對假設進行比較

即便可能性很低，你還是先接受了飛碟的假設，因為那時並沒有更好的解釋。然而，現在出現了另一個可能解釋：電影拍攝。於是你形成了一個**對立假設**。考慮對立假設的過程，就是根據你手上有的資料，對不同理論做出比較。

當你看到纜線、攝影團隊，還有額外的燈光時，你的資料就改變了，最新資料變成：

$$D_{\text{更新版本}} = \text{亮光、碟狀物、纜線、拍攝團隊、額外燈光等等}$$

在觀察到其他資料後，你改變了原本對這個情況做出的結論。我們將這個過程拆解成貝氏推理的流程：首先，你的第一個假設給出了一個解釋資料的方法，並解決你的困惑；但加上了這些新的觀測結果後，這個假設對原始資料的解釋就不夠充足了。我們能將之寫為：

$$P\left(D_{\text{更新版本}} \mid H_1, X\right) = \text{非常非常低}$$

你現在有了一個新的假設：H_2。這個假設能對資料做出更合理的解釋，寫為：

$$P\left(D_{\text{更新版本}} \mid H_2, X\right) >> P\left(D_{\text{更新版本}} \mid H_1, X\right)$$

這裡的關鍵是要理解，我們在將不同的假設相比，以了解它們對觀測資料做出的解釋孰好孰壞。當我們說：「資料發生的機率在第二個假設中比第一個假設還要高。」就是在說第二個假設較能完整解釋我們的觀察結果。這也帶我們走到了貝氏分析的核心：**檢驗你的信念對這個世界做出的解釋有多充足**。當我們說某個信念比另一個信念更準確時，就是因為它對我們觀測到的這個世界做出了更好的解釋。

在數學上，我們用這兩個機率的比例來表達這個概念：

$$\frac{P\left(D_{\text{更新版本}} \mid H_2, X\right)}{P\left(D_{\text{更新版本}} \mid H_1, X\right)}$$

如果這個比例的數值很大，好比說 1000，就表示「H_2 對資料的解釋比 H_1 合理 1000 倍。」由於 H_2 對資料的解釋比 H_1 好上很多，所以我們的信念從 H_1 更新為 H_2。這正是當你對觀測資料有不同的可能解釋時，會讓你改變想法的實際過程。現在，你相信看到的是電影拍攝的過程，正是因為在你觀察到的所有資料中，這是最合理的解釋。

資料傳遞信念；信念不應傳遞資料

最後要特別強調的是，在所有的例子中，唯一不容置疑的部分只有資料本身。假設會改變，生活經驗 X 可能會與他人不同，唯獨資料 D 是全體共享的不爭事實。

考慮下面兩個公式，我們在本章廣泛使用了第一個公式：

$$P\left(D \mid H, X\right)$$

讀作「根據我的假設和生活經驗，這個資料會出現的機率。」更簡單的說法是：「我的信念對自己的觀察結果做出多好的解釋。」

在我們的日常生活中，很常出現一種互換的情景，也就是：

$$P\left(H \mid D, X\right)$$

讀作「根據我的資料生活經驗，**我的信念**成立的機率。」或是「我的觀察結果有多支持自己的信念。」

通常在實際情況中，當你需要為一個抽象信念指派機率時，想想你願意為這個信念下多少賭注相當有幫助。你會願意用 1 賠 1000000 來賭明天太陽會升起，但對你最愛的棒球隊獲勝的情況，你可能會將賠率降低許多。在任一案例中，你都能用我們剛剛的計算過程，算出該信念發生機率的準確數字。

為拋硬幣測量信念

我們現在有一個利用賠率來為抽象概念判斷機率的方式了，但真正的耐受度測試才要開始呢：這個方法是否依然適用於我們之前用計數來計算的拋硬幣案例中呢？與其將單次拋硬幣視為一個**事件**，我們能換個說法，將問題變成「我有多相信下一次拋硬幣的結果會是正面？」現在我們不是在討論 $P(\text{正面})$ 了，而是一個對拋硬幣結果的假設，或說信念，也就是 $P(H_{\text{正面}})$。

