序

我們每天都在使用網路搜尋、社交網路，以及智慧型手機的 app，但您知道它們每天都在被「最佳化」嗎？乍看很華麗的技術，譬如 Internet、SaaS、DX（Digital Transformation），背後都用到了所謂「網站最佳化」的技術，說它是「現代的改善法（Kaizen）」也不為過，也是超重要的隱形技術。

簡而言之，就是準備 A 和 B 兩個版本，用來測試使用者更喜歡哪個版本（A/B 測試），然後改善產品使其更符合使用者需求。這樣的事情每天都在以驚人的速度進行著，不僅是在 GAFA（Google、Apple、Facebook、Amazon）這樣的大公司，在國內的網路企業以及矽谷的新創公司內也都是如此。

一旦了解這項技術，就會知道「工作時總是無意識地去開 app」和「孩子總離不開智慧型手機」的原因。無論怎麼想著「要忍耐」還是會去碰觸。要憑自己的意志力去對抗這項「最佳化」，就像要跟拳擊世界冠軍對戰一樣困難，這項技術就是如此重要而強大。

目前很少有書籍對這項超重要的隱形技術做系統性的整理。本書解釋了該技術的數學背景，用易於理解的方式，以程式碼具體說明其作用。內容基於作者的博士論文並進一步加強，作為教材來說是非常完整的，特別是對於理工科系及工程師來說，應該算容易閱讀。

碁峰 www.gotop.com.tw

這種概念對許多今後要進行數位化的產業來說不可或缺，也會從根本改變我們在行銷、產品開發和設計的思維。即使不是理工背景，也請從內容感受基於資料的最佳化世界的「氛圍」。

東京大學研究所工學系研究科
人工智慧工程研究中心 / 技術經營戰略學專攻 教授
松尾豐

碁峰 www.gotop.com.tw

前言

您是每天寫程式、為了最好的使用者體驗而努力的網站工程師嗎？或者是行銷人員，為了網站服務的成長而需要考慮並實踐各種方案？不論是哪種，只要參與網站服務的開發，就無法忽視「為使用者提供更好的體驗」這個課題。隨著網站服務為了滿足市場需求而不斷增長，越來越多的使用者更深入地使用該服務，沒有什麼比這更令人高興了。

但使用者與開發網站服務的您是不同的人，您對網站的善意改變，可能會對部分使用者產生致命的負面影響；而網站中的某些 bug，卻可能受到某些使用者歡迎。使用者使用該服務的原因各式各樣，有些可能想也想不到，畢竟您對該服務了解太深了，以致於無法如一般使用者操作。

例如您是否聽過 2010 年左右開始的 Burbn 服務呢？ Burbn 是基於位置的社交網路服務，允許使用者在附近的餐館和購物中心等地點簽到，並分享在該地點拍攝的照片，兩位開發者很熱心地增加各種功能，讓更多人可以享受服務，但結果卻不盡人意。當開發者仔細觀察使用者行為時，發現大多數使用者都沒有使用定位功能！使用者熱衷的反而是附加的照片共享功能，尤其是濾鏡。當時智慧型手機的相機畫質不像現在這麼好，需要可將照片變漂亮的功能，因此這兩位開發者決定放棄濾鏡和照片分享以外的所有功能，重新開始新的服務，而這就是今日的 Instagram。即使是現在全世界都在用的 Instagram，其開發者一開始也無法全盤洞見使用者的需求。

碁峰 www.gotop.com.tw

開發網站服務時，如果不能依靠自己的感覺，那應該依靠什麼呢？最好的方法之一，就是直接詢問使用者，畢竟知道自己需求的只有使用者自己。即使在 Instagram 的例子中，如果兩位開發者沒有觀察到使用者的行為[1]，也不會如此成功吧。

但這並不意味使用者所說的都是正確的，因為並非每個人都能精準表達需求，還會有「我以為有這個功能會很方便，但實際用了卻覺得很難用」的情形。作為開發者，即使將使用者的需求變成實際的功能，也無法保證使用者就真的會用。

在沒有明確的要求、也沒有明確方式去滿足的情況下，唯一的根據就是使用者實際的**行為**，因為這就是使用者實際體驗所呈現的結果。只有將使用者行為的變化以資料方式呈現，才能量化評估一個網站服務的增長，這就是透過與資料互動，找出開發者與使用者都未曾見過的理想網站的過程。開發者觀察使用者行為，根據所獲得的見解提出假說，並將新的措施導入網站；當使用者對該措施有所反應時，使用者行為就會發生變化。開發者觀察行為的變化來檢驗假說，從而得出新的見解和假說。如此重複**假說檢驗**的循環，雖然問題和解法都不明確，但的確是個能讓網站穩健成長的方法[2]。

假說檢驗的最簡單形式就是 **A/B 測試**。在 A/B 測試中，來訪網站的使用者被隨機分為 A、B 兩組，其中一組所用的服務版本沒有導入新措施，另一組則有；觀察這兩組的使用者行動有何差異，從而評估該措施的效果。這個方法除了用於使用者介面的外觀修改，也可用在包括後端在內的所有措施，例如顯示內容演算法的改變。

想法聽起來簡單，但實際上其中有許多深奧的主題。到底要根據多少使用者數量才能判斷、到底要測什麼、這個指標真的代表網站是否健全成長嗎、目前實驗的措施是否有考慮周全……這麼多的問題，其實都沒有明確的答案。更準確地說，如果不作各種嚴格的假設，就無法正確處理。

