眾數 = 30；

樣本中位數 = 33；

差距中點 $= \frac{1}{2}(20 + 98) = 59$

我們可從例 1.3-1 發現，有 8 位學生（約 72%）的數學期中考試成績不但低於平均值，且遠低於差距中點。由此例子看來，平均值與差距中點顯然未能正確反映數據的中心位置，它們反而給我們一個錯誤印象。那些不正常的極端大或極端小的資料值稱為**離群值** (outlier)。平均值與差距中點容易受樣本中一兩個比較極端的數值影響而增加或減少，但中位數較不受這些離群值影響。若數據嚴重不對稱，中位數是一個比較合適的中心位置統計量。

1.3.2 位置統計量

位置統計量 (location statistic) 是用來描述一組由小至大排列的數據的幾個重要位置，包括：最小值 (minimum)，最大值 (maximum)，第一四分位數 (first quartile) Q_1（又稱下四分位數，lower quartile），第二四分位數 (second quartile) Q_2 和第三四分位數 (third quartile) Q_3（又稱上四分位數， upper quartile）。它們被稱為一組數據的**五數綜合** (five-number summary)。

第一四分位數是中位數左邊（不含中位數）所有觀測值的中位數；第二四分位數即是中位數；第三四分位數是中位數右邊（不含中位數）所有觀測值的中位數。而四分位數間距 (inter-quartile range, *IQR*) 是指第一和第三四分位數兩者的差距，即 $Q_3 - Q_1$。而所有小於 $Q_1 - 1.5 \times IQR$ 或大於 $Q_3 + 1.5 \times IQR$ 的數值為離群值。

例 1.3-2 在例 1.3-1 中，最小值為 20，最大值為 98，第一四分位數 Q_1 為 24，第二四分位數 Q_2 為 33，第三四分位數 Q_3 為 42。所以，$IQR = 42 - 24 = 18$，$Q_1 - 1.5 \times IQR = -3$，$Q_3 + 1.5 \times IQR = 69$。因此數據中的 80 和 98 為離群值。

五數綜合可用一個**框線圖**（又稱為**盒鬚圖**或**箱形圖**，box plot / box-and-whisker diagram）來表達。框線圖並不顯示所有觀察值而只顯示五數綜合，它可以用作檢驗資料的離群值及分佈型態。如圖 1.13 所示，它是例 1.3-2 的框線圖，此圖可以鉛垂或水平的形式展示。

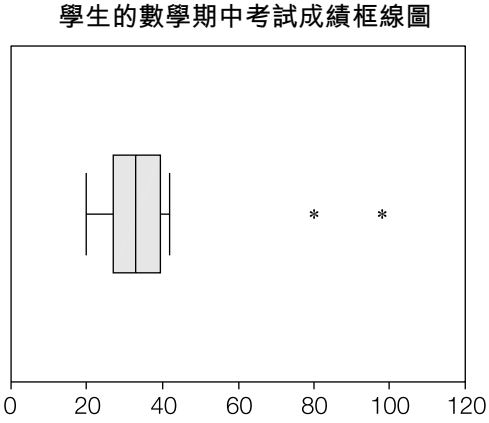

圖 1.13

框子的下界限是下四分位數，上界限是上四分位數，而框子的長度即代表四分位數間距。框中的鉛垂線「|」代表中位數，通常框內包含 50% 的觀察值，因此框子愈大，則表示觀察值散佈愈大。由框子上下界所延伸出來的線，稱為鬚 (whisker)，是用以連接非離群觀察值的最大值與最小值（上例中分別為 20 和 42）。此外，框線圖中亦可顯示出資料的離群值（以符號「*」表示）。

從框線圖可看出一組數據是對稱（兩邊鬚長度均等）、往右偏 (right-skewed)（右邊鬚較長）或往左偏 (left-skewed)（左邊鬚較長）。

1.3.3 離差統計量

除了數據的中心位置外，我們也要了解數據的分散幅度，這幅度稱為**離差** (dispersion)。

分佈域 (range) 是最簡單的離差統計量，通常以 R 表示，它是數據最大值及最小值之差。

2.1 引言

在上一章，我們已介紹如何把調查所得的資料表達出來，亦即是敍述統計。接下來，我們將討論推論統計。

推論統計是一門很有趣的課題，它是基於一些樣本數據作分析，從而推斷總體的參數，即是從僅有少量資料去推斷實際的情形。例如我們抽取 5 罐某牌子的汽水，量度它們的容量，發覺平均每罐的容量為 345 毫升，我們便推斷這牌子的汽水每罐容量為 345 毫升。至於這個結論是否合理，是否過於武斷，我們會在較後的章節中討論。