跟之前一樣，我們需要一個對立假設來比較我們的信念。我們可以單純地說對立假設就是沒有得到正面 $H_{\neg\,\text{正面}}$，但得到反面 $H_{\text{反面}}$這個選項比較貼近我們的日常生活，所以我們會用後者。畢竟我們最在乎的是要合情合理。不過，在這個討論中，還是必須要認知到：

$$H_{\text{反面}} = H_{\neg\,\text{正面}},\ \text{且}\ P\left(H_{\text{反面}}\right) = 1 - P\left(H_{\text{正面}}\right)$$

看看我們能如何將信念建模成這些競爭假設間的比例：

$$\frac{P\left(H_{\text{正面}}\right)}{P\left(H_{\text{反面}}\right)} = ?$$

還記得嗎？我們想要將此式讀作「我相信這個結果會是正面的信念比我認為會是反面的信念要強上幾倍？」隨著賭局開始，由於每個結果的不確定性完全一樣，所以唯一公正的賠率會是 1 賠 1。當然，我們也能挑任何賠率，只要兩個數字相同就行：2 賠 2、5 賠 5，或 10 賠 10。這些賠率全部都有一樣的比例：

這些結果呈現的是在 12 種可能結果中，會得到點數 6 的 2 種結果，與我們預期的的 *P*(6 點) = 1/6 相同。基於總共有 6 種結果符合拋出正面這個條件，並有 2 種結果符合擲出 6 點這個條件，我們可能會直接認為任一條件會發生的結果總共有 8 種。然而，正面　6 點在兩個清單中都有列出，也就是說如果我們單純地將 *P*(正面) 和 *P*(6 點) 相加，在重複計算下得到的答案其實是超量的。而事實上，在 12 種獨特結果中，只有 7 種符合我們的條件。

為了取得正確的機率，我們必須將所有機率相加後，減去兩個事件同時發生的機率，也就是用**聯集**來計算非互斥事件的機率，稱為機率的**求和定則**：

$$P(A) \text{ 或 } P(B) = P(A) + P(B) - P(A,B)$$

我們將每個事件發生的機率相加，並減去兩個事件同時發生的機率，以確保同屬於 *P*(*A*) 及 *P*(*B*) 的結果不會被重複計算。所以在我們擲骰子和拋硬幣的例子中，擲骰子的結果小於 6 點或是拋硬幣的結果為正面的機率為：

$$P(\text{正面}) \text{ 或 } P(6\text{點}) = P(\text{正面}) + P(6\text{點}) - P(\text{正面}, 6\text{點}) = \frac{1}{2} + \frac{1}{6} - \frac{1}{12} = \frac{7}{12}$$

最後，我們再用一個**聯集**的例子來鞏固這個概念。

範例：計算得到鉅額罰款的機率

想像一個新的場景：你正開著車，展開公路旅行，卻因超速而被攔停。由於已經很久沒有被要求靠邊停車了，你突然想到自己很可能忘了將新的行照和汽車保險證放在副駕駛座的儲物箱。兩樣之中要是少了任何一樣，罰單的金額就會變得更高。在你打開儲物箱之前，要如何算出缺少其中一份文件，並因此得到高額罰款的機率呢？

首先，你相當肯定自己有將行照放在車上，並對此事指派 0.7 這個機率。同時，你也相當確定自己將汽車保險證留在家裡的櫃台上，所以認為能在車內找到它的機率只有 0.2。現在的狀況是：

$$P(\text{行照}) = 0.7$$
$$P(\text{保險證}) = 0.2$$

這兩個數字都是你**能夠**在儲物箱中找到文件的機率，然而你真正擔心的是其中任何一樣文件**不在現場**。要得知文件不在車上的機率，只需簡單運用減法：

$$P(\text{缺少}_{\text{行照}}) = 1 - P(\text{行照}) = 0.3$$
$$P(\text{缺少}_{\text{保險證}}) = 1 - P(\text{保險證}) = 0.8$$

