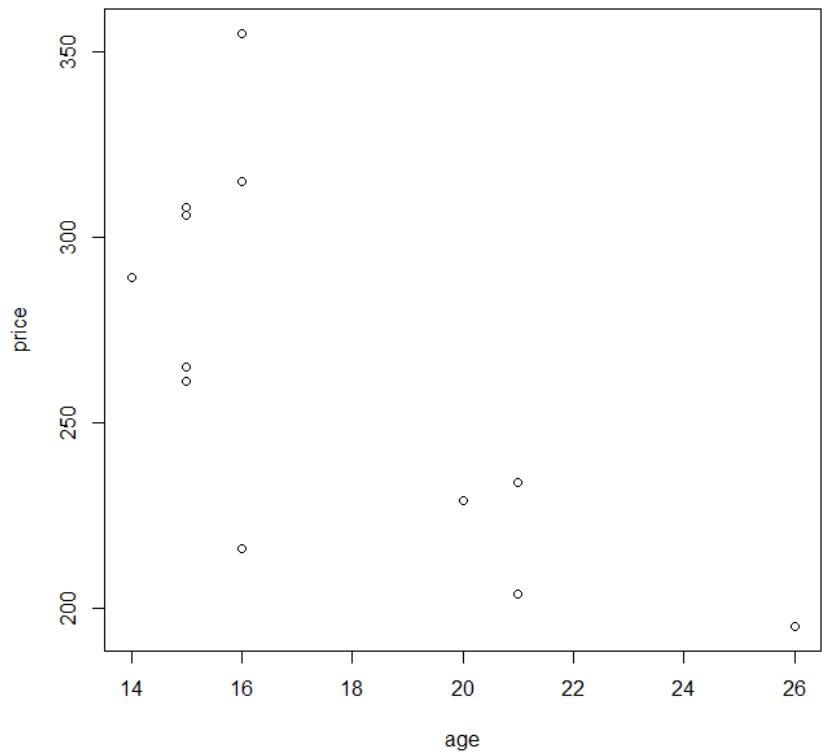
課程大綱

- 基本原理。
- 變數轉換與選擇。
- 一個案例。
- 類別型態自變數。

屋齡

- 屋齡也有可能影響房價。

價格（千元）	大小（平方公尺）	房間數	屋齡（年）
315	75	1	16
229	59	1	20
355	85	2	16
261	65	2	15
234	72	2	21
216	46	1	16
308	107	3	15
306	91	3	15
289	75	2	14
204	65	1	21
265	88	3	15
195	59	1	26



- 別管房間數了，讓我們來試試採用屋齡當自變數。

屋齡

- 對於房子 i，讓 y_i 做為房價、$x_{1,i}$ 做為大小，以及 $x_{3,i}$ 做為屋齡。假設他們之間是線性關係：

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{3,i} + \epsilon_i \text{。}$$

- 統計軟體給我們下列報表：

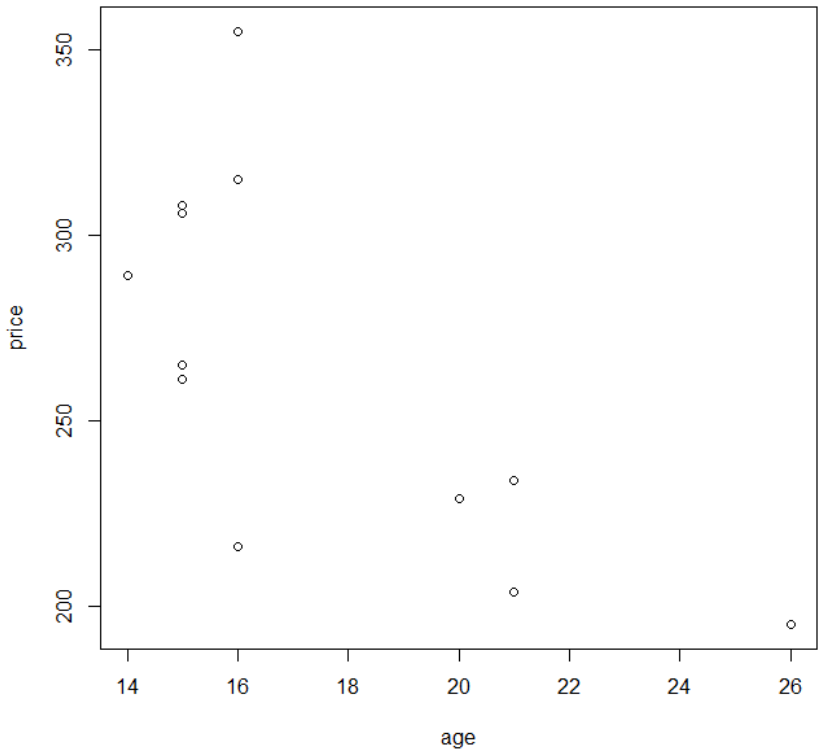
	Coefficients	Standard Error	t Stat	p-value	
Intercept	262.882	83.632	3.143	0.012	
Size	1.533	0.628	2.443	0.037	**
Age	−6.368	2.881	−2.211	0.054	*
			$R^2 = 0.696,\ R^2_{\text{adj}} = 0.629$		

- R^2 從 0.531（只有房子大小為自變數）上升到 0.629。屋齡在 90% 的信心水準下是顯著的。好像不錯！

「非線性」關係

- 可以再改進嗎？
- 根據散佈圖，或許可以試試「非線性」(nonlinear) 的關係：
  - 新房價錢跌得快，舊房則跌得慢。
- 不要假設線性關係式或許有幫助。
- 舉例來說，我們可以試著把屋齡改成屋齡的倒數：

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 \left(\frac{1}{x_{3,i}} \right) + \epsilon_i \text{。}$$



變數轉換

- 若是要用我們的樣本資料去估計

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 \left(\frac{1}{x_{3,i}} \right) + \epsilon_i.$$

  - 這個動作叫「fitting」。
  - 準備一個新變數，其值為 $\frac{1}{\text{age}}$。
  - 把價格、大小和房子屋齡的倒數放入迴歸模型，然後讀報表。
- 我們可以考慮任何的非線性關係（反正都是要製作一個新變數）。
- 這個技巧叫做變數轉換（variable transformation）。

價格（千元）	大小（平方公尺）	1/屋齡（1/年）
315	75	0.063
229	59	0.050
355	85	0.063
261	65	0.067
234	72	0.048
216	46	0.063
308	107	0.067
306	91	0.067
289	75	0.071
204	65	0.048
265	88	0.067
195	59	0.038

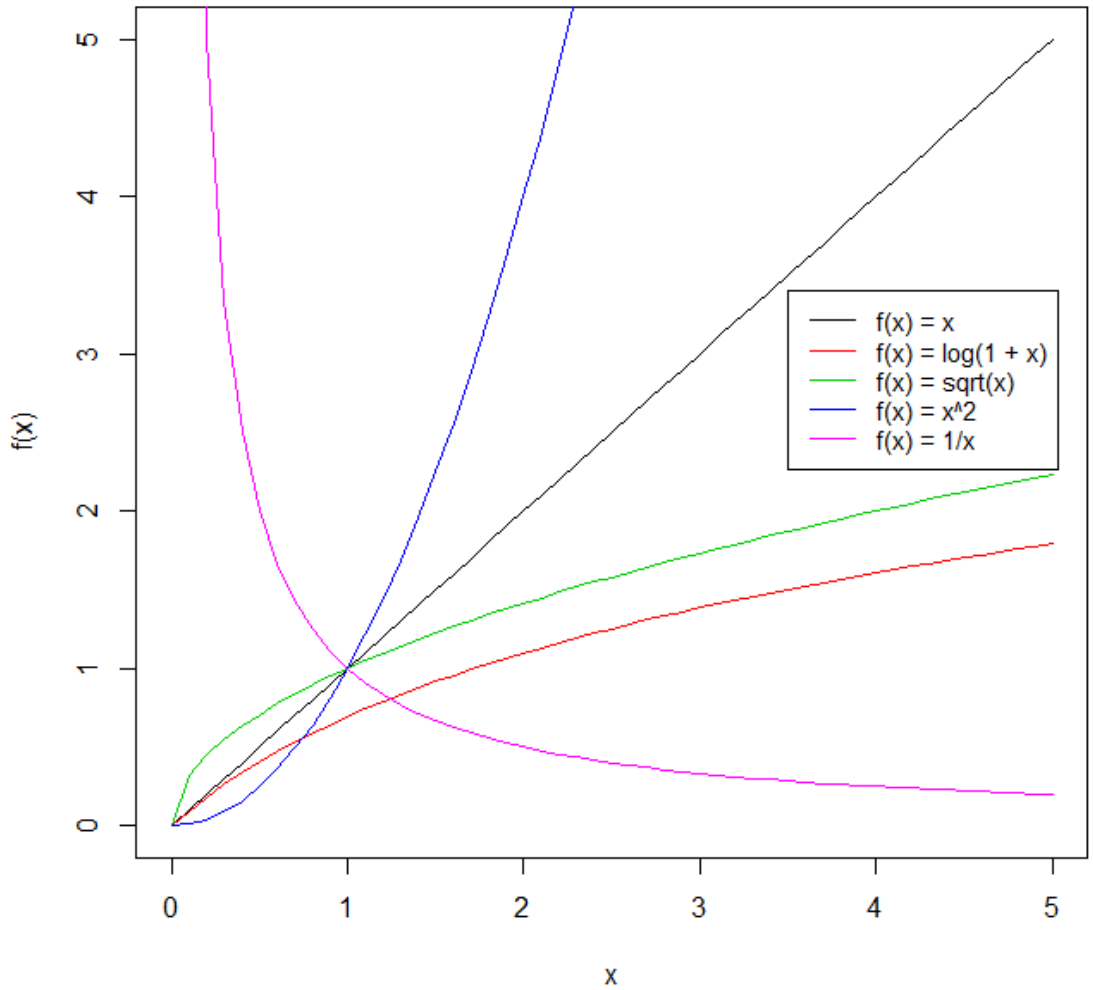
屋齡的倒數

- 統計軟體給我們下列的報表：

	Coefficients	Standard Error	t Stat	p-value	
Intercept	22.905	57.154	0.401	0.698	
Size	1.524	0.647	2.356	0.043	**
1/Age	2185.575	1044.497	2.092	0.066	*
			$R^2 = 0.685,\ R^2_{\text{adj}} = 0.615$		

