

第四章

隨機樣本及其他

4.6 敘述統計

在統計裡，我們常需要**收集資料**。資料收集到後便要做**資料分析**(data analysis)。由於資料常是以數字的形式出現，所以英文裡的data，在中文裡，除了稱為資料外，也常稱為**數據**。所謂讓數字說話，就是指對所收集到的資料，整理並分析，以提供有用的資訊。

對資料所做的初步整理，是屬於**敘述統計**(descriptive statistics)的工作。本書主要介紹**推論統計**(inferential statistics)。不過敘述統計是一種較直觀的來了解數據的方式。諸如給出樣本平均及樣本變異數，也都是屬於敘述統計的範疇。但以此二者，不見得是最好的了解數據的方式，往往也失之於簡略。適當地以圖形來展示資料，可以讓我們對資料更了解。有時圖形可以告訴我們數據裡隱藏的故事，引導我們進一步探討的步驟。所以統計圖形在資料分析的前置作業中，扮演著重要的角色。

本節我們介紹幾個簡單的統計圖示法。圖示法很多，並不限我們所介紹者。

4.6.1 點圖

如果資料數目並非很多，則可以點圖(dot plot, 或dot diagram)來了解資料。將數據以點表示在 x 軸上方，若有重複的值，則將點往上堆。例如，設有數據

28, 36, 37, 52, 36, 45, 39, 45, 38, 35, 36,

則其點圖如圖6.1。點圖可以讓我們對數據的散佈情況，有一粗略的概念。

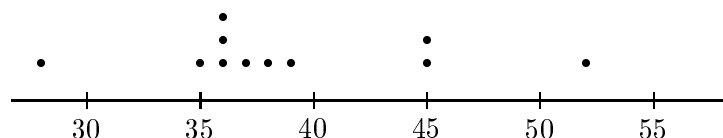


圖6.1 點圖

4.6.2 莖葉圖

仍以數據來說明。假設有下述數據：

9.5 10.8 8.8 11.2 10.2 10.3 10.2 11.3 10.0 8.8
 10.7 9.9 11.4 9.8 10.5 9.8 9.9 10.9 8.1 10.5
 10.6 8.2 8.6 9.2 9.9 10.0 11.0 9.2 10.7 10.9

莖葉圖(stem-and-leaf diagram)於選定適當的莖後，便可表示出數據。見圖6.2。

莖	葉
8	8 8 1 2 6
9	5 9 8 8 9 2 9 2
10	8 2 3 2 0 7 5 9 5 6 0 7 9
11	2 3 4 0

圖6.2 莖葉圖

點圖與莖葉圖其實未減縮任何資料，只是將資料以不同的方式呈現，以看得更清楚。

4.6.3 盒鬚圖

盒鬚圖(又稱**長鬚圖**, box-and-whisker plot, 簡稱boxplot), 是一很有效的表示資料的方法, 很多計算機程式中皆有此功能。盒鬚圖可用來了解資料的**偏斜性**(skewness)及**離群值**(outliers)。圖6.3為一典型的盒鬚圖。

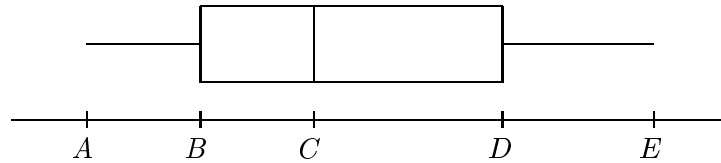


圖6.3 盒鬚圖

圖6.3中, A 稱為資料的最小值, E 稱為最大值, B 與 D 則分別為資料的下四分位數及上四分位數, 因此圖中盒子(盒子之高度並不重要)包含資料的中間50%部分。又 C 為資料的中位數。圖6.3包含一個盒子, 及二凸出來的鬚($\overline{AB}$ 及 $\overline{DE}$ 上方二線段), 這是此圖命名的由來。