如果我們用的是加法，而不是完整的求和定則來計算這個組合機率，得到的結果會大於 1：

$$P(\text{缺少}_{\text{行照}}) + P(\text{缺少}_{\text{保險證}}) = 1.1$$

這是因為兩件事並非互斥：兩份文件完全有可能都不在車上。也就是說，上面這個方法會害我們重複計算，而我們必須要找到同時缺少兩份文件的機率，並將這個值減去。關於這點，我們可以用乘積法則來找到答案：

$$P(\text{缺少}_{\text{行照}}, \text{缺少}_{\text{保險證}}) = 0.24$$

現在我們能用求和定則，算出至少缺少一份的文件機率了。這個過程與之前計算拋出正面或擲出 6 點的例子是相同的：

$$P(\text{缺少}) = P(\text{缺少}_{\text{行照}}) + P(\text{缺少}_{\text{保險證}}) - P(\text{缺少}_{\text{行照}}, \text{缺少}_{\text{保險證}}) = 0.86$$

套用機率密度函數來解決我們的問題

當我們將黑盒子案例中的資料數值代入貝他分布，並畫成圖表（圖 5-3），會發現這看起來像是圖 5-2 的連貫版本。而這個圖闡述的是貝他分布 Beta(14,27) 的機率密度函數。

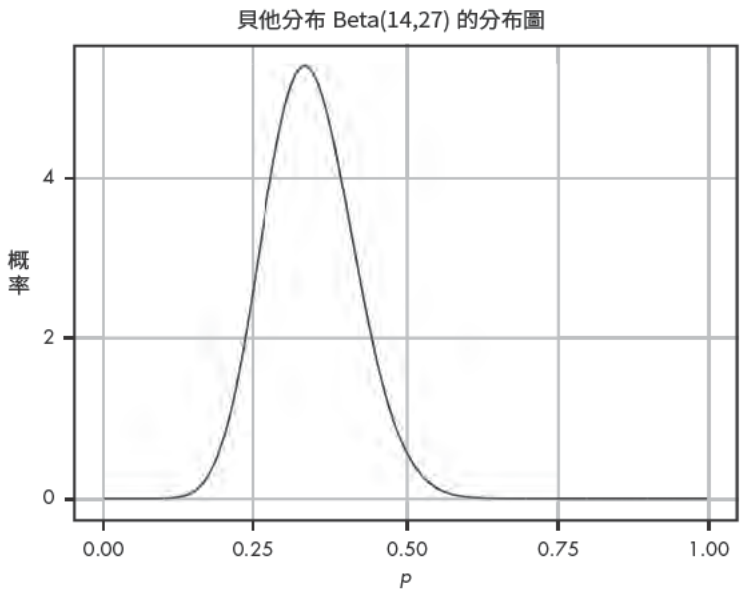

圖 5-3 根據我們收集到的黑盒子資料形成的貝他分布圖

如你所見，圖表中大部分的密度都小於 0.5。有鑒於此，我們預期放入一枚硬幣，會從黑盒子得到兩枚硬幣的機率低於五成。

這個圖表同時也顯示出，若說放入一枚硬幣，黑盒子會返還兩枚硬幣的情況至少有五成，這基本上是不太可能的。至於現在，在還沒有犧牲太多硬幣的情況下，我們已經發現了使用這個盒子，輸錢比贏錢的機率要高。雖然我們可以在這張圖表中看到自身信念的分布狀態，但我們還是希望能將「黑盒子會返還兩枚硬幣的實際機率小於 0.5 的概率」這個信念精準量化。要得出這個答案，我們只需用上一點微積分（和一些 R 語言）。

抽卡遊戲的逆向工程

在現實世界中，我們幾乎永遠不可能知道事件發生的真實機率。這就是為什麼貝他分布是能讓我們了解手邊資料的強力工具。在第四章的抽卡遊戲中，我們已經知道自己想要的每一張卡牌被抽中的機率，但在現實情況中，遊戲開發商幾乎不會把這樣的資訊告訴玩家（可能不想讓玩家算出抽到自己想要的卡牌這個機率有多低吧）。現在，假定我們有一個新的抽卡遊戲**頻率學派大戰！**。這個遊戲裡也有許多統計學家的角色，而這次，我們想要抽到的是布拉德利．埃夫隆（Bradley Efron）這個角色。