- 模型檢驗：
  - 變數都顯著 (雖然信心水準不同)。
  - 使用大小和屋齡比使用大小和屋齡的倒數好。
- 「房價在不同屋齡時的下降速率不同」這個假設不被樣本資料支持。
- 把 $\frac{1}{\text{age}}$ 換成 age^2 也沒有比較好。

常見的變數轉換



變數選擇與模型建立

- 有時候我們有非常多的候選自變數。
  - 大小、房間數、屋齡、離最近的公園的距離、離最近的醫院的距離、社區治安、學區...
  - 就算只考慮線性關係，p 個候選自變數就有 $2^p - 1$ 種組合。
  - 事實上每個變數都可以被轉換。
  - 之後甚至還可以討論變數間的交互作用。
- 如何找出「最好的」迴歸模型 (如果有的話)?

變數選擇與模型建立

- 世界上沒有「最好的」模型，但是有「好」模型。
- 一些建議：
  - 用散佈圖檢視每個自變數跟應變數間的關係，據此嘗試變數轉換。
  - 檢視自變數間的兩兩關係。如果某兩者高度相關，常常就有一個不需要。我們說它們之間有共線性 (multicollinearity)。
  - 一旦有了一個模型，檢視每個變數的 p-value，並試著移除不顯著的變數。要注意的是，這可能會影響到剩餘變數的顯著性。
- 反覆修正，直到你找不到更好的模型。
  - R^2 大、修正的 R^2_{adj} 大、p-value 們小。
  - 統計軟體通常可以 (部份地) 自動化上述流程，不過人為決策還是必要的。
  - 有時關鍵其實是去找尋新的自變數。
- 直覺與經驗可能會幫上忙 (或幫倒忙)。

課程大綱

- 基本原理。
- 變數轉換與選擇。
- 一個案例。
- 類別型態自變數。

一個案例：票券銷售

- 一個劇團過去六年做了近千場演出。
- 老闆想要增加票房賣座度。
- 關鍵問題：什麼是影響賣座度的關鍵因子？
  - 讓我們用售票張數來定義賣座度。
  - 潛在因子：演出年份、演出月份、演出於星期幾、演出時間 (早上、下午、晚上)、演出地點、演員、戲劇種類、票價...
- 老闆隨機抽出 100 場演出，給你這些演出的一些資訊。
  - 都在週末演出、公開售票、以同樣方式售票。
  - 每一場演出的票價都不隨時間改變。
- 做為一名顧問，如何透過統計與資料分析幫助劇團？

變數

- 共有六個變數：

變數	意義
$Year$	演出進行的年份（1、2、...、6）
$Time$	演出進行的時間（早上、下午、晚上）
$Capacity$	表演廳的座位數
$AvgPrice$	所有票種的票價平均數
$SalesQty$	總售出張數
$SalesDuration$	起售日期至演出日期的間隔天數

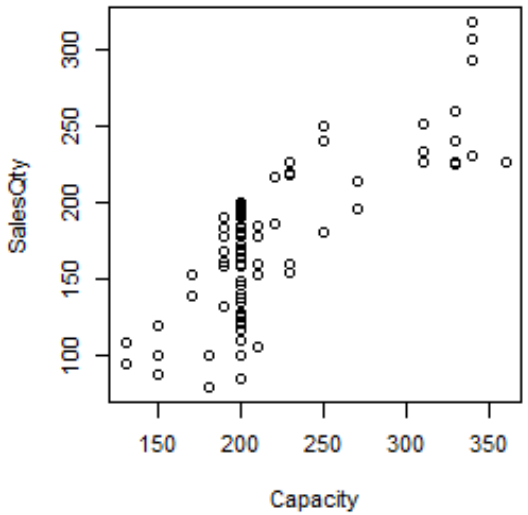
- 座位數後售票數已經被等比例調整（scaling）過了。

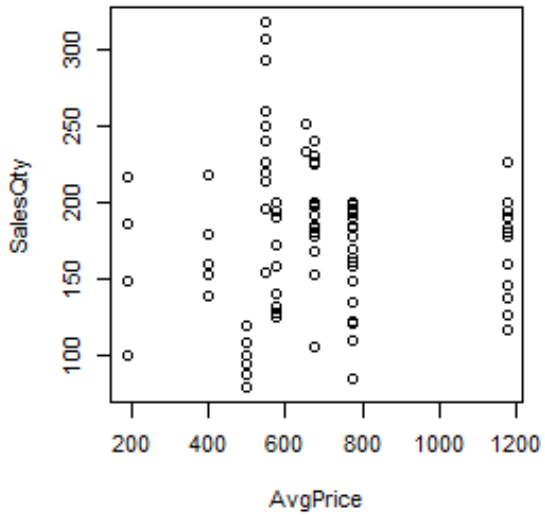
資料 (一部分)

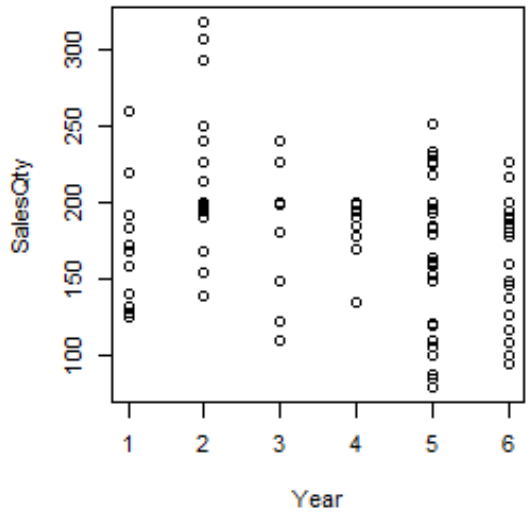
Yr.	Tm.	Cap.	A.P.	Qty	S.D.	Yr.	Tm.	Cap.	A.P.	Qty	S.D.
5	A	230	400	218	50	2	M	190	575	190	289
5	A	150	500	119	46	6	A	130	500	108	89
5	A	230	400	160	126	4	E	200	775	169	100
5	A	200	775	200	324	4	E	200	775	135	259
6	E	190	1175	178	115	5	A	310	650	251	346
6	A	190	1175	183	109	2	A	250	550	250	145
5	E	190	775	161	58	1	A	190	675	183	254
3	A	200	675	200	112	6	A	200	1175	146	110
5	E	200	775	158	323	1	M	200	575	140	94
1	M	200	575	128	360	4	A	200	775	195	255

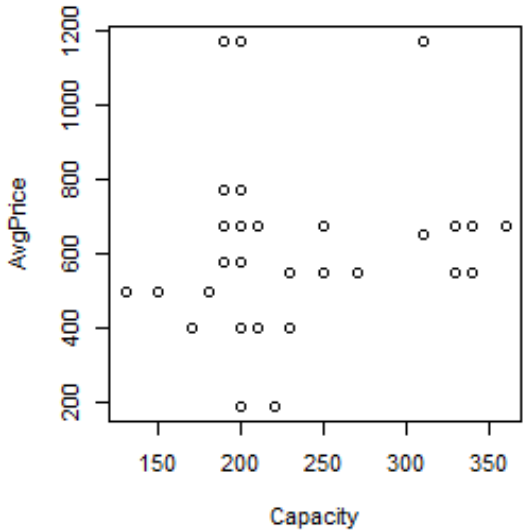
迴歸分析

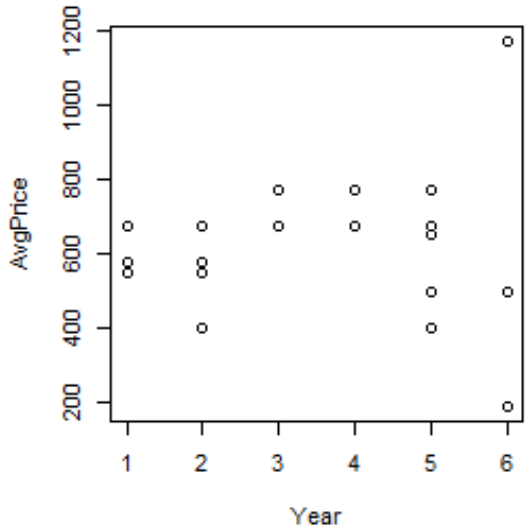
- 讓我們先來試幾個自變數。
  - 應變數：$SalesQty$.
  - 自變數：$Capacity$, $AvgPrice$, $Year$.
- 請注意 *Year* 是數值型資料：
  - 兩個值之間的距離有實際意義：$4-2$ 和 $5-3$ 都表示差兩年。
  - 值有單一變化方向。
  - 如果是*月份*，其值就會循環，那麼 $12-11$ 跟 $1-12$ 就完全不同。
- 散佈圖有用：
  - 變數選擇：哪個自變數可能有影響？
  - 變數轉換：一個自變數如何影響應變數？
  - 共線性：有沒有兩個變數高度相關？