樣本的中位數, 上四分位數及下四分位數, 以及一般的第 p 分位數, 皆可與分佈的分位數對應。設有數據 $y_1, \dots, y_n$, 將其按小至大排列, 而得 $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ 。關於**樣本第 p 分位數**(或說**第 p 樣本分位數**), 不同的書有不同的定義。有些書是定義為 $y_{([np]+1)}$, 其中 $[\cdot]$ 為最大整數函數。此定義雖然簡單, 但缺點為若 p 與 r 很接近, 則第 p 分位數與第 r 分位數可能會相等。另外, 亦可將樣本的第 p 分位數定義為

$$(6.1) \quad y_{([(n+1)p])} + ((n+1)p - [(n+1)p])(y_{([(n+1)p]+1)} - y_{([(n+1)p])}).$$

這其實是線性內插法的概念。我們以下例來說明。

例6.1 設有數據8, 20, 5, 15。求樣本之第0.37分位數。

解. 先將樣本按小至大排列, 而得 $y_{(1)} = 5$, $y_{(2)} = 8$, $y_{(3)} = 15$, $y_{(4)} =$

4 第四章 隨機樣本及其他

20。因 $n = 4$, $p = 0.37$, $[np] + 1 = [4 \cdot 0.37] + 1 = [1.48] + 1 = 2$ 。故以第一種定義所得之第0.37樣本分位數為 $y_{(2)} = 8$ 。又若 $p = 0.4$, 則 $[np] + 1 = 2$, 故第0.4樣本分位數亦為 $y_{(2)} = 8$ 。事實上從第0.25至第0.49樣本分位數皆為8。

現在來看第二種定義。四個樣本點, 將樣本的值域分成五個區間, 此四點便對應第0.2, 0.4, 0.6及0.8分位數。第0.37分位數介於第0.2及0.4分位數間。而 $y_{(1)} = 5$, $y_{(2)} = 8$, 因此第0.37樣本分位數為

$$5 + \frac{0.37 - 0.20}{0.40 - 0.20}(8 - 5) = 5 + 0.85 \cdot 3 = 7.55。$$

第二種定義的優點是不同的 p 值, 有不同的第 p 分位數。缺點是 p 不能很靠近0或1。此因若 $p < 1/(n+1)$, 則 $[(n+1)p] = 0$, 但我們並沒有 $y_{(0)}$ 的定義。又當 $p > n/(n+1)$, 則 $[(n+1)p] + 1 > n$, 因此 $y_{([(n+1)p]+1)}$ 也不存在。

當 $p = 1/2$, (6.1)式導致樣本中位數為

$$\begin{aligned} & y_{([(n+1)/2])} + \left(\frac{n+1}{2} - \left[\frac{n+1}{2}\right]\right)(y_{([(n+1)/2]+1}) - y_{([(n+1)/2]}) \\ &= \begin{cases} y_{((n+1)/2)} & , n \text{ 為奇數,} \\ (y_{(n/2)} + y_{(n/2+1)})/2 & , n \text{ 為偶數。} \end{cases} \end{aligned}$$

此與4.5節對樣本中位數所給的定義相符合。

底下我們對樣本第 p 分位數採第二種定義。如此樣本的第 $i/(n+1)$ 分位數, 即樣本的第 i 個順序統計量 $y_{(i)}$ (習題第4題)。

例6.2 瑞萊曾由不同的來源量測標準體積中氮(nitrogen)的重量(見Tukey (1977)), 而得到下述數據:

2.30143, 2.29816, 2.30182, 2.29890, 2.31017,
2.30986, 2.31010, 2.31001, 2.29889, 2.29940,
2.29849, 2.29889, 2.31024, 2.31030, 2.31028。

將數據表至小數第3位(四捨五入), 則圖6.4a, b分別為點圖及盒鬚圖。

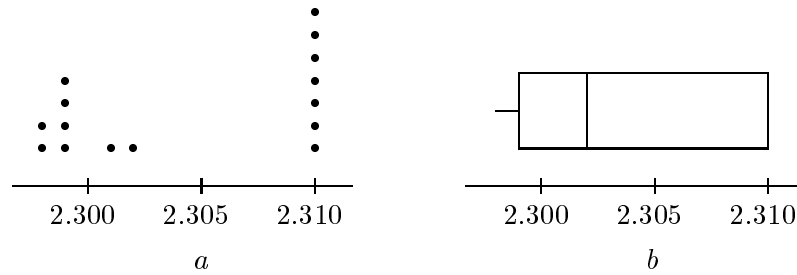


圖6.4 圖a及圖b分別為例6.2中之點圖及盒鬚圖

4.6.4 分位圖

利用樣本分位數來繪資料的圖，稱為**分位-分位圖**(quantile-quantile plot, 又稱Q-Q plot), 簡稱**Q-Q圖**。這是以圖形來檢驗數據是否來自某一分佈的方法。