我們不知道能抽到這張卡牌的機率為何，但我們真的很想要得到它，還希望能不只抽到一張。我們失心瘋地花了一堆錢，發現在 1200 次的抽卡中，得到了 5 張埃夫隆卡牌。我們有個朋友正在考慮是不是要為這個遊戲花錢，但他只打算在抽到埃夫隆卡牌的機率大於 0.005 這件事成立的概率大於 0.7 時，才願意掏錢。

朋友請我們找出答案，讓他知道是不是應該花錢抽卡。資料顯示，抽卡 1200 次，只得到了 5 張埃夫隆卡牌，所以我們將這個貝他分布 Beta(5,1195) 化成圖表，得到圖 5-4（別忘了抽卡總數為 $\alpha + \beta$）。

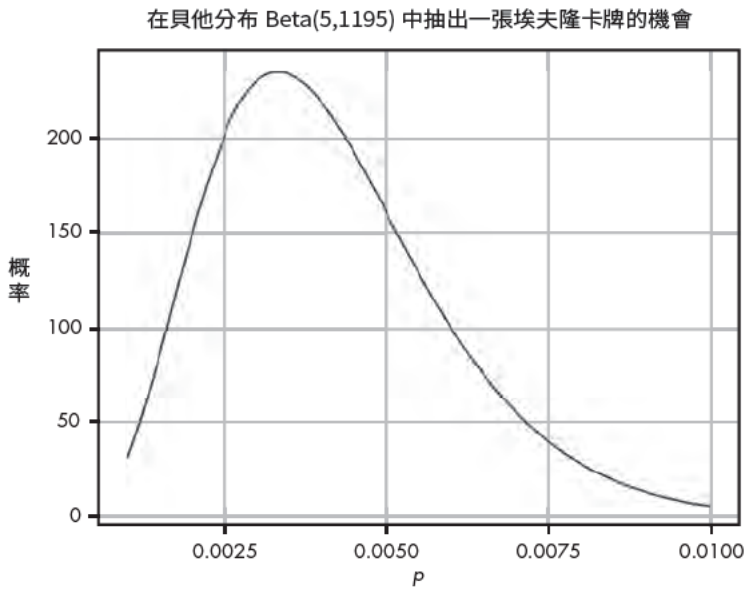

圖 5-4　根據資料，能抽到埃夫隆卡牌的貝他分布圖。

從這個圖表中，我們可以看到幾乎所有的機率密度都小於 0.01。我們需要知道的是大於 0.005 的確切概率是多少，畢竟這才是我們的朋友在意的數值。我們可以在 R 語言中將整個貝他分布積分，如前所述：

```
Integrate(function(x) dbeta(x,5,1195),0.005,1)
0.29
```

結果顯示，根據我們觀察到的證據，抽到埃夫隆卡牌的機率至少有 0.005 這件事，成立的概率只有 0.29。朋友說了，只有在機率接近或大於 0.7 時，他才考慮要花錢抽卡，所以根據我們蒐集到的資料建立出的證據，他最好不要賭這一把。

在本章中，我們要來看看如何用事前機率解題，以及各種能用機率分布，將信念描述為包含所有可能數值的值域而非單一數值的方式。機率分布會比單一數值合用有兩大原因。

首先，在現實生活中，我們要考慮的可能信念範圍通常不小。其次，描繪出機率的範圍能讓我們在一組假設集上描述信心水準。在第五章檢驗黑盒子的時候，其實就已經探究過這兩點了。

C-3PO 的小行星群疑慮

我們用《星際大戰：帝國大反擊》這部電影來舉例。片中在數據分析上有個令人印象深刻的錯誤：當韓索羅駕駛千年鷹號進入小行星群，試圖躲開敵軍時，全知的 C-3PO 提醒他，在機率上的贏面並不大：「先生，你能成功穿越小行星群的機率大約只有 1/3,720 啊！」