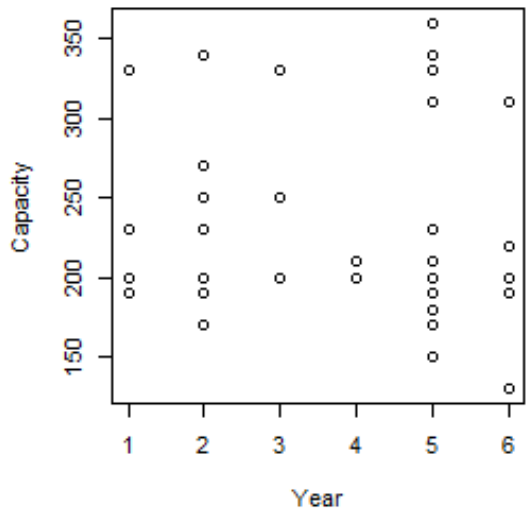

SalesQty
Capacity
100
150
200
250
300
150
200
250
300
350


SalesQty
AvgPrice
100
150
200
250
300
200
400
600
800
1000
1200


SalesQty
Year
100
150
200
250
300
1
2
3
4
5
6


AvgPrice
Capacity
200
400
600
800
1000
1200
150
200
250
300
350


AvgPrice
Year
200
400
600
800
1000
1200
1
2
3
4
5
6


Capacity
Year
150
200
250
300
350
1
2
3
4
5
6

迴歸分析

- 看起來 $Capacity$、$AvgSales$ 和 $Year$ 都值得一試。
- 如果我們將它們分別放進迴歸模型：
  - $SalesQty = 20.79 + 0.72Capacity$：$R^2 = 0.538$、$p$-value ≈ 0。
  - $SalesQty = 174.9 + 0.0028AvgPrice$：$R^2 = 0.0002$、$p$-value $= 0.885$。
  - $SalesQty = 203.6 - 6.77Year$：$R^2 = 0.063$、$p$-value $= 0.0115$。
- 如果我們將它們一起放進去：
  - 迴歸模型是

$$SalesQty = 24.742 + 0.702Capacity + 0.027AvgPrice - 4.696Year。$$

  - $R^2 = 0.57$、$R^2_{\text{adj}} = 0.556$，p-value 分別是 0、0.056 和 0.019。
- 不要分別放，要一起放。

加入 $Time$

- $Time$（早上、下午、晚上）也可能有影響。
- 但是它是類別資料。
  - 更精確地講，它是名目資料。
  - 就算我們把 $Time$ 編碼成 1、2 跟 3，我們也不能就把它當成數值資料。
- 對於一個類別變數，我們必須使用一或數個虛擬變數（dummay variable、indicator variables）。

加入 $Time$

- $Time$ (早上、下午、晚上) 也可能有影響。但是它是類別資料。
  - 為什麼不編碼成數值然後直接做迴歸分析？
- 假設我們把 (morning, afternoon, evening) 編碼成 $(1, 2, 3)$:
  - 迴歸模型是

$$SalesQty = 164.021 + 6.313Time。$$

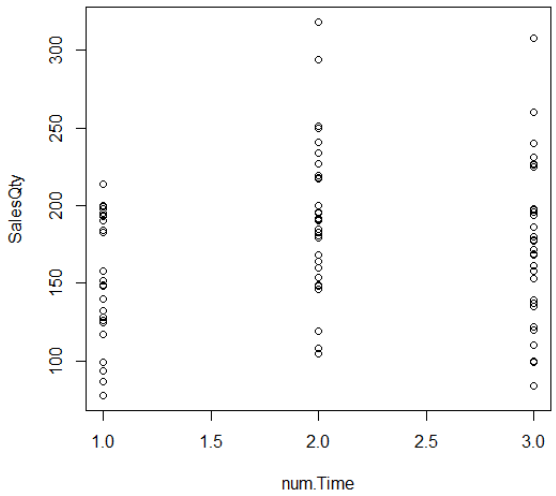
  - 這有錯嗎？

數值編碼沒有意義

- 不同的編碼就會給我們不同的迴歸模型！
- 我們也可以把 (morning, afternoon, evening) 編碼成 $(1, 2, 10)$ 或 $(3, 1, 2)$：

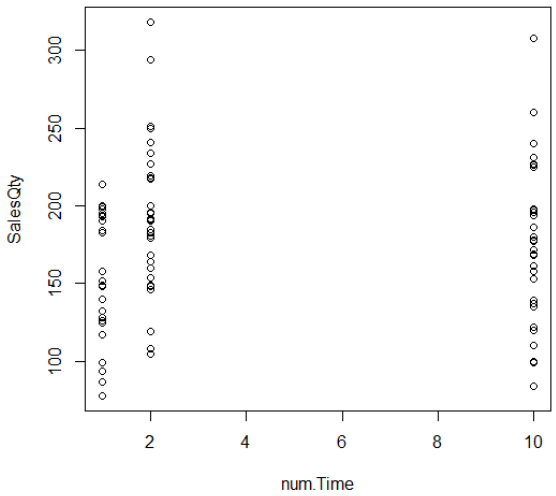
$$\textit{SalesQty} = 164.021 + 6.313\,\textit{Time}$$

p-value $= 0.294$



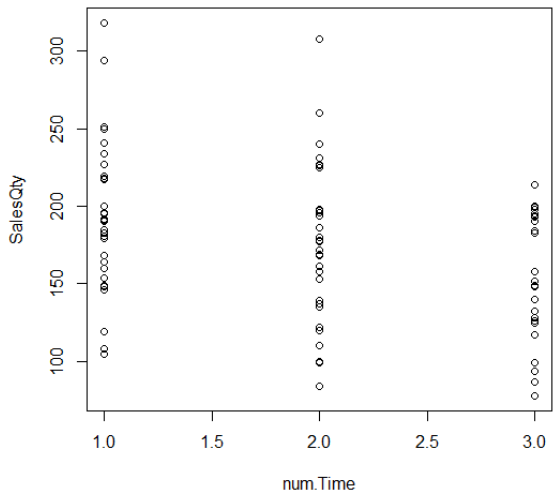
$$\textit{SalesQty} = 177.224 - 0.075\,\textit{Time}$$

p-value $= 0.95$



$$\textit{SalesQty} = 205.725 - 15.091\,\textit{Time}$$

p-value $= 0.0084$



課程大綱

- 基本原理。
- 變數轉換與選擇。
- 一個案例。
- 類別型態自變數。

二元變數

- 類別變數需要特別處理。
- 先看看一個特殊情況：如果一個類別變數是二元（binary）的，我們就可以直接將之編碼成 0 和 1 並且直接放進迴歸模型。
  - 男/女、生/死、買/沒買、公立/私立...
  - 編碼成 1 和 0、1 和 2 或 7 和 8 也都沒問題。
  - 編碼成 1 和 -1、1 和 5 或 4 和 8 就比較不好。
- 這是因為迴歸模型的係數代表「當其他自變數不變，而此自變數增加一單位」時，應變數會如何變化。
- 當一個二元變數被編碼成 0 和 1，它的係數就告訴我們「如果這個變數從 0 變成 1（且其他變數都不變），我們預期應變數會增加 $\hat{\beta}_i$。」
- 如果一個類別變數有超過兩個可能的值呢？

虛擬變數

- 假設 x 有三個可能的值 A、B 跟 C。
- 讓我們先選一個基準點 (reference level)，比如說 A。
- 接著創造兩個虛擬變數 (dummy variable、indicator variable) x^B 和 x^C:

$$x^B = \begin{cases} 1 & \text{若} x = \text{B} \\ 0 & \text{若為其他情況} \end{cases} \quad \text{和} \quad x^C = \begin{cases} 1 & \text{若} x = \text{C} \\ 0 & \text{若為其他情況} \end{cases}$$

換言之，我們有如下對應：

x	x^B	x^C
A	0	0
B	1	0
C	0	1

虛擬變數

- 現在我們把 x^B 和 x^C 放進迴歸模型

$$y = \hat{\beta}_0 + \cdots + \hat{\beta}^B x^B + \hat{\beta}^C x^C \text{。}$$

  - 如果 x 從 A 變成 B (而且其他變數都不變)，應變數預期將增加 $\hat{\beta}^B$ 。
  - 如果 x 從 A 變成 C (而且其他變數都不變)，應變數預期將增加 $\hat{\beta}^C$.
  - 如果 x 從 B 變成 C (而且其他變數都不變)，我們沒什麼結論。
- 我們用 x 把資料分成三組 (A、B 和 C)。
- 我們在問，再移除其他變數的影響之後，A 組和 B 組以及 A 組和 C 組間是否有顯著差異。

虛擬變數的通則

- 如果變數 x 有五個可能的值 M、N、O、P 和 Q。
  - 我們首先選擇一個基準點，比如說 P。
  - 我們接著創造四個虛擬變數：

x	x^M	x^N	x^O	x^Q
M	1	0	0	0
N	0	1	0	0
O	0	0	1	0
P	0	0	0	0
Q	0	0	0	1

  - 在 P 組和 M 組、P 組和 N 組、P 組和 O 組，以及 P 組和 Q 組間之否有顯著差異？
- 一個類別變數若是有 k 個可能的值，我們就需要 $k-1$ 虛擬變數。

$Time$ 的虛擬變數

- $Time$ 有三個值：morning、afternoon 和 evening。
- 讓我們選 afternoon 當基準點。
- 我們需要兩個虛擬變數：

$Time$	$Time^M$	$Time^E$
morning	1	0
afternoon	0	0
evening	0	1

- 用 $Time^M$ 和 $Time^E$ 做為自變數，我們會得到

$$SalesQty = 191 - 30.069 Time^M - 16.303 Time^E。$$

兩個變數的 p-values 各是 0.009 和 0.138。

- 如果把一場演出換時間從下午移到早上，我們預期會少賣 30.069 張票。

$Time$ 的虛擬變數

- 讓我們把手上有的變數都加進去：

$$SalesQty = 0.696Capacity + 0.027AvgPrice - 5.282Year$$
$$- 14.387Time^{M} - 21.328Time^{E}.$$