以**常態分佈的Q-Q圖**(所謂normal quantile-quantile plot, 簡稱normal Q-Q plot, 有時也稱normal quantile plot)為例。在實際的應用裡, 我們常會想知道收集到的數據是否來自常態分佈的母體。在第十章我們會引進較客觀的檢驗方法, 此處我們看如何由圖形來做初步的檢驗。先選取一些適當的 p , 再以(6.1)式得到樣本之第 p 分位數 y_p , 然後在座標平面上繪出所有 (x_p, y_p) , 其中 x_p 為 $\mathcal{N}(0, 1)$ 分佈之 p 分位數。如果樣本是來自常態分佈的母體, 則其Q-Q圖近似於一條直線, 否則便不是。這是一種簡易地檢驗**常態性**(normality)的方法。

如果有 n 個樣本, 通常取 $p_i = i/(n+1)$, $i = 1, \dots, n$, 則 $y_{p_i} = y_{(i)}$, 且常態分佈的Q-Q圖, 即所有點 $(x_{p_i}, y_{(i)})$, $i = 1, \dots, n$, 其中 $x_{p_i} = \Phi^{-1}(p_i)$ 。

圖6.5a, b分別為20個來自 $\mathcal{E}(1)$ 及 $\mathcal{N}(0, 1)$ 分佈的樣本點之常態分佈的Q-Q圖, 二者似乎都不像是直線。但若各產生300個樣本點, 並繪於圖6.5c及d, 則圖6.5(d)就像一條直線了。這樣的以視覺來判定點是否在一直線上, 難免失之於主觀。但圖示法是一種初步的檢定。

例6.3 設有9個樣本點

2.33, 5.69, 0.39, -2.28, -2.14, 3.31, -0.72, 1.63, -3.91。

我們想看是否源自於常態分佈。

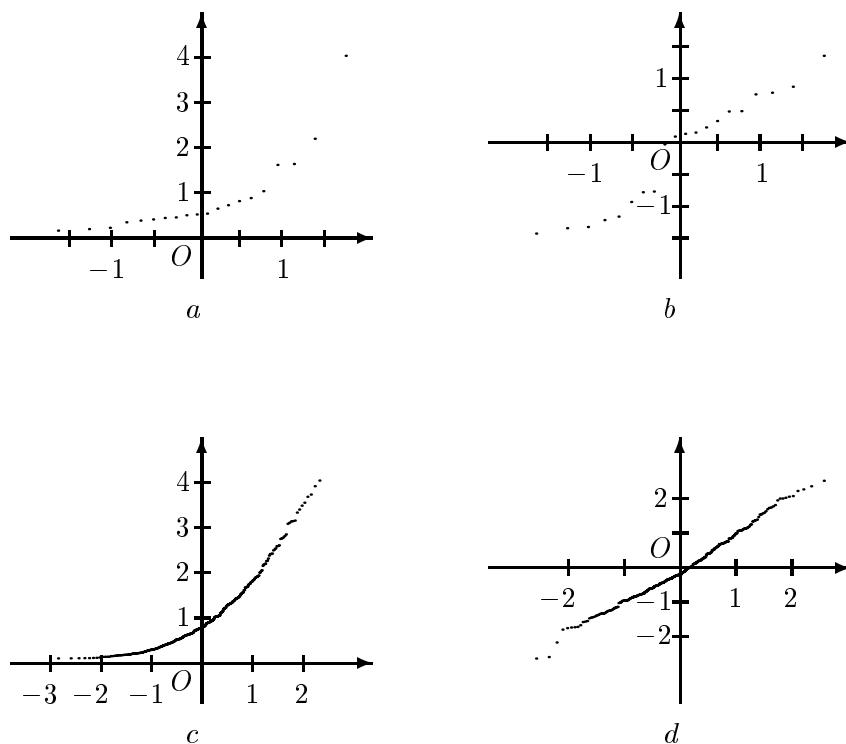


圖6.5 圖 a, b 分別為 $\mathcal{E}(1)$ 及 $\mathcal{N}(0, 1)$ 分佈20個樣本點之常態分佈的Q-Q圖,
圖 c, d 則為300個樣本點之常態分佈的Q-Q圖