「跟我談什麼機率！」韓回答。

表面上，這只是一個捨棄了「無聊」數據分析的有趣橋段，然而此處其實有個巧妙的難題。身為觀眾，我們知道韓是有能力完成這個挑戰的，但我們也不會因此否定 C-3PO 的分析結果。說真的，其實韓也知道這麼做很危險，他說過：「他們要是群瘋子才會跟著我們。」加上在電影中，緊追在後的鈦戰機隊全軍覆沒，也足以證明 C-3PO 提出的數值並非無憑無據。

C-3PO 在計算時忘了一件事，那就是韓本人是個亡命之徒！ C-3PO 並沒有做錯什麼，只是忘了加上這項非常重要的變數。現在，問題變成：在不將機率完全摒棄的情況下，我們是否能像韓提議的那樣，避免 C-3PO 的失誤？要回答這個問題，我們得先將 C-3PO 的思考模式和我們對韓的信念這兩項素材建模，再用貝氏定理來將兩者融合。

下段內容，我們會先談談 C-3PO 的推論過程，接著再套入韓的剽悍行徑。

確立 C-3PO 的信念

C-3PO 可不是隨口喊個數目而已，他能流利使用超過六百萬種溝通方式，而這種能力背後需要大量的資料來支撐。所以，我們可以假定他手上其實握有足以佐證「近似 1/3720」的確切數據。C-3PO 提供的是成功穿越小行星群的**近似**機率，有鑒於此，我們知道他提供給韓的資訊僅足以建立可能成功率的一個範圍。而要描繪出這個範圍，我們要看的是成功機率的信念**分布**，而不只是用單一數值呈現這個機率。

對 C-3PO 來說，結果只有正反兩面，要不成功穿越，要不完全失敗。我們將運用 C-3PO 擁有的數據，加上第五章學到的貝他分布，來計算所有可能的成功機率。使用貝他分布是因為這種方法能根據我們對成敗的已知資訊，對此事件所有的可能機率建立出正確的範圍。

別忘了，計算貝他分布時要將資料參數化為 α（觀測到的成功總數）及 β（觀測到的失敗總數）：

$$P(\text{成功率} \mid \text{成敗}) = \text{Beta}(\alpha,\beta)$$

這個分布表述的是根據我們手上有的資料，哪一個成功率最有可能發生。

要找出 C-3PO 的信念，我們得先猜猜他的資料源自何處。若說在 C-3PO 的紀錄中，曾經有 2 個人成功穿越了小行星群，但不幸的是，有 7440 人的旅程在華麗爆炸中結束！圖 9-1 繪製出 C-3PO 對真實成功率這個信念的機率密度函數。

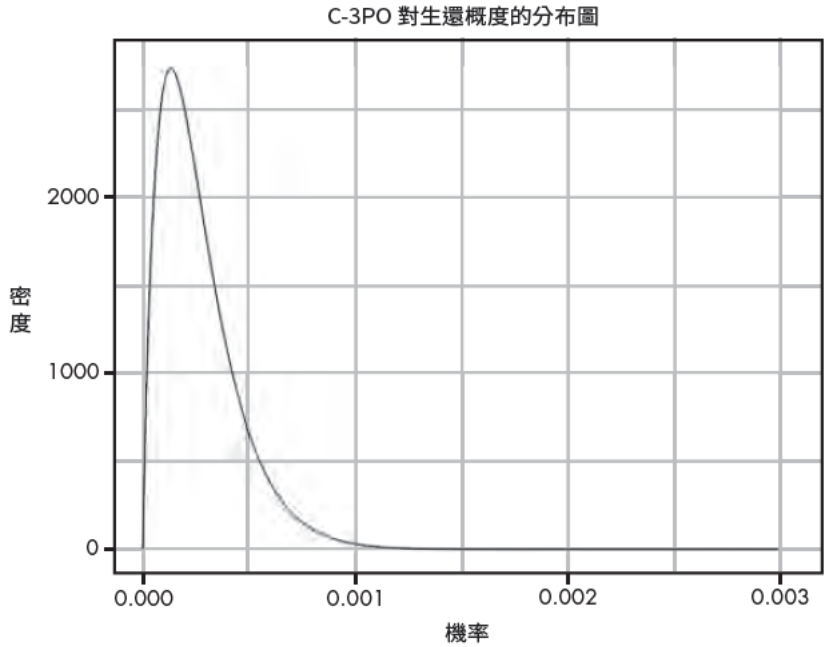

圖 9-1　貝他分布圖，表示 C-3PO 對韓能夠生還的信念。

對任何要進入小行星群的一般飛行員來說，這看起來糟透了。用貝氏定理的語言來陳述，就是 C-3PO 根據觀察結果，估計真實成功率為 3720：1，這個值也就是我們在第八章談過的概度。接下來，我們要來找出事前機率。