	Coefficients	Standard Error	t Stat	p-value	
Intercept	39.280	19.724	1.992	0.049	**
$Capacity$	0.696	0.069	10.263	0.000	***
$AvgPrice$	0.027	0.013	2.033	0.045	**
$Year$	−5.282	1.931	−2.735	0.007	***
$Time^{M}$	−14.387	7.784	−1.848	0.068	*
$Time^{E}$	−21.328	7.227	−2.951	0.004	***
				$R^2 = 0.608,\ R^2_{\text{adj}} = 0.587$	

結語

- 當遇到自變數是類別變數時，我們就需要加虛擬變數。
  - 虛擬變數的值非 0 則 1。
- 如果它有 k 個可能的值，我們就需要 $k-1$ 個虛擬變數。
  - 在原本變數中當基準點的值，在所有虛擬變數中都被設成 0。
  - 在原本變數中不是基準點的值，會有恰好一個虛擬變數被設成 1。
- 我們只是在（只能）檢定基準點和非基準點之間是否有顯著差異。
  - 對於兩個非基準點之間是否有顯著差異，我們一無所知。
  - 真的要知道，就要換基準點。
- 如果有任何一個虛擬變數是顯著的，而你因此想要把它留在迴歸模型中，那所有的為了同一個類別變數而產生的虛擬變數就都要留下。

給工程師的統計學與資料分析 123

第四單元：迴歸分析 (2)

孔令傑

國立臺灣大學資訊管理學系

2017 年 9 月 2 日

課程大綱

- 交互作用。
- 內生性與殘差分析。
- 羅吉斯迴歸分析。

變數間的交互作用

- 在迴歸模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

  中，β_i 衡量 x_i 對 y 的影響。
- 有時變數 x_i 對 y 的影響程度取決於另一個變數 x_j。
- 舉例而言：當我們比較房屋價格、大小與房間數量間的關係。
  - 當房屋大時，愈多房間會讓價值愈高。
  - 當房屋小時，太多房間就不好了。
- 再舉一例：考量商品的市場需求。
  - 當需求的價格敏感度高；價格提高時，需求下降得多。
  - 價格敏感度可能在男性、女性上有所不同。
- 在這種情況下，我們說變數 x_i 與 x_j 存在交互作用 (interaction)。

為交互作用建立模型

- 為了建立交互分析模型，首先我們必須用 x_i 與 x_j 組成新變數 $x_i x_j$，也就是兩變數的乘積。
- 在迴歸模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2 \cdots$$

中，$\beta_{1,2}$ 衡量變數 x_1 與 x_2 之間的交互作用。
  - 變數 x_1 對 y 的影響係數為 $\beta_1 + \beta_{1,2} x_2$。
  - 變數 x_2 對 y 的影響係數為 $\beta_2 + \beta_{1,2} x_1$。
- 迴歸模型中的平方項 x_i^2 是個特例：

$$y = \beta_0 + \beta_1 x_1 + \beta_1' x_1^2 + \cdots$$

在此，x_1 對 y 的影響係數隨著 x_1 不同而不同。

$Time$ 與 $AvgPrice$ 的交互作用

- 變數 $Time$ 與 $AvgPrice$ 間有交互作用嗎?
- 讓我們在模型中加入變數 $Time^M \times AvgPrice$ 與 $Time^E \times AvgPrice$：

	係數	標準差	t 檢定值	p 值	
Intercept	55.876	22.652	2.467	0.015	**
$Capacity$	0.676	0.068	9.950	0.000	***
$Year$	−6.192	1.966	−3.149	0.002	***
$Time^M$	−55.205	23.829	−2.317	0.023	**
$Time^E$	−19.105	21.81	−0.876	0.383	
$AvgPrice$	0.015	0.019	0.836	0.405	
$Time^M \times AvgPrice$	0.054	0.030	1.792	0.076	*
$Time^E \times AvgPrice$	−0.004	0.030	−0.136	0.892	
			$R^2 = 0.624,\ R^2_{\text{adj}} = 0.595$		

- 若我們想在模型中保留變數 $Time^E \times AvgPrice$，我們也必須保留變數 $Time^M \times AvgPrice$、$AvgPrice$、$Time^M$ 與 $Time^E$。

$Time$ 影響 $AvgPrice$ 的相關係數

- 讓我們看一下 $Time$ 與 $AvgPrice$:

	係數	標準差	t 檢定值	p 值	
$Time^M$	-55.205	23.829	-2.317	0.023	**
$Time^E$	-19.105	21.81	-0.876	0.383	
$AvgPrice$	0.015	0.019	0.836	0.405	
$Time^M \times AvgPrice$	0.054	0.030	1.792	0.076	*
$Time^E \times AvgPrice$	-0.004	0.030	-0.136	0.892	

- 在不同時間人們擁有不同的價格敏感度。當價格提升 1 元，我們預期：
  - 下午的銷量增加 0.015。
  - 早上的銷量增加 $0.015 + 0.054 = 0.069$。
  - 晚上的銷量增加 $0.015 - 0.004 = 0.011$。

$AvgPrice$ 影響 $Time$ 的相關係數

- 讓我們再看一次變數 $Time$ 與 $AvgPrice$:

	係數	標準差	t 檢定值	p 值	
$Time^M$	-55.205	23.829	-2.317	0.023	**
$Time^E$	-19.105	21.81	-0.876	0.383	
$AvgPrice$	0.015	0.019	0.836	0.405	
$Time^M \times AvgPrice$	0.054	0.030	1.792	0.076	*
$Time^E \times AvgPrice$	-0.004	0.030	-0.136	0.892	

- 當我們把一場演出從下午改時間到早上，我們預期銷量會增加

$$-55.205 + 0.054 AvgPrice。$$

若 $AvgPrice = 500$，我們預期銷量增加

$$-55.205 + 0.054 \times 500 = -28.205。$$

Time 與 *Year* 的交互作用

- *Time* 與 *Year* 會影響彼此對銷售量的影響嗎?

	係數	標準差	t 檢定值	p 值	
(Intercept)	39.597	22.31	1.775	0.079	*
Capacity	0.693	0.068	10.267	0.000	***
AvgPrice	0.024	0.013	1.799	0.075	*
$Time^E$	−2.696	18.562	−0.145	0.885	
$Time^M$	−25.114	18.303	−1.372	0.173	
Year	−4.703	2.944	−1.597	0.114	
$Time^E \times Year$	−4.841	4.302	−1.125	0.263	
$Time^M \times Year$	2.898	4.166	0.695	0.489	
			$R^2 = 0.620,\ R^2_{\text{adj}} = 0.591$		

- *Time* 及 *Year* 的交互作用不顯著。
  - 人們的對演出時間的偏好在不同年份沒有顯著差別 (每年偏好都相同)。
  - 我們可以移除交互作用項。

總結

- 兩個變數間的交互作用可以利用交乘項來放進模型。
  - 若交乘項的係數顯著地非零，則一變數的影響程度取決於另一個變數。
- 三種保留變數的規則是：
  - 高次項：若需保留 x^k，我們也須保留 x^{k-1}、x^{k-2} 直到 x。
  - 虛擬變數：對於一組為了一個類別變數而做出來的虛擬變數，要留一個就要留下全部。
  - 交互作用：若希望保留變數 $x_i x_j$，我們也須保留 x_i 與 x_j。
- 加入變數 $x_i x_j x_k$ 到迴歸模型中也是可以嘗試的做法。

課程大綱

- 交互作用分析。
- 內生性與殘差分析。
- 羅吉斯迴歸分析。

SalesDuration

- 讓我們考慮變數 *SalesDuration*。
  - 此變數是一場表演的開始售票日與實際表演日間的差異日數。
  - 也就是一場表演的公開售票日數。
  - 銷售期間 (*SalesDuration*) 愈長，銷售量會愈大嗎?
- 我們希望能加入變數 *SalesDuration* 到我們的迴歸模型中。
- 在此案例中卻有些困難：
  - 通常劇團在一年年終時就會決定下一年的表演日程。
  - 絕大多數的演出都是排好的。
  - 演出門票多在實際演出前的幾個月銷售完畢。
  - 然而，若該系列演出非常受歡迎，劇團或許會決定多加開幾場。
  - 額外加開的表演有較短的 *SalesDuration*，卻也有較高的銷售量 *SalesQty*。
- 簡言之，*SalesQty* 影響 *SalesDuration*。

內生性（endogeneity）

- 若一個迴歸模型中，自變數被應變數影響，我們說這個模型有內生性（endogeneity）問題。
  - 若我們加入變數 $SalesDuration$ 到我們模型中，就有內生性問題。
  - $Year$、$Time$、$Capacity$ 與 $AvgPrice$ 則通常沒有內生性問題。
  - 若這些變數會因為銷售量而被決定，則內生性問題同樣會產生。
- 內生性會導致有偏差的預測或錯誤的解釋。
- 若新增 $SalesDuration$ 到模型中，我們可能會蓄意延後開始售票日！

範例：促銷電話

- 假設有間銀行讓它的員工打電話邀請潛在客戶來存款（或借款）。
- 很多因素可能影響這個結果（成功或失敗）：
  - 受訪人員的性別、年齡、職業、教育水準等等。
  - 打電話的員工的性別、年齡、經驗等等。
  - 電訪日期、電訪時間、當天天氣等等。
- 這些過去的電訪資訊都有被錄音記錄下來。
- 每一通電話的長度也有被記錄下來：
  - 我們發現這個變數與電訪成功或失敗有高度相關性。
  - 然而這不能在以成功或失敗為應變數的迴歸模型中被當作自變數。
  - 因為此變數會被結果影響：一旦客戶答應存款了，談話會因為需要談論存款細節而加長。
- 若我們加入談話長度進入模型，我們會鼓勵電訪人員說愈慢愈好。