由於樣本數 $n = 9$, 我們給出 $(x_i, y_i), i = 1, \dots, 9$, 之圖, 其中 x_i 為 $\mathcal{N}(0, 1)$ 分佈之第 $i/10$ 分位數, 即

$$P(Z \leq x_i) = \frac{i}{10}, i = 1, \dots, 9,$$

y_i 為樣本第 i 個順序統計量。如此得下述9點:

$$(-1.28, -3.91), (-0.84, -2.28), (-0.52, -2.14), (-0.25, -0.72), \\ (0, 0.39), (0.25, 1.63), (0.52, 2.33), (0.84, 3.31), (1.28, 5.69)。$$

圖6.6給出常態分佈之Q-Q圖, 大致是線性, 只有圖的右上方較偏離。這告訴我們說不定該進一步檢查資料是否登記有誤, 或是否資料無

誤，但有離群值存在。事實上上述數據為由 $\mathcal{N}(0, 9)$ 分佈所模擬產生，但將數據3.69改為5.69。若將5.69改回3.69，則 (x_i, y_i) , $i = 1, \dots, 9$ ，便大致落在直線 $y = 3x$ 上（對 $\mathcal{N}(\mu, \sigma^2)$ 分佈，其理想的Q-Q圖為直線 $y = \sigma x + \mu$ ）。

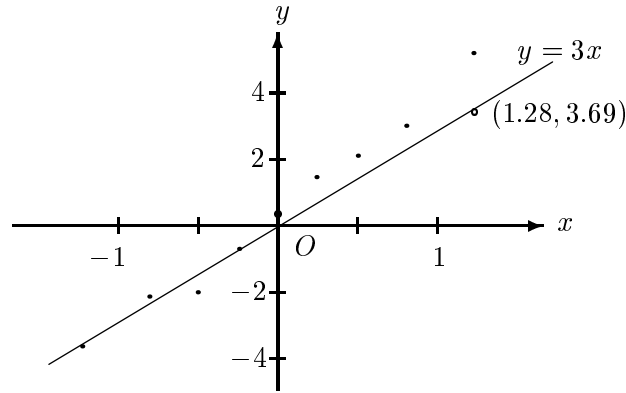


圖6.6 常態分佈之Q-Q圖

當有 n 個樣本點，有些書的常態分佈之Q-Q圖為繪點 $(x_{p_i}, y_{(i)})$, $i = 1, \dots, n$ ，其中 $p_i = (i - 1/2)/n$ ，而 x_{p_i} 為 $\Phi^{-1}(p_i)$ 。有些書則取 $p_i = (i - 3/8)/(n + 1/4)$ 。

如果是 $(p_i, y_{(i)})$, $i = 1, \dots, n$ ，的圖，便稱**分位圖**(quantile plot, 不過分位-分位圖有時也簡單的稱為分位圖)。其中 p_i 則如前通常取為 $i/(n + 1)$ (或 $(i - 1/2)/n$ ，或 $(i - 3/8)/(n + 1/4)$)。與盒鬚圖相比，分位圖展示所有的觀測，及較多的分位數。

有Q-Q圖亦有**P-P圖**(probability-probability plot)。仍以常態分佈為例，點 $(p_i, \Phi(y_{(i)}))$, $i = 1, \dots, n$ ，之圖形，即為P-P圖。又在上述Q-Q圖及P-P圖的定義中，常態分佈可改為其他分佈。如對 χ_r^2 分佈，則得 χ_r^2 之Q-Q圖及P-P圖等。