解釋韓的剽悍行徑

C-3PO 的分析蘊含的問題，是資料來源為所有飛行員，但他忘了韓可不是大眾認知下的一般飛行員。如果我們無法用一個數來表示韓的剽悍行徑，那麼我們的分析就行不通了。不只是因為韓成功通過了小行星群，也因為我們相信他做得到。「統計」這個工具能協助並組織我們對世界推理和信念的過程，如果我們的統計分析不僅違背自己的推理和信念，還無法動搖固有的想法，代表這個分析一定出了什麼差錯。

藉由取機率密度函數曲線下的累積面積，累積密度函數就能得到這個機率值（如果你比較習慣用微積分的話，累積密度函數就是機率密度函數的**反導函數**）。我們可以將這個過程歸納為兩個步驟：(1) 為機率密度函數的每個數值計算出曲線下的累積面積，及 (2) 繪製出這些數值。這就是我們的累積密度函數了。任意 x 軸數值對應到的曲線數值就是 x 或小於 x 的值出現的機率。在的值為 0.0065 時，曲線對應到的數值為 0.008，而這跟我們之前的計算出結果一模一樣。

要理解這是如何運作的，讓我們將這個問題的機率密度函數分割成每份 0.0005 的小塊，並聚焦在我們的函數中機率密度最高的那個區域：0.006 至 0.009 之間。

圖 13-2 呈現了貝他分布 Beta(300,39700) 其機率密度函數曲線下的累積面積。如你所見，我們的曲線下累積面積顧及了所有在左側的小塊面積。

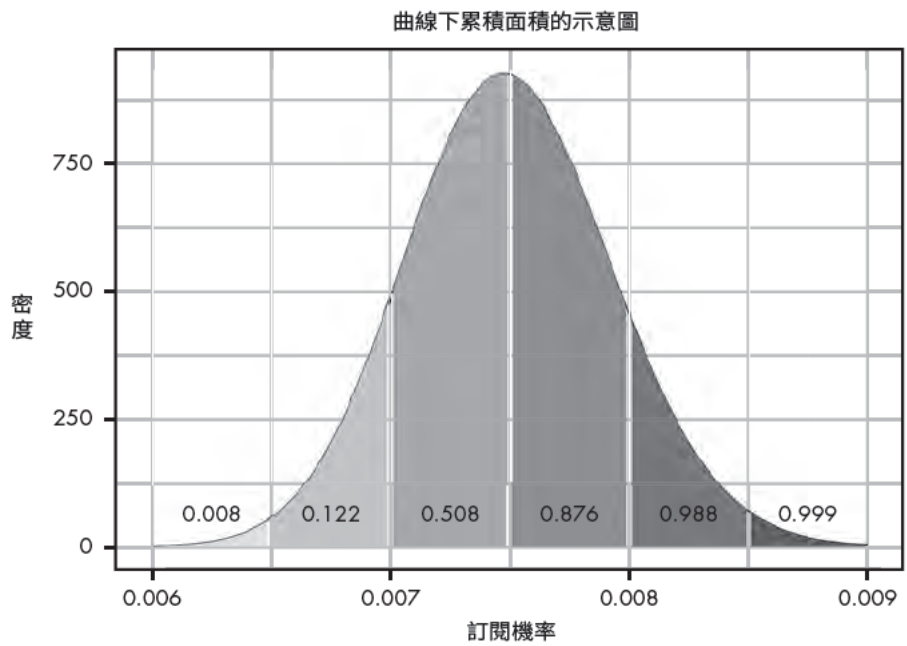

圖 13-2 曲線下累積面積的示意圖