避免內生性

- 避免內生性的方法：
  - 移除整個自變數。
  - 移除自變數中會被應變數影響的部份。
- 在票務銷售案例中：
  - 我們可以移除變數 $SalesDuration$。
  - 我們可以移除額外加開的表演。
- 在電訪案例中：
  - 我們可以移除通話時間長短這個變數。
  - 我們可以只使用受訪者答應之前的時間。

殘差分析

- 當做迴歸模型時：
  - 我們嘗試發掘變數間的潛藏關係。
  - 我們假設（相信）真實世界是

$$y = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$$

    並用資料來猜測模型的係數。
  - 我們用 R^2、R^2_{adj} 和 p-值來檢驗模型。
- 若模型良好，隨機誤差 ϵ 應該真的是隨機的。
  - 殘差 ϵ 不應該有系統性的規律（systematic pattern）。
- 我們需要做殘差分析（residual analysis）。

四項假設

- 假設變數 x 與 y 之間有一個係數 β_0 與 β_1 未知的線性關係

$$y = \beta_0 + \beta_1 x + \epsilon \text{ ,}$$

  其中 ϵ 是隨機誤差。
- 理想上，殘差 (residual) ϵ 應該符合四個假設:
  - 期望值為零：給定任何 x 的值，預期的 ϵ 值都是 0。
  - 變異數一致：給定任何 x 的值，ϵ 的變異數都相同。
  - 彼此獨立：ϵ 在不同 x 下的值都是互相獨立的。
  - 常態分佈: 給定任何的 x 值，ϵ 都遵守常態分佈。
- 對於一個迴歸模型，我們會需要：
  - 以預測為目的：需要前三項假設。
  - 以解釋為目的：需要所有的假設。

檢驗假設

- 假設我們手上有樣本資料 $\{(x_i, y_i)\}_{i=1,\ldots,n}$。
- 線性迴歸幫助我們找到 $\hat{\beta}_0$ 與 $\hat{\beta}_1$ 並得出以下迴歸公式

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i \ ,$$

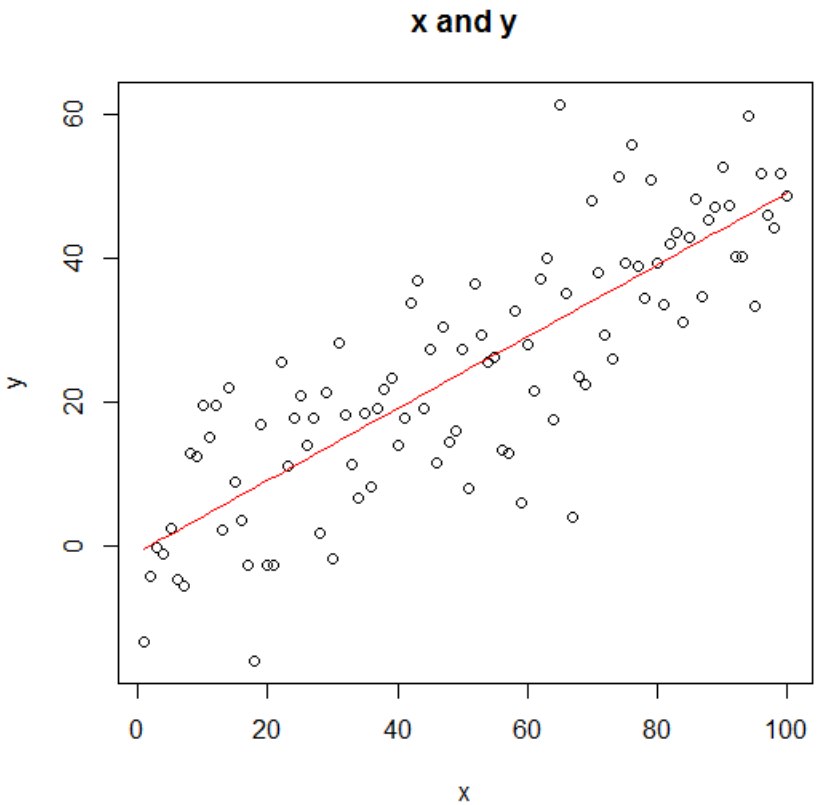
  其中 ϵ_i 就是預測值 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 與實際值 y_i 的殘差 (residual)。
- 藉由做殘差分析，我們檢測這些 ϵ_i 是否符合四項假設。
- 雖然學術上有一系列相關的統計檢定，這邊我們只介紹用圖形方式做直觀的檢測。

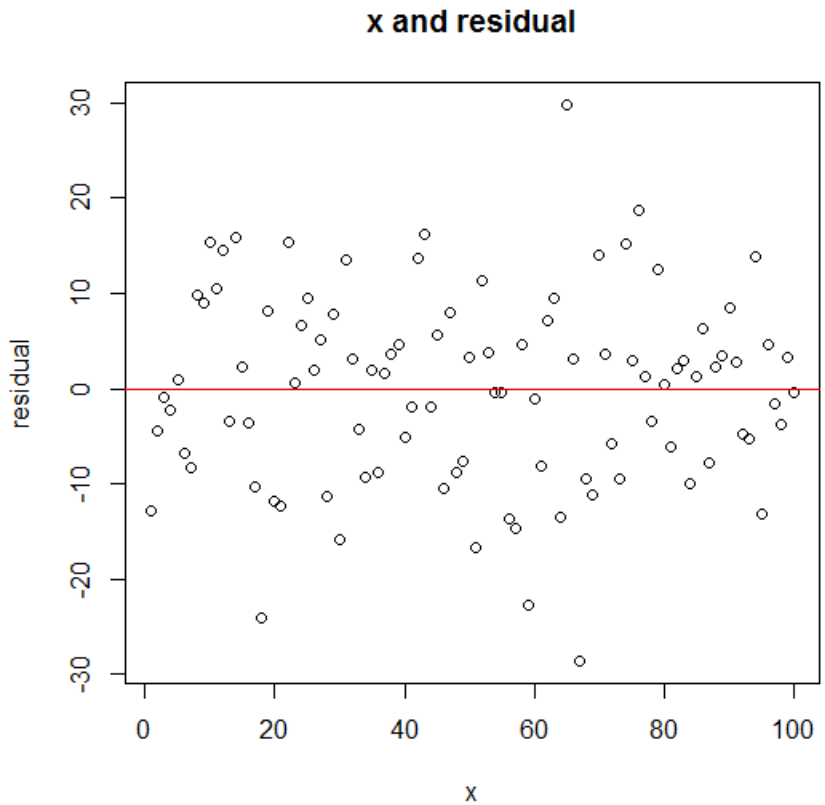
殘差散佈圖與殘差直方圖

- 我們可以根據不同 x_i 繪製 ϵ_i 的殘差散佈圖 (residual plot)。
  - 可以做「期望值為零」、「變異數一致」、「彼此獨立」這三個假設。
  - 從圖形中應該看不出系統性規律。
- 我們可以做殘差直方圖 (residual histogram)。
  - 可以檢測常態分佈。
  - 圖形應對稱且符合鐘形分佈。
- 一般來說：
  - 正確的圖形不保證好的模型。
  - 但錯誤的圖形通常表示模型有問題!