統計學上，任何一個記錄 —— 無論是數據抑或是現象，都稱為**觀察** (observation)。例如上述所量出 5 罐汽水的容量分別是 342、346、347、349、341（毫升），則這五個觀察便構成一個**觀察集** (set of observation)。又例如投擲硬幣 4 次，得出 H、T、T、T，其中 H 代表正面向上而 T 代表反面向上，那麼 {H, T, T, T} 是一個觀察集（這是一個**重集** (multi-set)，即其元素縱然相同也看成不一樣，因為每一個元素代表不同的投擲）。所有可以產生觀察的過程稱為**試驗** (experiment)。投擲硬幣是一個試驗，觀察的結果只可能有兩個 —— 正面向上或反面向上，這些結果我們稱為觀察值。街頭訪問市民對外傭加薪的意見也是一個試驗，但它的觀察值卻有很多。

在討論推論統計之前，我們必須了解一些基本知識，包括一些基本的數學語言及概念，也包括統計學最基本的概念 —— 概率。

2.2 集合

集合是數學最基本的概念，它是用來描述或規限我們所討論事物的範圍。在這書中我們不會對集合論作很深入的討論，而只會介紹一些簡單的定義和作一些集合的基本運算。

簡單地說，集合是一個包含一些「物件」的個體，而這個體可以清晰地界定那些東西是屬於它的。19 世紀中葉的德國數學家康托爾 (Cantor, 1845 – 1918) 曾為集合下了一個定義，他認為當我們把一些清晰可分的事物看成一體時，這整體便稱為**集合** (set)。

我們通常以大階英文字母，如 A，B，C，· · · 等來代表集合，而以小階英文字母，如 a，b，c，· · · 等代表「物件」。我們所說的「物件」是很廣泛的，它可以是數字或觀念，也可以是真實的物件。

定義 2.2.1

設 A 為集合，物件 a 在 A 中稱為 A 的**元素** (element)，並記作 $a \in A$，讀作 a 屬於 A。若物件 b 不是 A 的元素，則記作 $b \notin A$。

定義 2.2.2

一個集合 A 包含有限個（不同的）元素稱為**有限集** (finite set)，否則稱為**無限集** (infinite set)。我們以 $|A|$ 代表 A 所含元素的個數，若 A 為無限集，則 $|A| = \infty$。

我們一般以一對大括號「{}」表示一個集合，並把集合的元素放進去。例如 V = {a, e, i, o, u} 是英文字母所有元音的集合。把一個集合的所有元素表列出來當然是最清楚，但有時是不可能的。例如當一集合包含的元素太多甚至是無限個，這時我們便會用一種稱為**結構式**的方法來表達該集合。其形式如下：

$$\{x \mid P(x)\}$$

其中，$P(x)$ 是有關元素 x 的一些命題，這集合包含所有滿足 $P(x)$ 的物件。例如 $A = \{n \mid n$ 是整數，$1 < n < 100\}$，當中 n 為 A 的任何一個元素，它的特性是介乎 1 及 100 之間的整數。換句話說，A 是由 2 至 99 所有整數的集合。注意，在結構式表示法中特性的部分，「，」代表「和」的意思。

教學活動 4.2

目標：探究以二項分佈模擬香港特別行政區行政長官的支持度

人數：25 人或以上

步驟：

1. 將同學分成六小組，每組同學負責訪問一個級別。
2. 在每個級別隨機抽出 40 名同學，詢問他們是否支持現任行政長官。把結果記錄如下表。

   年級：_____

意見	支持	不支持
頻數		

表 4.3

3. 比較不同級別對香港特別行政區行政長官的支持度。
4 作出對香港特別行政區行政長官的民間支持度調查報告。

進一步活動

坊間組織每隔一段時間便會作出對香港特別行政區行政長官的民間支持度調查報告，試找出最近一次的類似報告。當中的兩個參數 n 和 p 為何？

教學活動 4.3

目標：探究以正態分佈模擬學校學生的身高分佈

人數：25 人或以上

步驟：

1. 將同學分成六小組，每組同學負責一個級別。透過老師的協助，收集每個級別學生的身高數據，完成下表。

身高 (cm)										
頻數										

表 4.4

2. 將所得的數據製成相對頻數分佈圖。
3. 描述所得的分佈圖。
4. 所得的分佈圖是否接近某一個正態分佈？若是，對應的兩個參數 μ 和 σ 為何？
5. 比較不同級別的身高分佈。