1 像 Burbn 這樣的大型服務，改變服務的方向稱為 **pivot**（**關鍵轉折**）。因為 pivot 而成功的網站服務故事，請參見 [Holiday14] 和 [Ries11]。

2 基於假說檢驗的網站服務成長方法，通常稱為 **lean startup**（**精實創業**）。有關 lean startup 的更多資訊，請參見 [Ries11]。

碁峰 www.gotop.com.tw

在這種情況下，統計學和機器學習的知識，就是從所獲得的使用者行為資料中，獲得一點真確結果的有力武器。利用統計學和機器學習的知識，可以從因為各種因素（包括人類使用者的個性！）而導致的差異性資料中，得出一定的結論。而對分析背後所做的各種假設進行量化處理，也有助於理解分析的局限性。

本書將**網站最佳化（website optimization）**定義為「透過開發者與使用者之間的假說檢驗循環，以最大化或最小化網站服務的某些指標」，並解釋正確進行網站最佳化的數學方式。原本「網站最佳化」一詞有各種含意，譬如最小化網站顯示所需時間（效能最佳化），或是最小化在搜尋引擎顯示的排序（搜尋引擎最佳化）等，但本書並不涵蓋這些主題。

我原先是一位網站工程師，對基本的統計知識也不太有把握，但為了給使用者更好的體驗並根據資料進行決策，不知不覺就接觸到了這門學問。作為網站工程師的我從 A/B 測試開始，學習各種數學的假說檢驗方式，而這本書作為機器學習與統計學基礎的入門書，所遵循的就是當時我所走過的路。希望本書能夠透過網站最佳化，為開發網站服務或手機 app 的相關人員提供機器學習世界的新視角。

目標讀者

本書的目標讀者如下。

想要入門統計學或機器學習的網站工程師

平常專注於網站的前端開發或伺服器端開發，在工作中沒有太多機會直接接觸機器學習的工程師。透過實際執行本書的範例程式碼以加深理解，獲得使用機器學習系統來改善網站使用者體驗的知識與感覺。

參與網路行銷的網站負責人、網路行銷人員

平常就在思考並實行各種措施以增加網站使用人數的網站負責人或行銷人員。透過本書所介紹的各種網站最佳化方案，幫助您了解適合的演算法及思路，解決所面臨的問題。希望能幫助您制定更好的措施，並提高團隊內部的溝通效率。

碁峰 www.gotop.com.tw

對機器學習的應用感興趣的學生，特別是人機互動領域

影像辨識和自然語言處理等主題是機器學習的典型應用，但網站最佳化這類處理與人互動的系統應用比較少被提及。透過本書介紹的最佳化演算法及其開發流程，希望能以不同於其他書籍的角度來看待機器學習技術。

本書範例程式碼使用的語言是 Python 3。如果您是 Python 新手，建議先閱讀『Python Tutorial』[Rossum11] 等書籍和網站簡單學習一下。本書假定您已經掌握了高中學過的線性代數（向量、矩陣）和微積分等數學知識，我們將適當補充必要的數學，請放心閱讀。在「附錄 A 矩陣運算的基礎」也整理了閱讀本書所需的線性代數知識，請視情況閱讀參考。

本書架構

本書的架構如圖 1 所示。

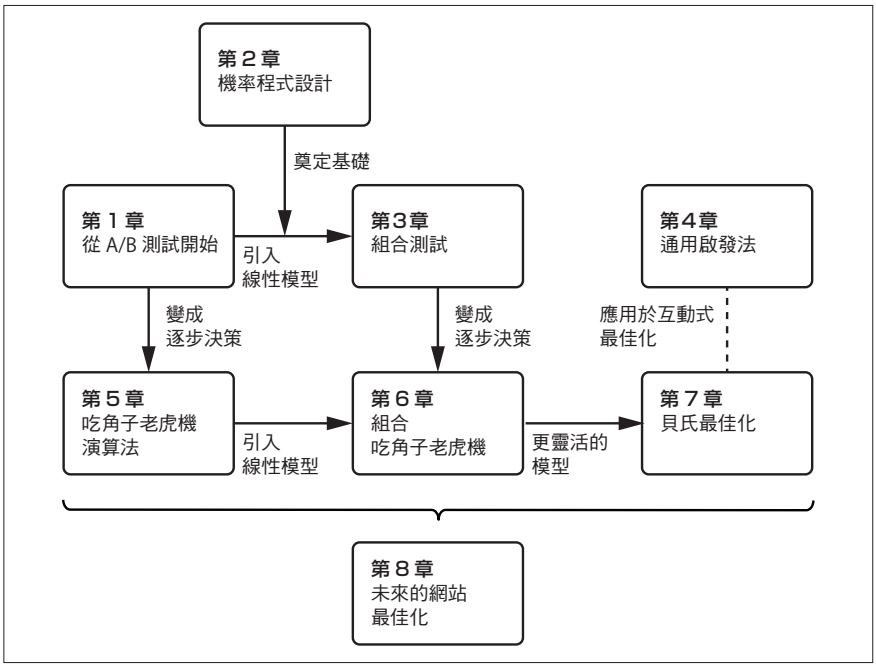

圖 1　本書架構

碁峰 www.gotop.com.tw

「第 1 章　從 A/B 測試開始：使用貝氏統計入門假說檢定」介紹假說檢驗的最簡單形式：A/B 測試，並解釋如何以數學方法（特別是貝氏推論）從獲得的資料中做出決策。

「第 2 章　機率程式設計：尋求電腦的幫助」討論馬可夫鏈蒙地卡羅法（MCMC）並將其作為貝氏推論的方法，側重於程式而非展開數學式。這裡會學到如何進行靈活的統計建模，特別是使用 MCMC 函式庫 PyMC3[3]。

