前言

資料科學是一項令人興奮的工作，要求從混亂的資料中提取洞見的能力，對於商業、醫療、政策等各類決策皆具高度價值。本書《資料科學學習手冊》的目標，是讓讀者具備從事資料科學工作的能力，為了達成這個目標，本書設計以下幾項特色：

著重基礎概念

科技日新月異，工具推陳出新。儘管本書採用特定技術，但我們的目標仍是讓讀者掌握資料科學的基本構成要素。本書藉由揭示思考資料科學的問題與挑戰，並介紹各項技術背後的基本原理，使讀者即使在技術更迭之際，仍能從中獲益。

涵蓋完整的資料科學生命週期

本書並非僅聚焦於例如操作資料表格、應用機器學習技術等單一主題，而是涵蓋完整的資料科學生命週期，從提出問題、取得資料、理解資料，到理解這個世界。對資料科學家而言，完整執行這整個流程，往往是最具挑戰性的部分。

使用真實資料

為了讓讀者能為真實世界的問題做好準備，本書認為有必要從使用真實資料的範例中學習，即使資料並不完美。本書所使用的資料集經過精心挑選，來自實際進行且具有影響力的資料分析，而非過度清理或人工合成的資料。

透過案例研究應用概念

書中納入多篇延伸案例研究，這些案例不是來自其他資料科學家的實際分析，就是從其基礎上擴充，藉此展示在實際情境中應用資料科學生命週期的各個階段。

結合運算與推論思維

在工作實務中，資料科學家必須能預見撰寫程式時所做的決策，及資料集大小可能對統計分析造成的影響。為了使讀者能應對未來的工作挑戰，《資料科學學習手冊》融合運算思維與統計思維，並以模擬實驗而非數學證明的方式說明統計概念。

本書的文字與程式碼皆為開放原始碼，可在 GitHub（*https://github.com/DS-100/textbook/*）上取得。

預期背景知識

閱讀本書前，讀者應已熟悉 Python，並了解使用內建資料結構如串列（list）、字典（dictionary）與集合（set）的方式；也能夠匯入並使用其他套件中的函數與類別；並能從頭撰寫函數。本書亦直接使用 `numpy` 套件，不會詳細介紹，也不要求讀者具備太多使用經驗。

若讀者具備一些機率、微積分與線性代數的基礎，將能從本書得到更多收穫，但本書仍會努力於以直觀方式說明數學概念。

本書架構

本書共包含 21 章，分為六個部分：

第一部分（第 *1–5* 章）

說明資料科學生命週期的概念，並以基礎等級完整走過一次生命週期，介紹全書會使用的術語，最後以一則公車到站時間的簡短案例收尾。

第二部分（第 *6–7* 章）

介紹資料框架（dataframe）與關聯（relation），以及使用 `pandas` 和 SQL 撰寫程式來操作資料。

第三部分（第 *8–12* 章）

專注於取得資料、發掘其特徵並找出其中問題的方式。理解這些概念後，讀者便能拿出一份資料檔案並向他人說明該資料集的有趣特性。最後以空氣品質的案例收尾。

第四部分（第 *13–14* 章）

探討常見的替代資料來源，如文字、二進位格式與來自網路的資料。

第五部分（第 *15–18* 章）

著重於運用資料來理解世界，內容涵蓋推論相關主題，如信賴區間與假設檢定，以及模型擬合、特徵工程與模型選擇；最後以肯亞獸醫預測驢子體重的案例收尾。

第六部分（第 *19–21* 章）

完成對監督式學習的探討，涵蓋邏輯斯迴歸與最佳化。最終也以案例收尾，說明預測新聞報導中真實或虛假陳述辦法。

本書結尾附有延伸學習資源，讀者可進一步深入本書所介紹的多項主題，同時也列出書中所使用的全部資料集。

本書編排慣例

本書使用以下排版慣例：

斜體字（*Italic*）

表示新術語、網址、電子郵件地址、檔名與副檔名。中文以楷體字呈現，其對應的英文則以斜體表示。

等寬字體（`Constant width`）

用於程式碼範例，亦用於段落中提及的程式元素，例如變數或函數名稱、資料庫、資料型別、環境變數、敘述與關鍵字。

粗體等寬字體（**`Constant width bold`**）

表示使用者應照字面輸入的指令或文字。

斜體等寬字體（*`Constant width italic`*）

表示需由使用者提供的值，或根據情境決定的值。

範例：處理餐廳安全違規資料

現在以能展示許多資料整理技巧的範例作為本章總結。回想第 8 章提到，舊金山的餐廳稽查資料儲存在三個資料表中，分別是商家／餐廳資訊：`bus`、稽查資訊：`insp` 以及安全違規事項：`viol`。其中 `viol` 資料集記錄稽查期間發現的詳細違規描述，我們希望擷取這些資訊的一部分，並將其與 `insp` 資料表中的稽查分數連結起來，這樣可以讓違規資訊與個別稽查結果結合。