進一步活動

試將所得的身高數據以性別區分，重複以上步驟，討論主要的分別。

若 $\alpha = 0.01$，則 $t_{16,\ 0.005} = 2.921$。另外，我們知道 $n = 18$，$\bar{x} = 46.33$，$\Sigma x_i^2 = 41\ 612$ 及 $\Sigma x_i = 834$（見例 8.2-3）。

由 (8-8) 式，我們可算出 β_0 的 99% 置信限為

$3.191 \pm 2.921 \times 5.770\ 310\ 33 \sqrt{\dfrac{41\ 612}{18 \times 41\ 612 - 834^2}}$，即

3.191 ± 14.8705。

也就是說 β_1 的 99% 置信區間為 (–11.6795, 18.0615)。

由 (8-5) 式，我們可算出 β_1 的 99% 置信限為

$0.549 \pm 2.921 \times \dfrac{5.770\ 310\ 33}{\sqrt{41\ 612 - 18 \times 46.33^2}}$，即 0.549 ± 0.309。

也就是說 β_1 的 99% 置信區間為 (0.240, 0.858)。

現在，我們嘗試用 0.01 的顯著水平，檢驗年齡會否影響人體脂肪比率。即我們想檢定 β_1 是否等於零，所以有以下的雙側檢驗。

設 $H_0: \beta_1 = 0$ 及 $H_1: \beta_1 \neq 0$。

若 H_0 為真，則由 (8-5) 式可算出 $T = \dfrac{b_1\sqrt{\Sigma x_i^2 - n\bar{x}^2}}{S} \sim t_{n-2}$。

因為是雙側檢驗，我們考慮 T 的絕對值，即

$$|t| = \left|\frac{0.549\sqrt{41\ 612 - 18 \times (46.33)^2}}{5.770\ 310\ 33}\right| \approx 5.1899 > t_{16,\ 0.005} = 2.921。$$

於是在 0.01 的顯著水平下，我們拒收 H_0，即年齡會影響人體脂肪比率。

我們也可以利用 β_1 的 99% 置信區間來決定接收或拒收 H_0。現在我們可注意到 0 落在 β_1 的 99% 置信區間 (0.240, 0.858) 之外。所以在 0.01 的顯著水平下，我們拒收 H_0，即年齡會影響人體脂肪比率。

8.5 相關分析

線性迴歸是把變量之間的線性關係表示出來，而**相關分析** (correlation analysis) 是把變量之間的線性關係之強度表現出來。在迴歸分析中，我們想把整個關係式估計出來，而相關分析則只想求出一個數字，從而表現出兩變量之間的線性關係，這個數字我們稱它為**線性相關係數** (linear correlation coefficient)，記作 ρ。例如若要知道香港恆生指數和美國杜瓊斯指數的關係有多強便可應用線性相關係數。

ρ 是用來形容變量 X 及 Y 之間的線性關係強度，它的定義為

$$\rho = \frac{\frac{1}{N}\Sigma(x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

其中 N 是總體容量，μ_X 及 μ_Y 分別為 X 及 Y 的總體平均值，σ_X 及 σ_Y 分別為 X 及 Y 的總體標準差。

$\text{Cov}(X, Y) = \frac{1}{N}\Sigma(x_i - \mu_X)(y_i - \mu_Y)$，稱為 X 及 Y 的總體**協方差** (covariance)。

估計總體相關係數 ρ 的統計量通常用

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

$$= \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}\frac{\Sigma(y_i - \bar{y})^2}{n-1}}}$$

$$= \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{S_X S_Y}$$

其中 n 為樣本容量，S_X 及 S_Y 分別為 X 及 Y 的樣本標準差。