「第 3 章　組合測試：分解為元素思考」對簡單的 A/B 測試進行擴充，處理當措施具有某種組合結構的情形。這裡會引入線性模型，介紹一種方法，在較少資料量時仍能有效檢驗假說。

「第 4 章　通用啟發法：不用統計模型的最佳化方法」討論通用啟發法，也是一種尋找最佳措施的方法，異於至今介紹的統計方法。我們會發現即使該方法沒有假設任何模型，也能做到高效率搜尋，同時也能應用在互動式最佳化，也就是人和電腦互動找出最佳解。

「第 5 章　吃角子老虎機演算法：面對測試中的損失」又回到了開頭的 A/B 測試問題，並且改變思維，在實驗過程中逐步做出決策，而不是得到所有試驗資料後才做出決策。接著將介紹這種問題的固定形式「多臂吃角子老虎機問題」以及其解法。

「第 6 章　組合吃角子老虎機：吃角子老虎機演算法遇到統計模型」考慮的是引入線性模型的吃角子老虎機問題。這是將**第 3 章**的內容擴展，成為逐步做出決策的吃角子老虎機演算法。這裡將說明如何加速該方法，以及擴展至個人化。

「第 7 章　貝氏最佳化：處理連續值的解空間」說明如何將線性模型擴展到更靈活的模型「高斯過程」，以將吃角子老虎機演算法應用於更複雜的問題。正如**第 4 章**所解釋的，我們希望貝氏最佳化能成為一種解題方式，解決那些難以對問題作出特定假設的問題，特別是處理人類感性的交互式最佳化問題。

3　PyMC3 文件 https://docs.pymc.io/

碁峰 www.gotop.com.tw

最後的「第 8 章　未來的網站最佳化」涵蓋了未來網站最佳化的相關問題。本書所介紹的方法中有各種假設，以便將人類使用者和網站之間的互動視為數學問題來處理。我們將逐一分解這些假設，了解網站最佳化的未解決問題。

Python 的設定

本書所有的範例程式碼都以 Python 3 寫成，並假設在 Google Colaboratory（Colab）上執行。Colab 是可以在網頁瀏覽器上使用的 Python 執行環境，不但免費還有各種功能，譬如能以 Markdown 格式撰寫註解、以 Google Drive 分享、透過 GPU/TPU 加速執行等。Colab 是互動式執行環境，可以執行每一段範例程式碼並檢查結果以幫助學習。

要使用 Colab，請先造訪 https://colab.research.google.com/，會看到如圖 2 的畫面，點擊「新增筆記本」。

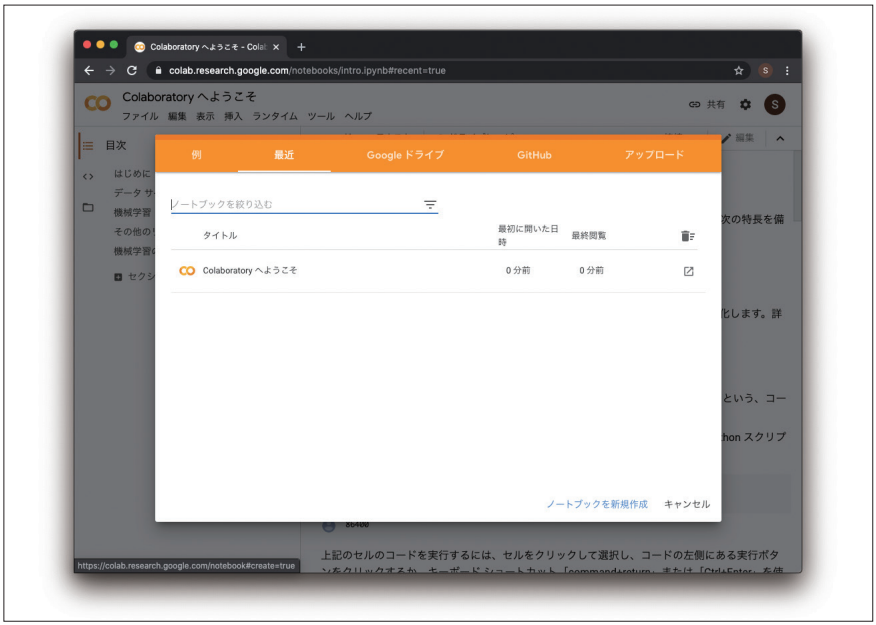

圖 2　Google Colaboratory 的首頁

碁峰 www.gotop.com.tw

第 1 章

從 A/B 測試開始：使用貝氏統計入門假說檢定

1.1 A/B 測試的影響

像 A/B 測試這種簡單的實驗，真的可以對使用者行為產生很大的影響嗎？網站設計的細微變化真的會影響我們的行為嗎？為了進一步理解，先來介紹一個 A/B 測試的應用實例。

2008 年 11 月的美國總統選舉中，民主黨候選人巴拉克 · 歐巴馬（Barack Obama）大獲全勝，擊敗了共和黨候選人約翰·麥肯（John McCain）。Internet 在這場選舉中發揮了重要作用，各方陣營在進行宣傳時都熱衷於使用 Twitter 和 Facebook 等社交媒體。

歐巴馬陣營也使用網路進行選舉活動，官方網站有一個功能是讓支持者登記接收最新消息。登記了電子郵件地址就能接收志工招募和募款等訊息。對歐巴馬陣營來說，重點是要提高支持者的註冊率，確保能與更多人溝通。