殘差散佈圖與殘差直方圖

- 這邊是一些用人造資料做的範例：

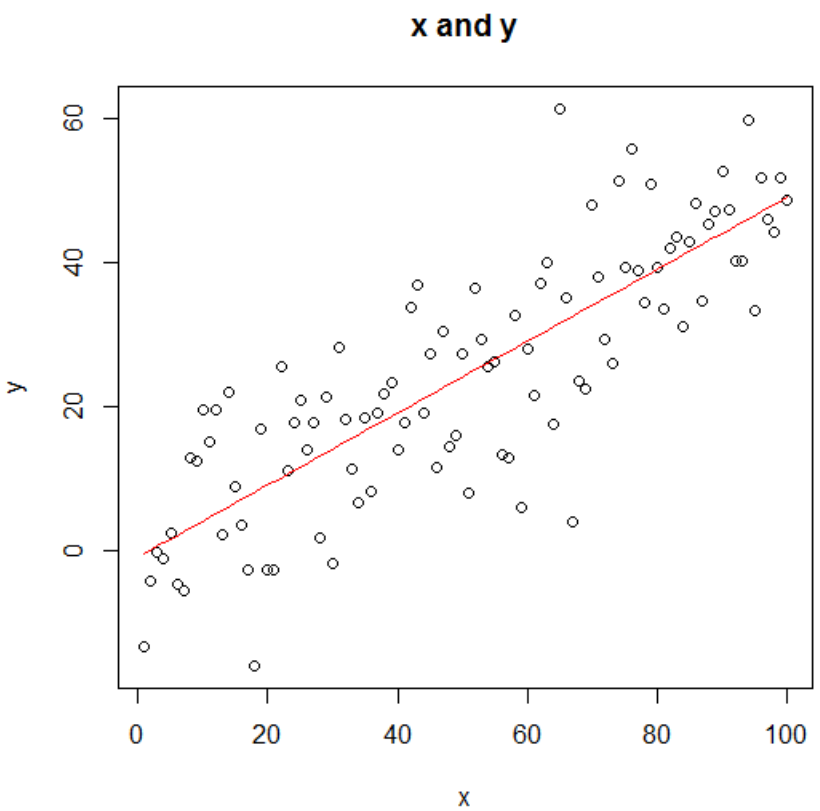


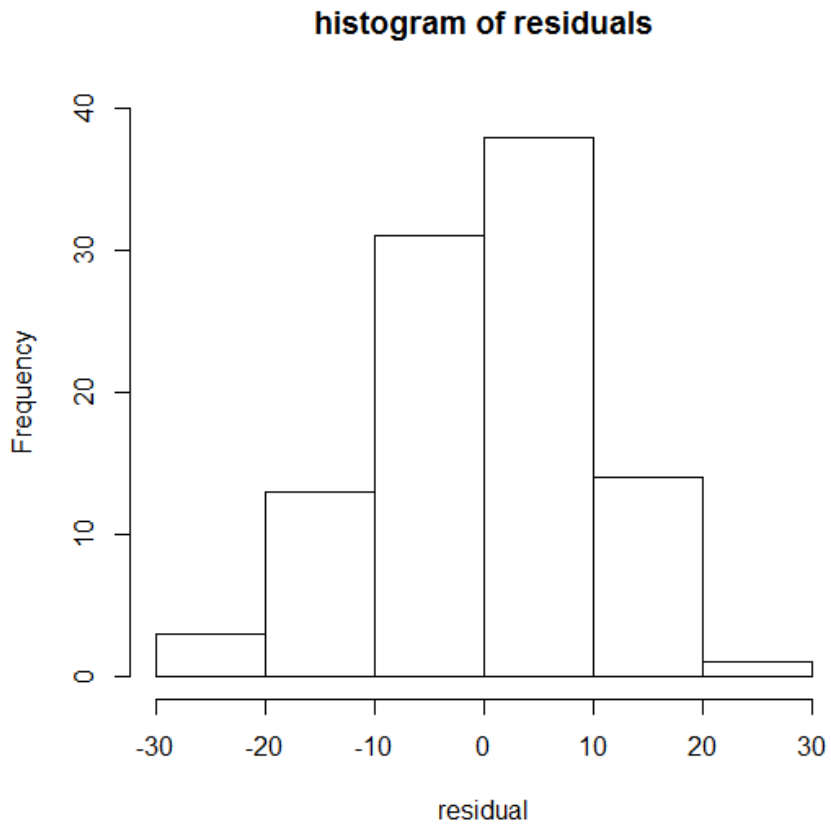


- 看不出系統性規律，很好！

殘差散佈圖與殘差直方圖

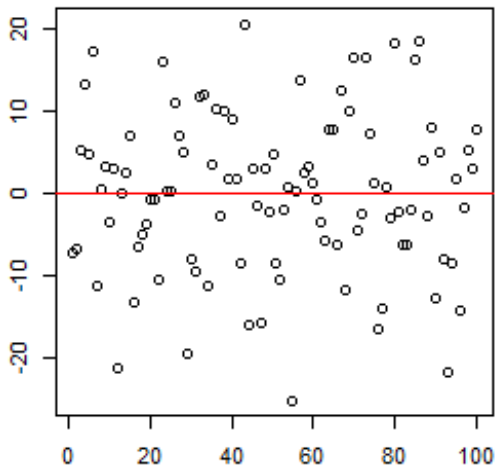
- 這邊是一些用人造資料做的範例：

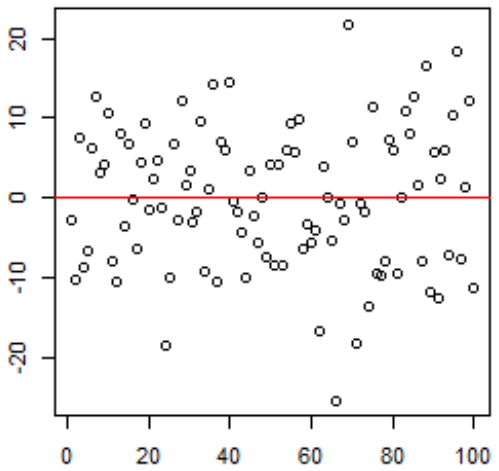


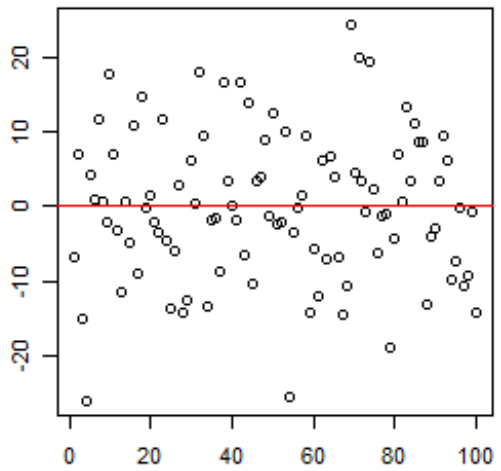


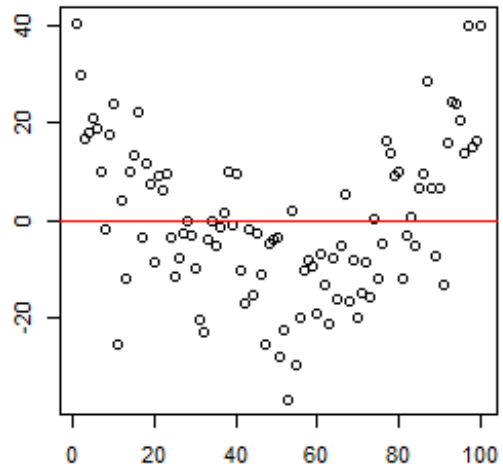
- 直方圖呈鐘形分布且對稱，很好！

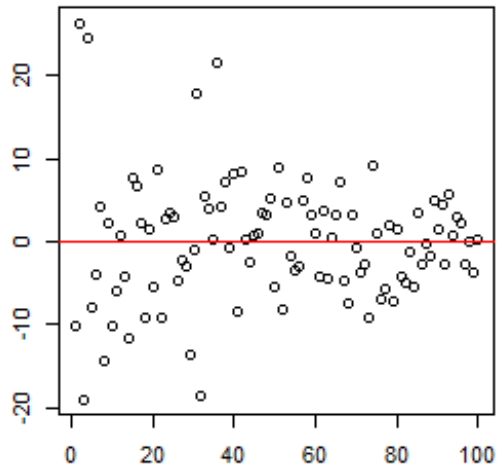
成功與失敗的殘差散佈圖

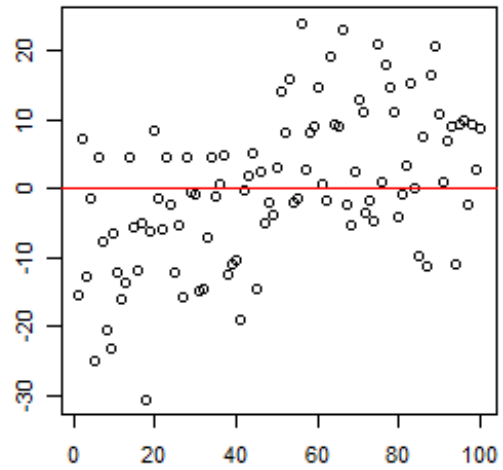




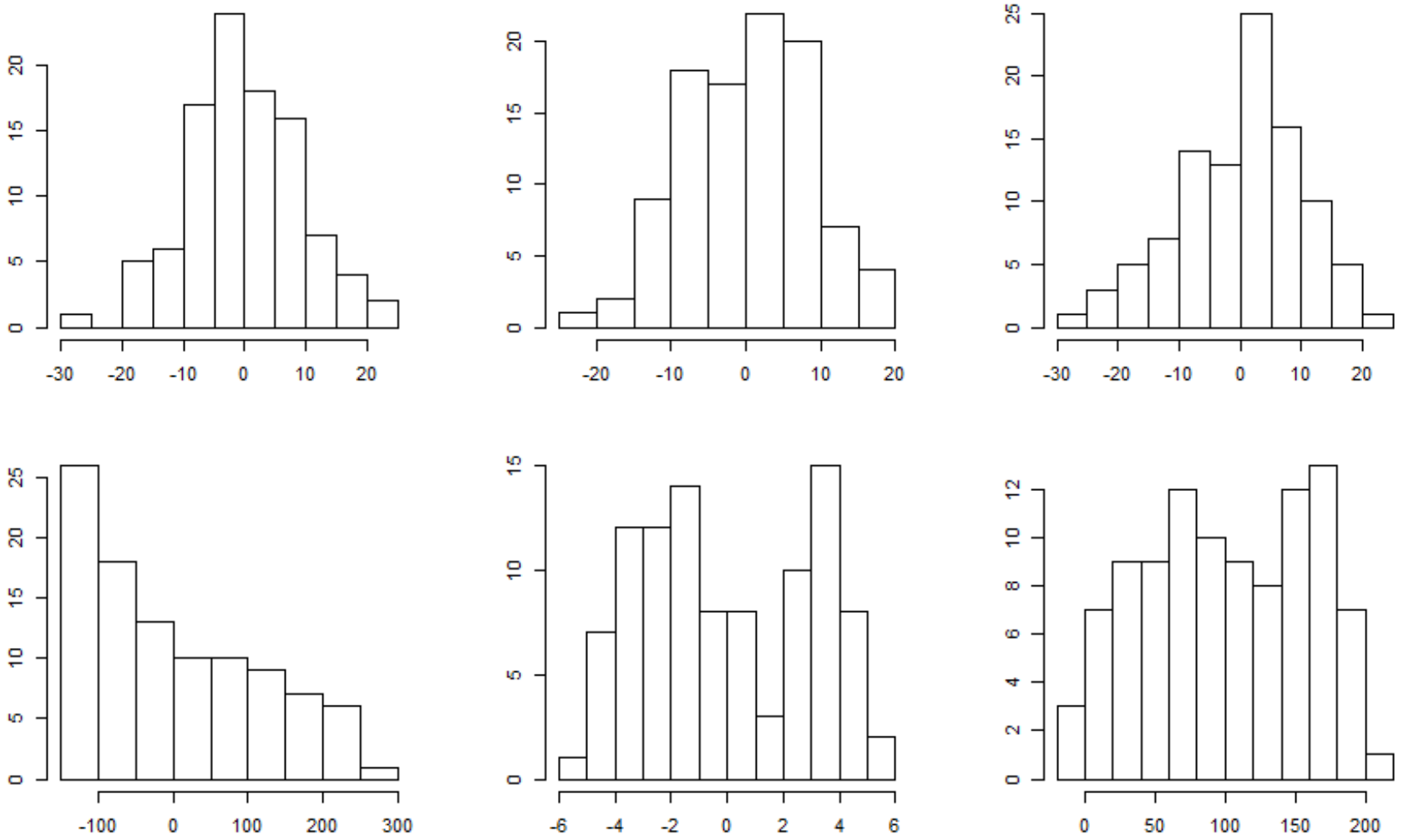








成功與失敗的殘差直方圖



殘差分析與複迴歸

- 若我們有一個複迴歸模型

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_p + \epsilon \ ,$$

  我們也應該做殘差分析。
- 需要繪製很多個殘差散佈圖。
  - 縱軸是 ϵ。
  - 橫軸是 $(x_1, x_2, ..., x_p)$ 的各種函數。
  - 至少應該個別測試第 k 個自變數 x_k，以及預測值 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_p$。