目標是要了解哪些類型的違規事項會造成較低的餐廳安全評分。這個範例涵蓋數個與結構轉換有關的重要概念：

- 過濾（filtering）：針對特定區段的資料分析
- 匯總（aggregation）：改變資料表的粒度
- 合併（joining）：整合分散在不同表格的資訊

此外，此範例也會示範將文字資料轉換為可供分析的數值型量值辦法。

第一步先簡化結構，只保留某一年內的稽查資料，原始資料中包含四年紀錄。以下程式碼會統計稽查表中每年的資料筆數：

```
pd.value_counts(insp['year'])

year
2016    5443
2017    5166
2015    3305
2018     308
Name: count, dtype: int64
```

將資料限制在某一年，可以讓分析更加簡潔。當然，之後若有需要，也可以再回頭納入所有四年的資料。

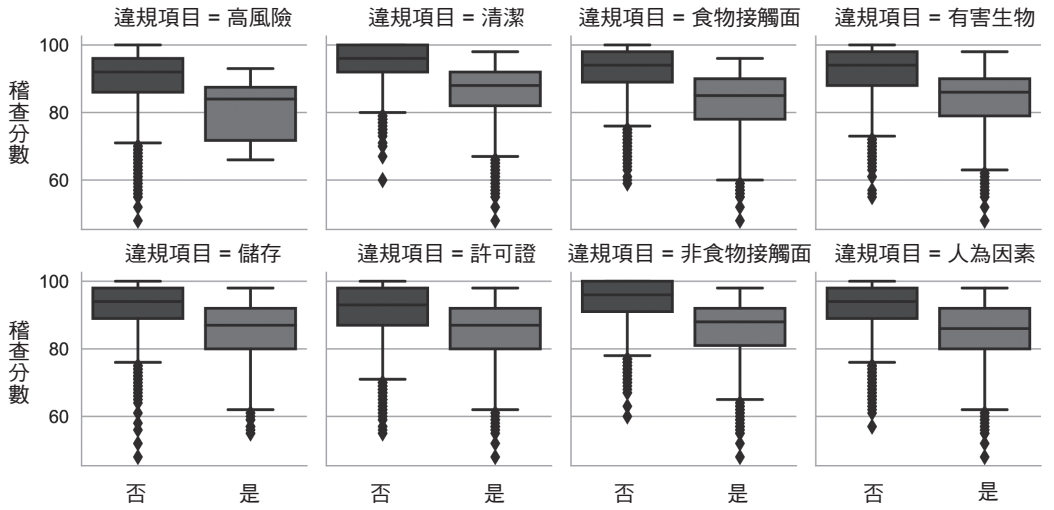
聚焦與匯總違規項目

將資料範圍限定在 2016 年的餐廳稽查資料，並使用 `pipe` 函數對稽查和違規這兩個資料框架套用相同的過濾條件：

```
def subset_2016(df):
    return df.query('year == 2016')

vio2016 = viol.pipe(subset_2016)
```

為了檢視每類違規與稽查分數之間的關係，可繪製一系列箱型圖（box plot），比較是否存在該違規類型時的分數分布情況。由於此處重點在於資料的樣式，而非視覺化程式碼，因此程式碼部分略去；可於線上檢視完整版本（*https://oreil.ly/go29H*）。



總結

資料整理是資料分析中不可或缺的一環。若缺乏這個過程，容易忽略資料中的問題，而這些問題可能對後續分析造成重大影響。本章涵蓋幾個幾乎每次分析都會用到的重要資料整理步驟。

本章說明將資料讀入資料框架後應檢查的內容。品質檢查有助於發現資料中的問題，為了找出錯誤與缺漏值，可採用多種方式：

- 檢查摘要統計量、分布情形與數值計數，第 10 章將提供更多使用視覺化與摘要統計量來檢查資料品質的範例與指引。本章簡要提及幾種方法。針對某個欄位列出唯一值的計數表，可以揭示意料之外的編碼與失衡的分布，例如某個極為罕見的值。百分位數對於揭露極高或極低值的比例也十分有用。
- 使用邏輯運算式來辨識欄位值超出合理範圍，或相關欄位之間的關係不合邏輯紀錄。簡單地計算不通過品質檢查的紀錄數，就能迅速了解問題的規模。

加入情境資訊

本章在圖表中使用文字，以提供有意義的座標軸標籤，包含計量單位、類別的刻度標籤與標題，分享視覺化圖表時這是一個很好的做法。理想目標是讓圖表本身就能表達足夠的情境資訊，也就是說，讀者應該可以在不用查閱其他說明的情況下，就能大致了解圖表內容。儘管如此，統計圖表中的每個元素都應具有特定用途，多餘的文字或圖表元素通常會稱為**圖表雜訊**（*chartjunk*），應該予以剔除。本節簡要介紹幾種可為圖表增添有用情境資訊的方式，並舉例說明加入情境資訊以建立具備出版品質的圖表方式。