因此歐巴馬陣營開始了一項提高支持者註冊率的實驗。想法是將歐巴馬在首頁上發布的圖片和註冊按鈕的文字做各種組合，從中找出註冊率最高的。他們準備了 6 種圖片（也包括影片）和 4 種註冊按鈕，如圖 1-1（本書開頭的圖 1）所示，總共有 $6 \times 4 = 24$ 種變化。然後隨機將網站訪客分配到不同的組合，計算各自的註冊率。那麼到底哪個組合的註冊率最高呢？

碁峰 www.gotop.com.tw

圖 1-1 歐巴馬在 2008 年美國總統大選期間進行實驗的圖片和按鈕變化實例。圖片引用自 [Dan10]

答案是歐巴馬與家人的合照以及註冊按鈕標有「LEARN MORE」的組合。據報導，採用這種組合成功讓註冊率從 8.26％提高到 11.6％。假設支持者擔任志工的比例以及每位支持者的捐款額在實驗前後是固定的，那麼相當於增加了約 290,000 名志工及 6000 萬美元的捐款。這個例子充分說明了網站設計的細微變更能產生的巨大影響。

順帶一提，工作人員最希望放的似乎是歐巴馬演講的影片，但實驗結果顯示此做法反而導致註冊率變差。由此可以看出，無論網站製作者多專業，都很難確定什麼措施才真正能吸引使用者。這也代表，與使用者實際合作檢驗假說的做法是很重要的。若想知道更多關於此實驗的資訊，請參見 Dan Siroker 的部落格文章 [Dan10]，他負責此項宣傳的分析。

碁峰 www.gotop.com.tw

1.2 Alice 和 Bob 的報告

如前所述，A/B 測試是一項強大的技術，雖然簡單且容易實作，但也能對業務產生巨大影響。不過，要正確地實行並非那麼容易。我們以下面的虛構問題為例，了解如何使用 A/B 測試，用數學做出決策。

> 在某個網站的營運公司 X 工作的網路行銷人員 Alice 和 Bob 帶來了以下的 A/B 測試報告。為了提高兩人各自負責的產品介紹頁面上「索取資訊」按鈕的點擊率，他們各自準備了 A 和 B 兩個提案並顯示給使用者，下面是實驗數據。不過 Alice 和 Bob 負責不同的頁面，各自的 A 案和 B 案也都不同。

lice 的報告	A 案	B 案
顯示次數	40	50
點擊次數	2	4
點擊率	5%	8%

ob 的報告	A 案	B 案
顯示次數	1280	1600
點擊次數	64	128
點擊率	5%	8%

> 兩份報告都是 B 案的點擊率比 A 案高，對於他們各自的報告，是否能基於其結果，得出結論說 B 案的點擊率均高於 A 案（因此應採用 B 案）？

若只看點擊率，Alice 的報告和 Bob 的報告具有相同的值（都是 5％和 8％），但它們的顯示次數和點擊次數是不同的。可以看到，Bob 的報告有更多的使用者反應筆數。像這樣透過實驗觀測到的資料稱為**樣本（sample）**，樣本中包含的資料筆數稱為**樣本大小（sample size）**，這裡的顯示次數就相當於樣本大小。

Bob 的報告中，因為顯示次數比較多，看起來似乎更可靠，但僅憑這點還不能確認 Bob 的報告結果是否可靠，也許 Alice 的報告顯示次數也足以提供可靠的結果，若果真如此，那麼 Alice 這邊就能用較少使用者來檢驗假說，反而會是更優秀的。我們該如何確定這些報告足以讓我們決定採用 B 案（或不採用）？這時就該統計學大展身手了。

碁峰 www.gotop.com.tw

1.2.1 整理資料產生的過程

首先要了解報告背後發生了什麼。Alice 和 Bob 對他們負責的網頁各自有 2 個設計方案，而他們想主張不同方案會導致不同點擊率。而不同方案有不同點擊率，就代表每個方案會有一個固有的點擊率，也就是每個方案有各自的潛力。可能的確 B 案的點擊量比較高，也可能兩案的點擊量其實一樣，只是這份報告得到的結果正好是 B 案比較高。我們試圖從資料來推論其真實性。

這裡我們把某個設計方案的固有點擊率表示為 θ。θ 是點擊率，也就是機率，所以是 0 到 1 之間的值。亦即 $0 \leqq \theta \leqq 1$。

將這個設計方案顯示給使用者看以後，使用者會有所反應，可能會點擊目標按鈕，也可能不會。這裡將是否有點擊設為 r，如果有點擊就是 1，沒有就是 0。有點擊會讓人開心，所以取 reward（報酬）的第一個字母 r。

如果 θ 較大，那麼就有較高機率 $r = 1$，反之較小就是有較高機率會是 $r = 0$。極端一點，若 $\theta = 1$，那麼不管顯示幾次，使用者都會按，也就是 $r = 1$。相反地，$\theta = 0$ 時，不管顯示幾次都會是 $r = 0$。像 r 這樣能以隨機方式取各種值的變數，稱為**隨機變數（random variable）**。

而將某個設計案顯示給使用者，也就是進行某種嘗試並觀察隨機變數的值，就是所謂的**試驗（trial）**。有試驗就會伴隨著結果，這裡所謂的結果就是是否發生點擊。亦即觀測 r 具體的值是 0 或是 1。

這就像是去賭場玩吃角子老虎機一樣，如果拉下吃角子老虎機的拉桿，就會得到掉出硬幣（或沒有硬幣掉出）的結果。可能會有慷慨大放送的吃角子老虎機，也可能會有不怎麼掉硬幣的吝嗇吃角子老虎機。假設將吃角子老虎機有多慷慨（掉硬幣的機率）視為點擊率 θ，將硬幣視為點擊 r，就能看出這個問題的構造與 A/B 測試是一樣的。