課程大綱

- 交互作用分析。
- 內生性與殘差分析。
- 羅吉斯迴歸分析。

羅吉斯迴歸分析

- 至今我們一直使用的迴歸模型是以數值變數（quantitative variables）作為應變數。
  - 這類迴歸也被稱作 ordinary regression。
- 以類別變數（qualitative variables）作為應變數時，不能用 ordinary regression。
- 解決方法之一是使用羅吉斯迴歸分析（logistic regression）。
  - 羅吉斯迴歸分析允許應變數為類別變數。
  - 這裡我們只討論二元類別變數（binary variables）。
- 讓我們先了解為何 ordinary regression 在變數為分類資料時不適用。

範例：存活機率分析

- 有 45 人在登山時困在暴風中，並且部分成員在風暴中不幸喪生[1]。
- 我們想瞭解性別與年齡如何影響成員的存活機率。

Age	Gender	Survived	Age	Gender	Survived	Age	Gender	Survived
23	Male	No	23	Female	Yes	15	Male	No
40	Female	Yes	28	Male	Yes	50	Female	No
40	Male	Yes	15	Female	Yes	21	Female	Yes
30	Male	No	47	Female	No	25	Male	No
28	Male	No	57	Male	No	46	Male	Yes
40	Male	No	20	Female	Yes	32	Female	Yes
45	Female	No	18	Male	Yes	30	Male	No
62	Male	No	25	Male	No	25	Male	No
65	Male	No	60	Male	No	25	Male	No
45	Female	No	25	Male	Yes	25	Male	No
25	Female	No	20	Male	Yes	30	Male	No
28	Male	Yes	32	Male	Yes	35	Male	No
28	Male	No	32	Female	Yes	23	Male	Yes
23	Male	No	24	Female	Yes	24	Male	No
22	Female	Yes	30	Male	Yes	25	Female	Yes

[1]資料來源為教科書 *The Statistical Sleuth*，作者為 Ramsey 與 Schafer。故事因本課程需求被修正過。

敘述統計

- 整體存活機率是 $\frac{20}{45} = 44.4\%$。
- 存活率似乎與性別有關：

性別分群	存活人數	分群大小	存活機率
Male	10	30	33.3%
Female	10	15	66.7%

- 存活率似乎與年齡有關：

年齡分群	存活人數	分群大小	存活機率
$[10, 20)$	2	3	66.7%
$[21, 30)$	11	22	50.0%
$[31, 40)$	4	8	50.0%
$[41, 50)$	3	7	42.9%
$[51, 60)$	0	2	0.0%
$[61, 70)$	0	3	0.0%

- 我們可以做得更好嗎？比如說，我們可否預測或解釋存活與否?

Ordinary regression 的問題

- 假設我們建構迴歸模型

$$survival_i = \beta_0 + \beta_1 age_i + \beta_2 female_i + \epsilon_i \text{ ，}$$

其中 age 代表一人的年齡，$gender$ 為 0 代表是男性，為 1 代表是女性；$survival$ 為 1 代表存活，為 0 代表死亡。

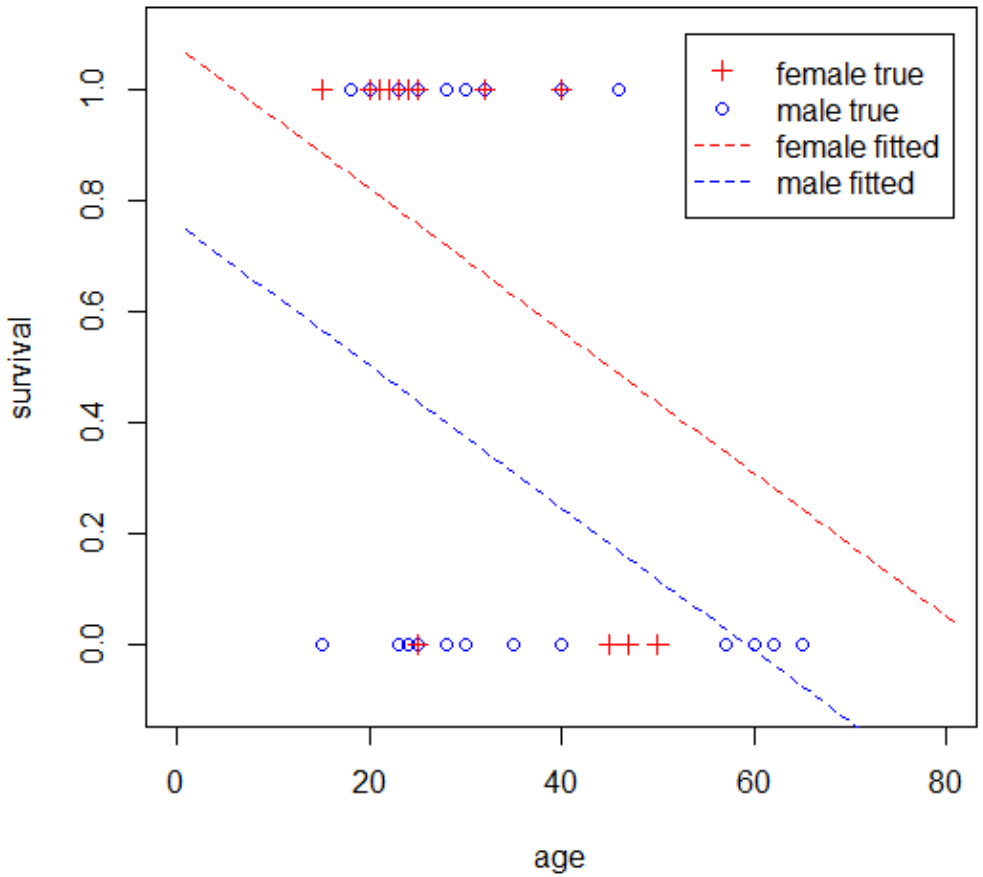
- 若將我們的資料放入 ordinary regression 模型，會得到

$$survival = 0.746 - 0.013age + 0.319female \text{。}$$

雖然 $R^2 = 0.1642$ 不高，但兩個變數都顯著。

Ordinary regression 的問題

- 我們理當可以用迴歸模型來得到存活機率的「預測值」。
  - 但模型會告訴我們，80 歲男性存活機率為 $0.746 - 0.013 \times 80 = -0.294$。
- 通常 ordinary regression 都無法產生結果介於 0 到 1 之間的機率值。

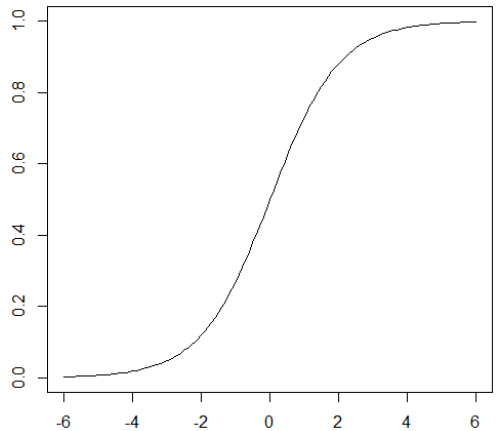


羅吉斯迴歸分析

- 正確的方式是使用羅吉斯迴歸分析。
- 在年齡存活範例中：
  - 我們仍猜測年齡較小較有可能存活，但年齡對存活機率的影響應該非線性。
  - 真實情況應是當一個人已經很年輕了，歲數再減幾歲也沒有什麼幫助。
  - 年齡降低的邊際效益 (marginal benefit) 遞減。
  - 年齡升高的邊際損失 (marginal loss) 也會遞減。
- 上述情況可以使用下列表達式表達

$$y = \frac{e^x}{1+e^x} \quad \Leftrightarrow \quad \log\left(\frac{y}{1-y}\right) = x$$

  - x 可以是介於 $(-\infty, \infty)$ 的任何值。
  - y 在 $[0, 1]$ 間。



羅吉斯迴歸分析

- 假設自變數 x_i 影響 $\pi = \Pr(y = 1)$：

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \text{。}$$

- 利用這樣的羅吉斯迴歸模型，會得到迴歸分析如下：

	預估值	標準差	z 值	p 值	
age	−0.078	0.037	−2.097	0.036	*
female	1.597	0.755	2.114	0.035	*

- 兩個變數都顯著。

羅吉斯迴歸分析曲線

- 模型給我們的預期曲線為

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.633 - 0.078age + 1.597female$$