文字情境資訊包括**標籤**（*label*）與**圖說**（*caption*）。持續地為刻度與座標軸使用具資訊性的標籤實務上是很好的做法，例如，座標軸標籤通常會因加入計量單位而更加清楚。圖表應在需要時包含標題與圖例，對於那些需要觀看與解讀的圖表來說，資訊性的標籤尤其重要；然而，即便僅為個人進行探索性資料分析，也常希望圖表保有足夠情境資訊，以便日後回顧分析時，能輕鬆辨識圖表的內容。

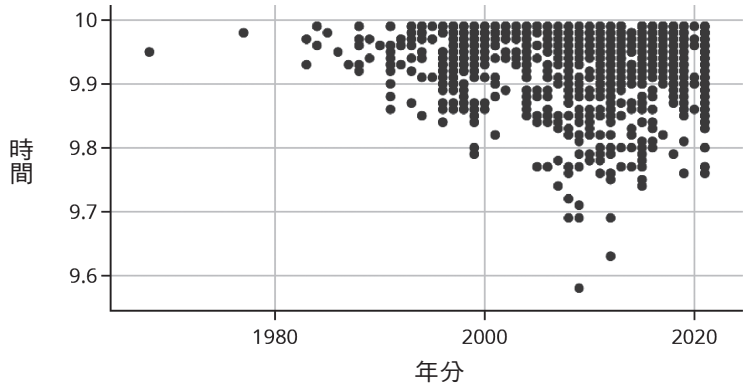
圖說具有多重功能，能描述圖表所繪內容、為讀者提供導引；圖說也能指出圖表中的重點特徵並評論其意涵。重複文本資訊的圖說也並無不可。因為讀者在瀏覽出版物時，經常只關注章節標題與圖表，因此圖說應具有足夠相關資訊。

參考標記（*reference marker*）則為圖表區域增添額外情境資訊，提供基準點、歷史數值或其他外部資訊的參考點與參考線，有助於讀者比較與解讀。例如，常在分位數 - 分位數圖（quantile-quantile plot）中加入斜率為 1 的參考線。亦可在時間序列圖中加上垂直線來標示特定事件，例如自然災害。

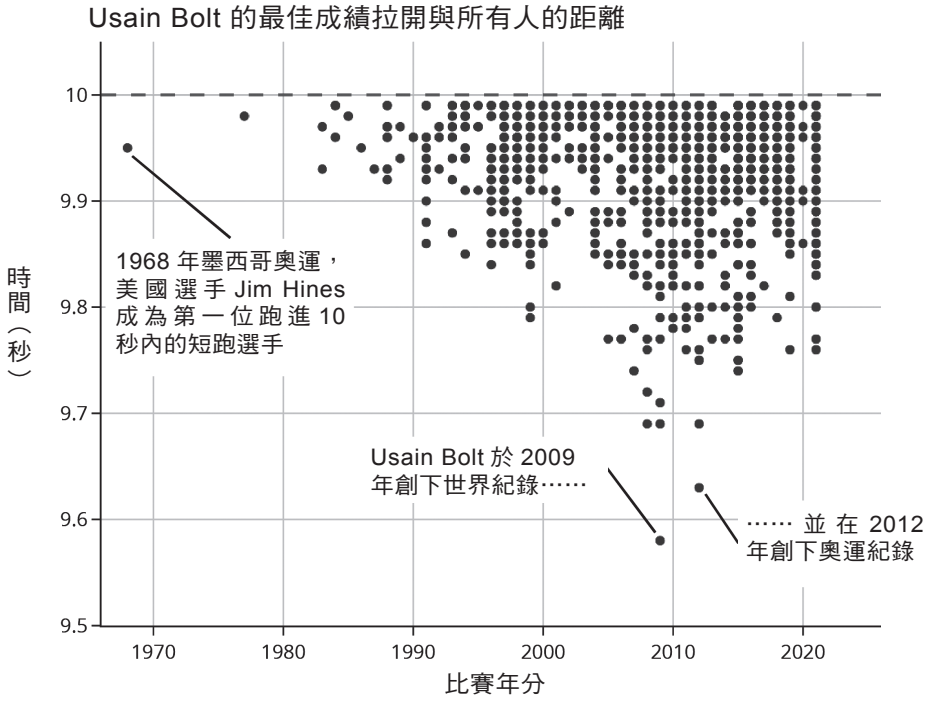
以下範例將展示將這些情境元素加入圖表的方式。

範例：100 公尺短跑成績

下圖顯示自 1968 年以來男子 100 公尺短跑的比賽成績，這些資料僅包含使用電子計時、於戶外正常風速條件下舉行的比賽，且僅納入跑進 10 秒內的選手，這是一張基本散布圖，顯示比賽時間與年分之間的關係。從這張圖出發來進一步增補內容，以製作出 FiveThirtyEight（*https://oreil.ly/pxHr4*）報導 100 公尺短跑時所刊登的圖表：