根據以上討論，從某個方案中產生點擊的過程如圖 1-2 所示。

碁峰 www.gotop.com.tw

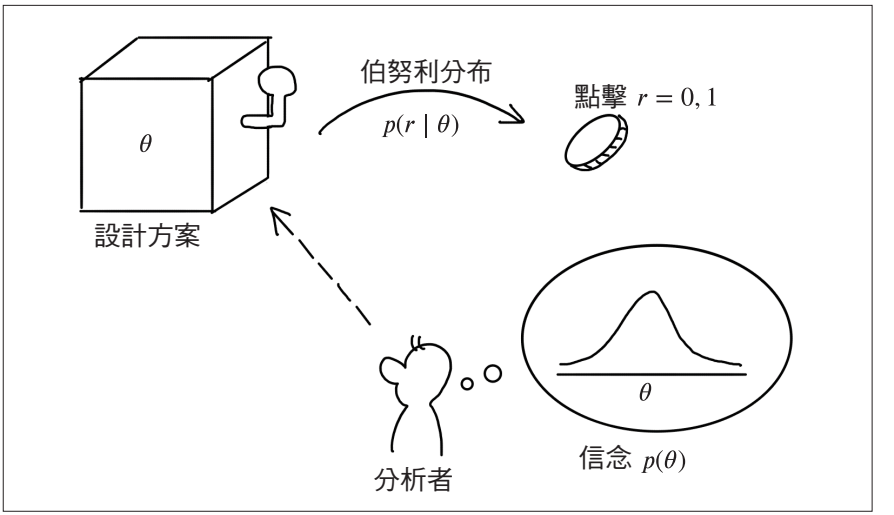

圖 1-2 點擊資料的產生過程

在這裡用吃角子老虎機作比喻，把設計方案以吃角子老虎機的形式表示。該設計方案有個固有點擊率 θ，每次顯示給使用者時，就會產生 r，表示是否有點擊。這裡以硬幣的形式來表示點擊，與前面的比喻相對應。像這樣將結果用 0 或 1 這兩種數值表示，且無論重複多少次、成功機率 θ 都不會改變的試驗，稱為**伯努利試驗（Bernoulli trial）**。而在下面看著這一切的，就是身為**分析者**的我們。

請注意這裡其實有個很重大的思維轉變。實際上是由使用者做出點擊的，但是我們將其視為**是設計方案產生了點擊**。某些使用者可能會到處點擊，也有些可能很謹慎很少點擊。回家慢慢瀏覽的使用者與在工作中急於收集資訊的使用者，行為傾向會有很大的不同。這些資訊有出現在圖中嗎？

一言以蔽之，我們這裡先暫時**無視**該資訊。我們想從這些資料中知道的是設計方案各自的固有點擊率。當然每位使用者都有不同的行為傾向，進行點擊的明明是使用者卻沒出現在圖中，的確有些奇怪。但這裡我們先只關心感興趣的部分以簡化討論。

碁峰 www.gotop.com.tw

的確每個使用者都有不同的行為傾向，這是各種複雜因素的組合，譬如時段、地點等，但我們將使用者隨機分配到不同的方案，是想將這些因素當作與想知道的效果無關。這在設計實驗時是非常重要的想法。另外關於透過比較以推論某些效果的其他原則，可參見本章末尾的專欄。

我們忽略了使用者特徵，但難道進行 A/B 測試時就無法考慮使用者特徵嗎？根據使用者特徵並提出最佳措施的架構，就跟所謂的**推薦**（recommendation）和**個性化**（personalization）有關。**第 6 章**末尾的專欄中，將討論考慮到這些使用者特徵的實驗。

觀察圖 1-2 下方，分析人員正在看著設計方案。設計方案的固有點擊率 θ 隱藏其中，分析者無法直接得知。分析者所知道的只有進行試驗的次數以及所得到的點擊次數，分析者需要根據實際觀察到的資料推測 θ。

如果將分析者對於固有點擊率 θ 的想法量化表示，會像「$\theta = 0$ 的機率約為 0.1，而 $\theta = 0.5$ 的機率約為 0.3 吧」這樣，這就是對話框內的**信念**（belief）$p(\theta)$。這種信念以**機率分布**（probability distribution）的形式表達。

1.3　機率分布

機率分布決定了一個隨機變數取某個特定值的機率。目前已經有好幾個機率開頭的詞彙，可能已經把大家搞糊塗了。這裡我們先梳理幾個例子。

1.3.1　離散值的機率分布

首先我們考慮一顆骰子，除非骰子有作弊，否則擲出每一面的機率應該都相同。假設擲骰子時出現的面 X 是隨機變數，那麼我們可以繪製隨機變數 X 的機率分布，如圖 1-3。