或者

$$\pi = \frac{\exp(1.633 - 0.078age + 1.597female)}{1 + \exp(1.633 - 0.078age + 1.597female)},$$

其中 $\exp(z) = e^z$。

羅吉斯迴歸曲線

- 模型給我們合理的預測值。
- 對 80 歲男人，π 是

$$\frac{\exp(1.633-0.078\times 80)}{1+\exp(1.633-0.078\times 80)}$$

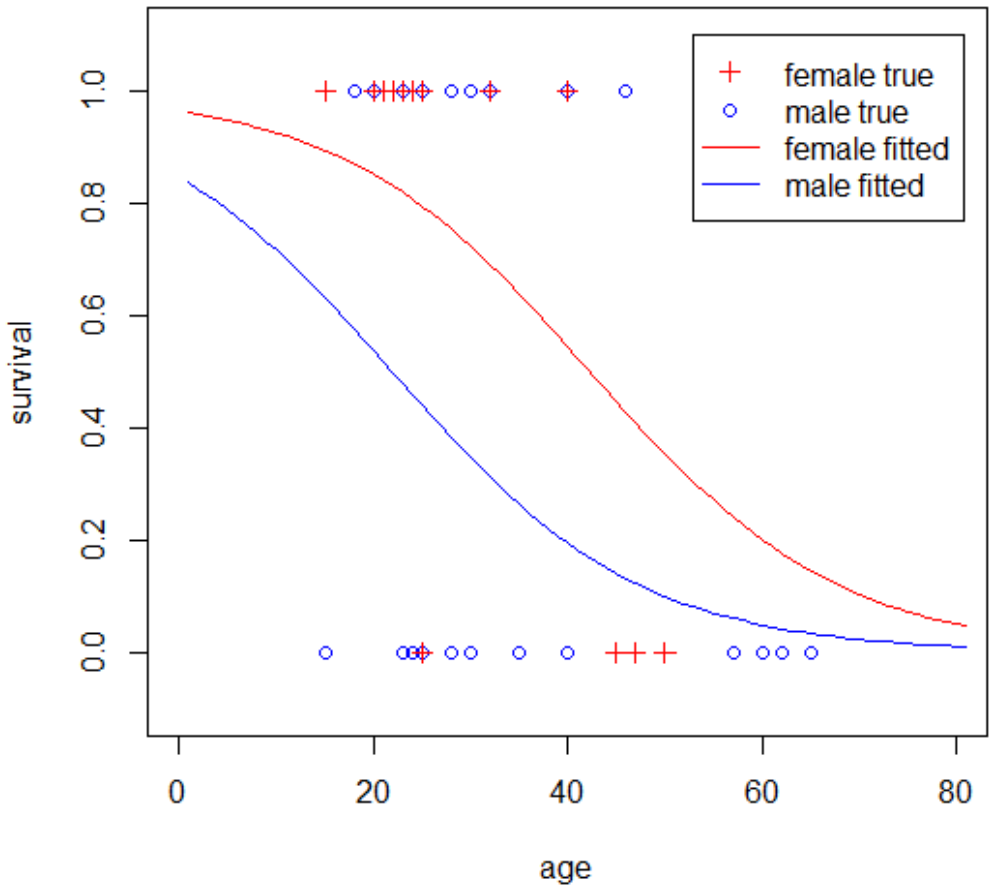
也就是 0.0097。

- 對 60 歲女人，π 是

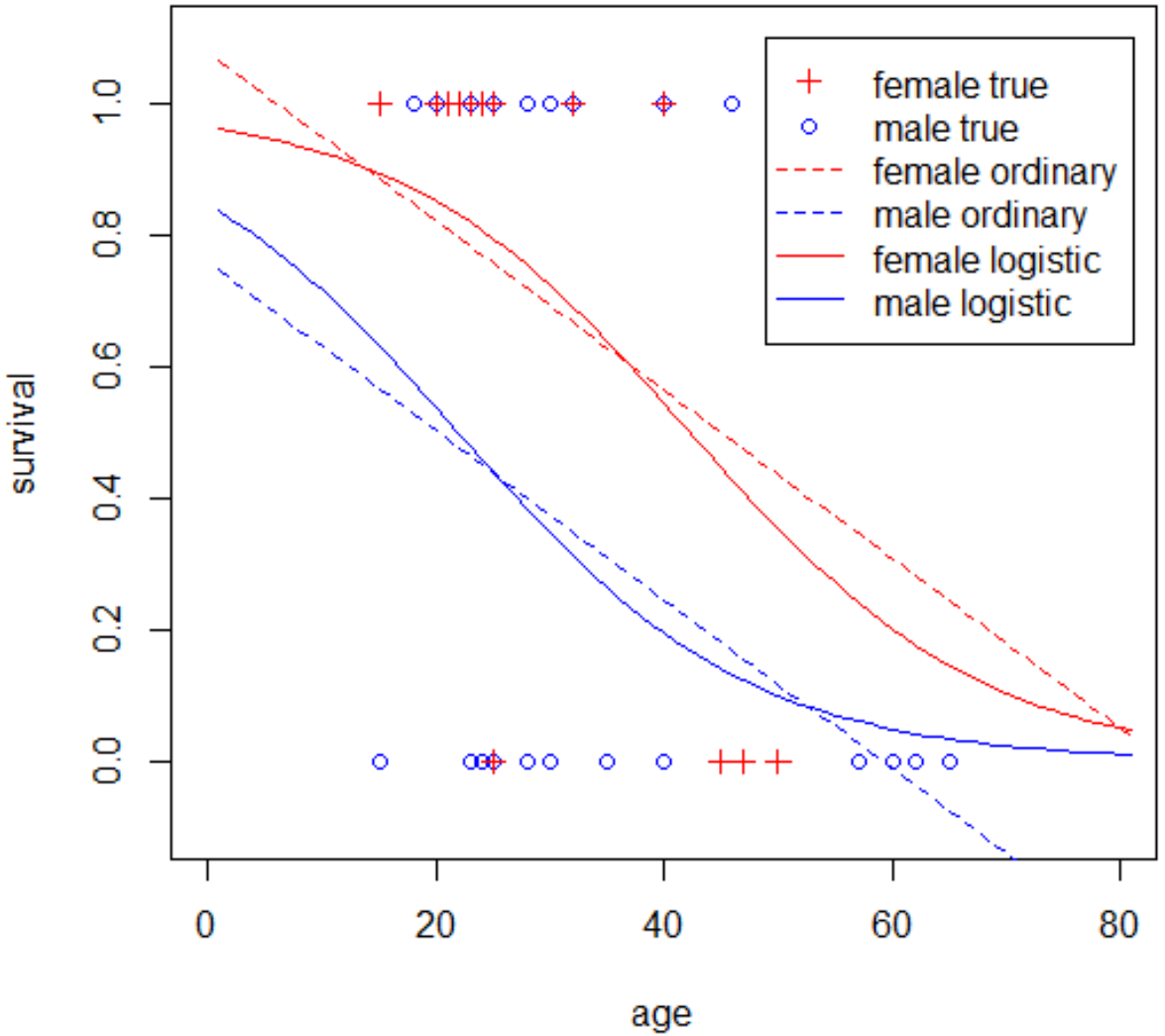
$$\frac{\exp(1.633-0.078\times 60+1.597)}{1+\exp(1.633-0.078\times 60+1.597)}$$

也就是 0.1882。

- π 永遠在 $[0,1]$。用 π 來當存活機率是沒有問題的。



比較


female true
male true
female ordinary
male ordinary
female logistic
male logistic
survival
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
age

模型的詮釋

- 模型給我們的預估曲線為

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.633 - 0.078age + 1.597female。$$

這裡面有任何意涵嗎？

  - $-0.078age$：年輕人較易存活。
  - $1.597female$：女人較易存活。

- 一般而言：
  - 使用p 值決定變數的顯著性。
  - 使用係數正負號做質化詮釋。
  - 使用預估曲線做量化預測。

模型選擇

- 回想在順序尺度應變數迴歸模型中，我們使用 R^2 與 R^2_{adj} 去評估模型有效程度。
- 在羅吉斯迴歸中，我們沒有這兩個值；我們使用 deviance。
  - 在迴歸模型報告中，null deviance 可被視為是不使用任何自變數時的總估計誤差。
  - residual deviance 可被視為是使用選定的自變數時的總估計誤差。
  - 理想上，residual deviance 應該愈小愈好。

在迴歸報告中的 deviance

- 報告中有 null deviance 與 residual deviance 的值。
- 針對 `glm(d$survival ~ d$age + d$female, binomial)`，我們有

```
    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
```

- 讓我們嘗試幾個其他模型：

自變數	Null deviance	Residual deviance
age	61.827	56.291
$female$	61.827	57.286
age, $female$	61.827	51.256
age, $female$, $age \times female$	61.827	47.346

  - 單獨只使用 age 比使用 $female$ 為佳。
- 如何比較不同變數個數的模型?

在迴歸報告中的 deviance

- 加入變數永遠會減少 residual deviance。當變數數量不同，我們可以使用 Akaike Information Criterion (AIC) 來比較模型。
- AIC 也在迴歸報告中有呈現出來:

自變數	Null deviance	Residual deviance	AIC
age	61.827	56.291	60.291
$female$	61.827	57.286	61.291
age, $female$	61.827	51.256	57.256
age, $female$, $age \times female$	61.827	47.346	55.346

- AIC 只用來比較互為巢狀關係 (nested) 的模型。
  - 若一個模型的變數是另一個模型的變數的子集合，則兩個模型互為巢狀。
  - 模型 4 優於模型 3 (基於 AIC 值做判斷)。
  - 模型 3 優於模型 1 或 2 (基於 AIC 值做判斷)。
  - 模型 1 與模型 2 不能被用 AIC 比較。