圖6.7為 $\mathcal{E}(1)$ 及 $\mathcal{N}(0, 1)$ 分佈之P-P圖。

4.6.5 直方圖

機率密度函數的直方圖，在1.4節已引進。對於數據，直方圖也是常用的圖示法。一般可將數據按數值相近而分成若干群。再將水平軸分成若

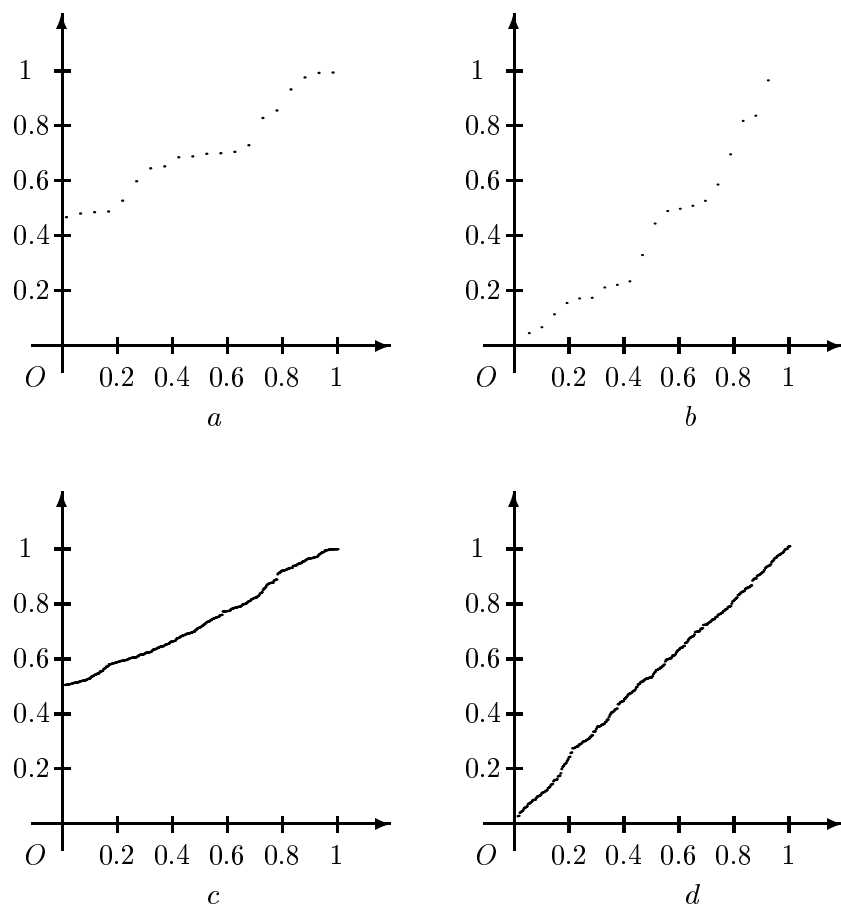


圖6.7 圖a, b分別為 $\mathcal{E}(1)$ 及 $\mathcal{N}(0, 1)$ 分佈20個樣本點之常態分佈的P-P圖,
圖c, d則為300個樣本點之常態分佈的P-P圖

于區間，每一區間的之中心為每一群之中點，半徑為每一群數據寬度之半。然後在該區間繪一長方形，長方形的面積與該群數據的頻率(即個數)成正比。

要注意的是直方圖是以面積來反映頻率大小，而非高度。而且每一長方形之底部寬度也可不同。又在繪直方圖時，有時難以決定的是該分幾群？每群之寬度是否要相同？要知不同的分法，影響到平均值及變異數等。

例6.4 一袋中有紙牌分別為 $0, 1, \dots, 9$ 。每次隨機地取一張, 每次取出後放回, 共363次。 $0, 1, \dots, 9$ 分別得到的次數為

35, 44, 25, 40, 50, 26, 33, 38, 37, 35。

圖6.8給出其直方圖。

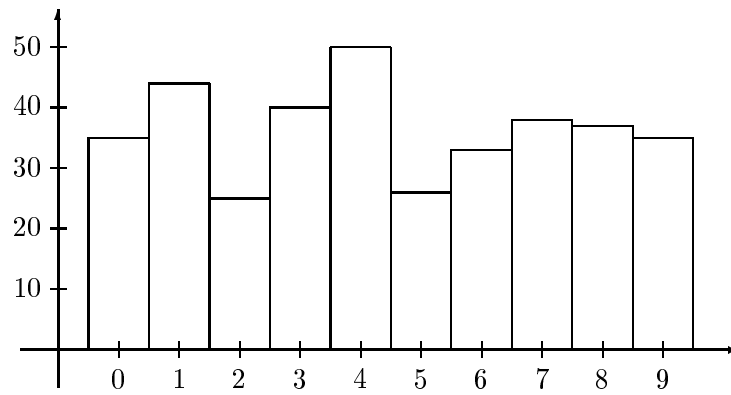


圖6.8 例6.4中之直方圖