碁峰 www.gotop.com.tw

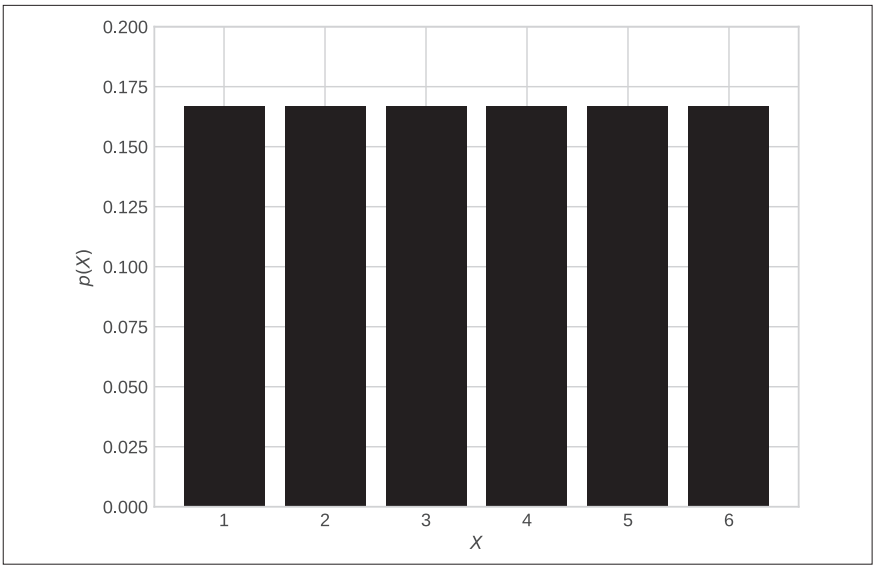

圖 1-3　理想骰子出現各面的機率

對於 1 到 6 每個面來說，出現機率都一樣是 $1/6$。所有值出現機率都相等的這種機率分布，稱為**均勻分布（uniform distribution）**。

再舉一個例子，讓我們思考在某商店街的幸運抽獎中，贏得的獎金 X（銘謝惠顧、三等獎、二等獎、一等獎）的機率分布。當然，一等獎比較難抽，所以我們自然會認為這種情況不是均勻分布。例如假設它看起來如**圖 1-4**。

像骰子和抽獎的例子這種定義為多個選項的機率分布，一般稱為**類別分布（categorical distribution）**。抽獎獎項的類別分布是以各個獎項出現機率來表示的：$p_{\text{銘謝惠顧}} = 0.6, p_{\text{三等獎}} = 0.3, p_{\text{二等獎}} = 0.09, p_{\text{一等獎}} = 0.01$。而代表某個機率分布的變數，稱為**參數（母數，parameter）**。

碁峰 www.gotop.com.tw

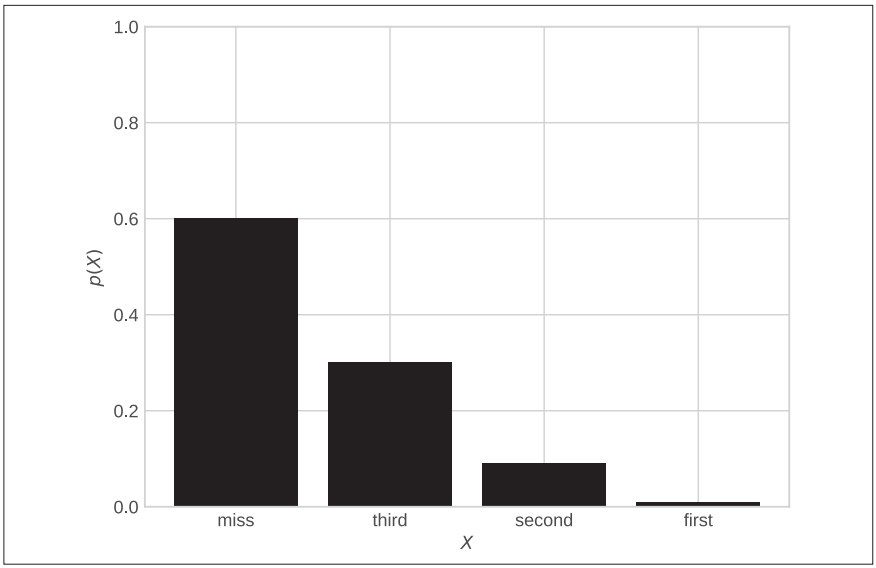

圖 1-4 抽獎獎項的機率分布

在進行關於機率分布的溝通時，參數是很有用的工具。例如在抽獎的例子中，即使不說明抽獎細節，只要說是「參數為 $\boldsymbol{p} = (0.6, 0.3, 0.09, 0.01)$ 的類別分布」，就能重現相同的機率分布。但請記得，構成機率分布的參數取決於機率分布的類型。

參數 $\boldsymbol{p}$ 以粗體顯示，表示它是一個**向量（vector）**。類別分布的參數以向量表示的原因將在 2.2 節詳述。

機率分布的形狀沒有規則，但必須滿足一些條件。首先考慮隨機變數由有限數量的不連續值（即離散值，例如 1, 2, 3, ...）表示的情況。

離散值的機率分布稱為**離散機率分布（discrete probability distribution）**。必須滿足的條件之一，是隨機變數取任何值的機率之和為 1。換句話說，如果隨機變數 X 的可能值是 $X_1, \cdots, X_n$ 的這 n 種類型，那麼就必須滿足

$$p(X = X_1) + \cdots + p(X = X_n) = \sum_{k=1}^{n} p(X = X_k) = 1$$

碁峰 www.gotop.com.tw

的條件。此處的 $p(X = X_i)$ 表示隨機變數 X 取某個值 X_i 的機率。

還有一個條件是，任何事件發生的機率都不會是負值。亦即，必須滿足

$$p(X = X_i) \geqq 0 \quad (i = 1, \cdots, n)$$

這個條件。反過來說，任何滿足這些規則的函數都可以用在離散機率分布。

1.3.2 連續值的機率分布

至此已討論了隨機變數是離散值時的機率分布，接著也來考慮隨機變數為連續值時的機率分布，即**連續機率分布（continuous probability distribution）**，如圖 1-5 的均勻分布的連續值版本（連續均勻分布）。連續均勻分布以兩個參數表示，即隨機變數 x 的最小值 a 和最大值 b。此例為 $a = 0,\ b = 1$（實線）和 $a = 0,\ b = 0.5$（虛線）時的連續均勻分布。