想要為其他人準備圖表時，需要考量圖表所要傳達的重點。在此案例中，主要訊息有兩點：過去 50 年來最頂尖的選手速度越來越快，以及 Usain Bolt 在 2009 年創下的驚人紀錄 9.58 秒至今仍無人能破；事實上，第二快的成績也由 Bolt 所創。在加入標題以直接點出圖表主要訊息、在 y 軸標籤中標示單位，以及在散布圖上標註幾個關鍵點，包含 Bolt 的兩項最佳成績之後，為這張圖表補充不少情境脈絡。此外，也加入一條 10 秒的水平參考線，以說明圖表僅繪製 10 秒以內的成績，並以特殊符號標示世界紀錄，以吸引讀者注意這個關鍵點：



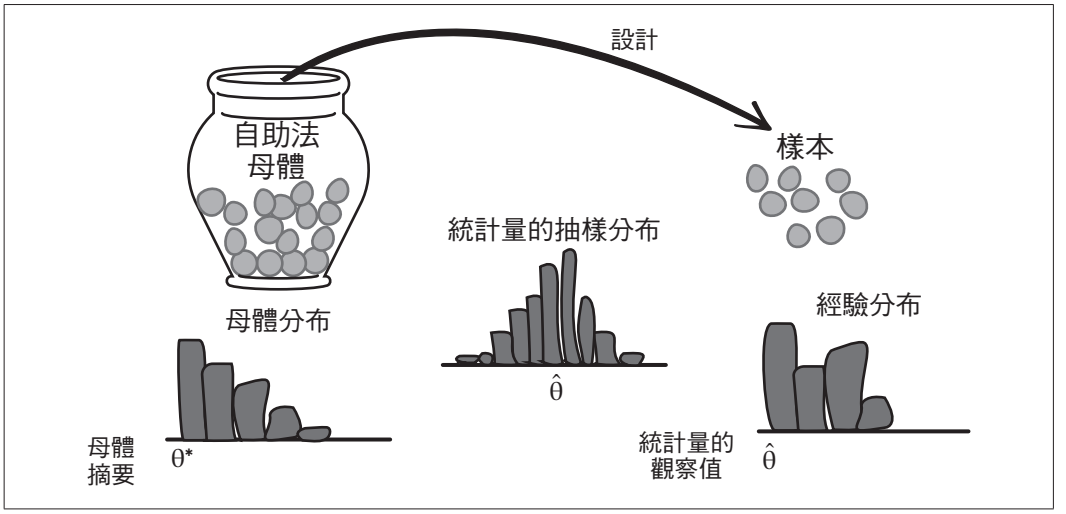


圖 17-1　資料產生過程的示意圖

為了簡化，這裡只考慮測量一個特徵。圖中瓶子下方是該特徵的**母體直方圖**（*population histogram*），代表整個母體中該特徵值的分布情形，最右側是實際樣本的**經驗直方圖**（*empirical histogram*），代表該特徵值在真實樣本中的分布情形。可注意到這兩個分布的形狀相似，這種情況會出現在抽樣機制能產生具代表性的樣本時。

我們經常關心的是樣本測量值的某個摘要統計，例如平均值、中位數或是簡單線性模型的斜率等等。這些摘要統計量（summary statistic）通常是用來估計某個母體參數，例如母體的平均值或中位數，圖中左側的母體參數記為 θ^*，右側則是從樣本中計算出的摘要統計量 $\hat{\theta}$。

產生樣本的隨機機制如果再執行一次，可能會給出另一組不同的資料；但如果實驗設計得當，則預期這個樣本仍能代表母體，換句話說，可以根據樣本計算的摘要統計量來推論母體參數。圖中央的**抽樣分布**（*sampling distribution*）是關於統計量的**機率分布**（*probability distribution*），代表對不同樣本所計算出的統計量可能出現的值以及各自機率。第 3 章已使用模擬法來估計多個例子的抽樣分布，本章將重訪這些例子，並對分析進行形式化。

最後再補充一點關於這三種直方圖的說明：如第 10 章所述，直方圖中的每個長條代表落入該區間的觀測值比例，在母體直方圖中，這是整個母體中該區間的比例；在經驗直方圖中，面積代表樣本落於該區間的比例；而在抽樣分布中，面積則代表資料產生機制產出某統計值落在該區間的機率。