與離散值一樣，連續值的機率分布必須滿足以下條件：隨機變數取任何值的機率之和為 1。但請注意，在連續機率分布中，「機率之和」是用積分 $\int$ 而不是總和 $\sum$。

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

由於積分是找到特定區域面積的過程，所以換句話說，條件就是：機率分布的函數與水平軸之間所圍面積為 1。例如在圖 1-5 中，$a = 0,\ b = 1$ 的連續均勻分布（實線）下的矩形寬度為 1、高度為 1，所以面積為 1。同樣地，$a = 0,\ b = 0.5$ 的連續均勻分布（虛線）的矩形寬度為 0.5、高度為 2，所以面積為 1。

碁峰 www.gotop.com.tw

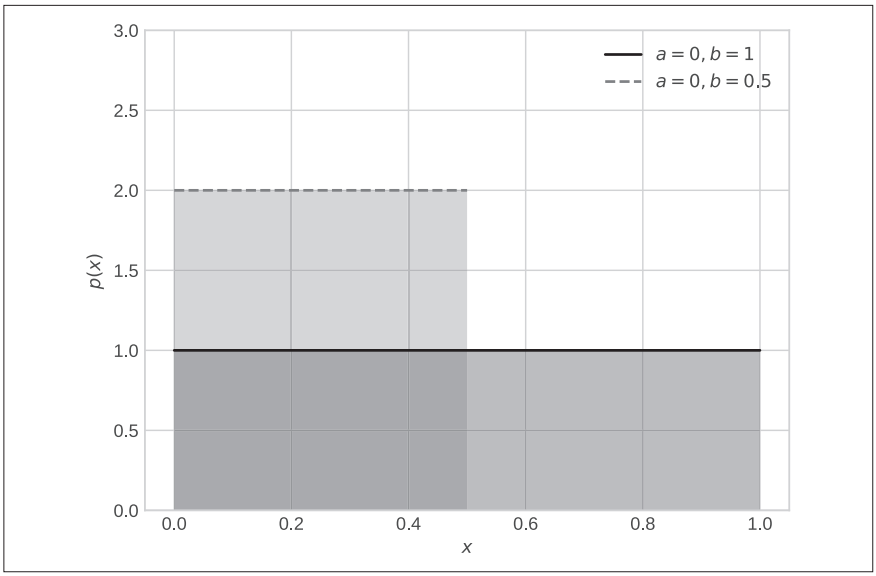

圖 1-5 連續均勻分布的範例

機率分布的另一個條件也一樣，即連續機率分布也必須符合對於任意隨機變數的值 x，滿足

$$p(x) \geqq 0$$

的條件。

在了解連續機率分布時，須特別注意無法從給定的機率分布直接得到隨機變數取某值的機率。在剛剛的離散機率分布的例子中，機率分布對應到每種事件發生的機率，譬如骰子出現 2 的機率，可以從圖 1-3 的 $X = 2$ 對應的長條高度看出，所以知道是 $1/6$。但在連續機率分布不能這樣解讀，而是要用來找出隨機變數落在一定範圍內的機率。例如在 $a = 0,\ b = 0.5$ 的均勻分布中，隨機變數 x 落在 0.2 和 0.3 之間的機率為寬度 0.1、高度 2 的矩形面積，即 0.2。

碁峰 www.gotop.com.tw

在離散機率分布中，對應垂直軸的值解讀為機率沒有問題，但在連續機率分布中，垂直軸對應的值是機率**密度**，只有考慮所關心的隨機變數所取的值範圍，才能得到機率。在 $a = 0,\ b = 0.5$ 的均勻分布中，函數值可以超過 1，正是因為它表示的是密度而非機率。即使使用相同的機率分布，也會根據隨機變數是離散值還是連續值而有所不同，因此我們將代表離散機率分布的函數稱為**機率質量函數（probability mass function, PMF）**，而表示連續機率分布的函數稱為**機率密度函數（probability density function, PDF）**，作為區別。

1.3.3 以離散化得到機率密度函數的近似

在電腦上處理連續密度函數時，將其離散化（discretize）進行近似處理的方式可能會更容易。換句話說就是將連續值的隨機變數以一定間隔離散化，將其作為機率質量函數處理。

例如考慮這個函數，

$$f(x) = \begin{cases} x + 1 & (-1 \leqq x \leqq 0) \\ -x + 1 & (0 \leqq x \leqq 1) \end{cases}$$

該函數定義在 $-1 \leqq x \leqq 1$ 的範圍內。這個函數與橫軸之間形成的區域是底為 2、高為 1 的三角形，可知面積為 1（即滿足機率密度函數的條件）。

首先，將此範圍劃分為多個區域，然後計算每塊區域的中心對應的機率密度函數的值。如果畫一個矩形，以劃分的區域為寬，機率密度函數的值為高，就能以大量的細長矩形來近似原始函數 $f(x)$，如圖 1-6。

碁峰 www.gotop.com.tw

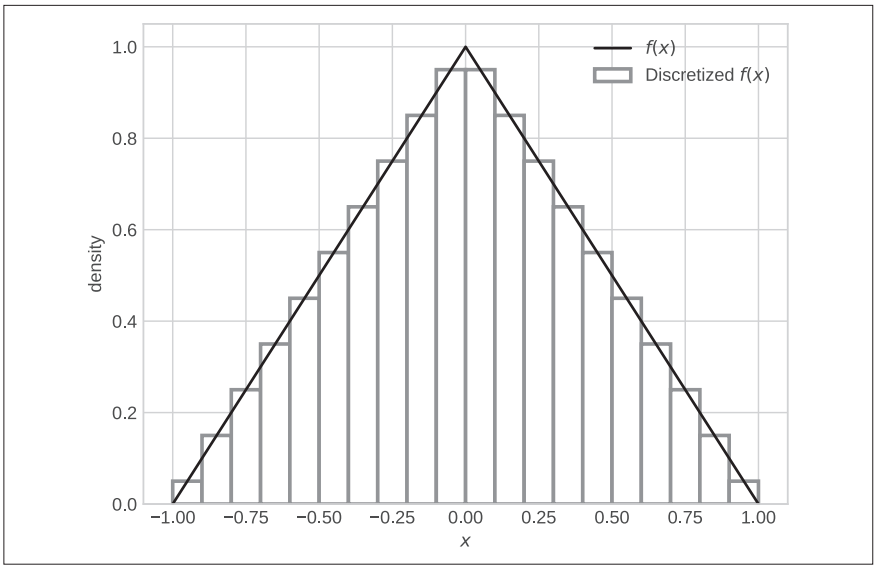

圖 1-6　機率密度函數 $f(x)$ 和以離散化近似的函數

該圖中將 $-1 \leqq x \leqq 1$ 的範圍分為 20 個寬度 0.1 的區域，可得知每個區域分別是

- 區域 1：$-1 \leqq x \leqq -0.9$（中心：$x = -0.95$）
- 區域 2：$-0.9 \leqq x \leqq -0.8$（中心：$x = -0.85$）
- ……
- 區域 20：$0.9 \leqq x \leqq 1$（中心：$x = 0.95$）

以每個區域的中心計算 $f(x)$ 的值，並將其當作對應區域的矩形高度。分割的寬度越窄，近似的結果就越好。但請注意這畢竟只是近似，因此矩形面積之和不見得會是 1，這表示可能會無法滿足機率密度函數的條件。

接著為了將此函數轉換為機率質量函數，將各區域中心設為新的離散隨機變數，並將各矩形面積作為與之對應的機率質量。可以想像成是將各區域的機率質量**濃縮**到各區域中心去。

碁峰 www.gotop.com.tw

此外如前所述，這些細長矩形的面積總和不一定是 1，所以我們將各機率質量都除以「各隨機變數值所對應機率質量之總和」，將機率質量總和調整為 1（為了滿足機率質量函數的條件）。像這樣將函數之和拿來進行除法以滿足機率分布條件的做法，稱為**正規化（normalization）**，這個方法在之後的章節會不時看到。

經過上面的步驟，以機率質量函數近似機率密度函數 $f(x)$ 的結果如**圖 1-7**。從圖中可以看出結果是一個機率質量函數，該函數保留了與原始函數 $f(x)$ 近似的形狀。看垂直軸可以發現有的值比**圖 1-6** 中的小，這是因為**圖 1-6** 處理的是機率**密度**，而**圖 1-7** 處理的是機率**質量**。

原本的機率密度函數 $f(x)$ 定義了 $-1 \leqq x \leqq 1$ 範圍內的所有值，而近似後的機率質量函數則只定義了各區域中央所對應的 20 個值，請注意這點。例如問這個機率質量函數「 $x = 0.72$ 的機率是多少」是不會有答案的，因為 $x = 0.72$ 不在那 20 個值（-0.95, -0.85, ……, 0.95）中。乍看很不方便，但如果將分隔寬度設得夠細，實際上就能覆蓋分析時所需的值。另外，將其視為機率質量函數還有一個優點，就是能避免積分計算。

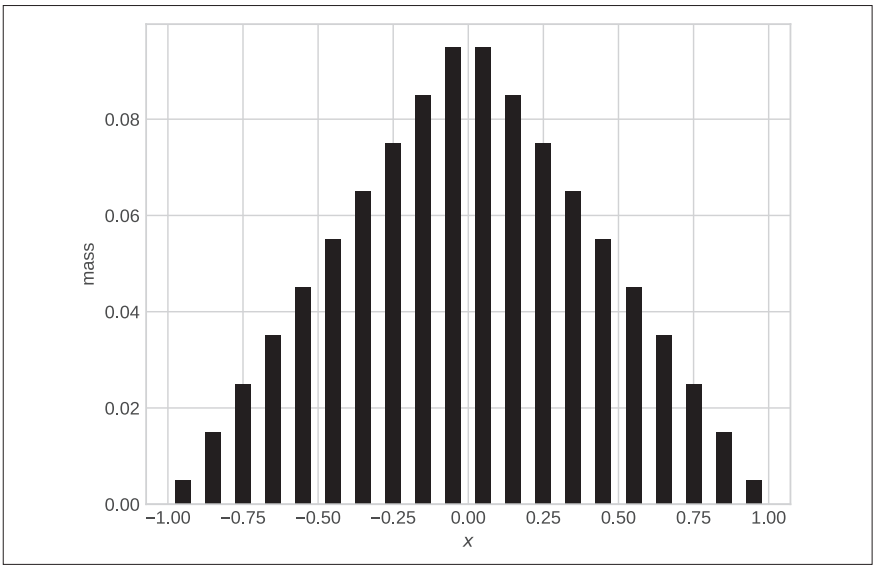

圖 1-7　以機率質量函數近似 $f(x)$ 的結果

碁峰 www.gotop.com.tw