詳細解答

STEP II 2000，第 12 題

概率與統計 • 正態變量之和

> **STEP II 2000，第 12 題**
>
> 繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：
>
> step.maths.org/questions/00-s2-q12

關鍵思路

Henry 遲到，正正就是他的總行程時間超過他在 0900 之前所餘時間的時候，所以對每條路線，我們把兩段獨立的正態行程時間相加，得出單一個正態變量(normal variable)。標準化（standardising）之後，「遲到」就變成「一個標準正態變量超過某個數」，而由於 Φ 是遞增的，兩條路線毋須查表便可比較。餘下的工作就是全概率定律（law of total probability）和貝葉斯定理（Bayes' theorem）。

解答

比較兩條路線

我們假設在每條路線上，火車行程與其後的行程互相獨立（independent）。

King's Cross。設 $X_1 \sim N(55, 25)$ 為火車行程時間，$Y_1 \sim N(30, 144)$ 為其後的行程時間。由於它們是獨立的正態變量，它們的和也是正態的，而平均值（mean）與方差（variance）各自相加：

$$T_1 = X_1 + Y_1 \sim N(85,\, 169), \qquad \text{標準差}13.$$

火車在 0715 開出，所以 Henry 有 105 分鐘，而當 $T_1 > 105$ 時他便會遲到。設 $Z \sim N(0,1)$，則

$$\mathrm{P}(\text{遲到} \mid \text{King's Cross}) = \mathrm{P}\left(Z > \frac{105-85}{13}\right) = 1 - \Phi\left(\tfrac{20}{13}\right) = A.$$

Liverpool Street。同樣地，$T_2 = X_2 + Y_2$，其中 $X_2 \sim N(65, 16)$ 與 $Y_2 \sim N(25, 9)$ 互相

獨立，所以

$$T_2 \sim N(90, 25), \qquad \text{標準差5.}$$

這班火車在 0720 開出，餘下 100 分鐘，所以

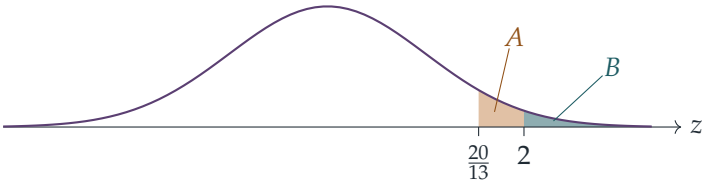
$$\mathrm{P}(\text{遲到} \mid \text{Liverpool Street}) = \mathrm{P}\left(Z > \frac{100-90}{5}\right) = 1 - \Phi(2) = B.$$

想一想 不查表，怎樣比較 A 與 B？

Φ 是一個累積分佈函數（cumulative distribution function），其密度為正，所以它是嚴格遞增的。由於 $\frac{20}{13} < \frac{26}{13} = 2$，我們有 $\Phi(\frac{20}{13}) < \Phi(2)$，因此

$$A = 1 - \Phi\left(\tfrac{20}{13}\right) > 1 - \Phi(2) = B.$$

所以 Henry 若乘搭往 King's Cross 的火車，遲到的機會較大，這正是所要證明的。（有趣的是，這條路線平均而言反而較快；它之所以落敗，是因為由 King's Cross 出發的那段地鐵行程變化太大。）



標準正態曲線下塗上陰影的尾部：A 是整個陰影部分的面積，B 是 2 以外顏色較深的部分。

遲到次數的期望值

在單一次行程中，設 K 為 Henry 乘搭往 King's Cross 的火車這一事件，故 $\mathrm{P}(K) = p$，並設 D 為他遲到這一事件。根據全概率定律，

$$\mathrm{P}(D) = \mathrm{P}(D \mid K)\,\mathrm{P}(K) + \mathrm{P}(D \mid K')\,\mathrm{P}(K') = pA + (1-p)B.$$

假設 M 次行程互相獨立，遲到次數便服從參數為 M 和 $\mathrm{P}(D)$ 的二項分佈（binomial distribution），所以

$$L = M(pA + (1-p)B) = M(B + p(A-B)).$$

由於 $A - B > 0$ 且 $0 \le p \le 1$，表達式 $B + p(A-B)$ 隨 p 遞增，由 $p = 0$ 時的 B 增至 $p = 1$ 時的 A。因此對 p 的每一個可能值，

$$BM \le L \le AM,$$

左邊的等號僅在 $p = 0$ 時成立，右邊的等號僅在 $p = 1$ 時成立。用文字來說：L/M 是 A 與 B 的加權平均（weighted average），所以它必定介乎兩者之間。

估計 p

在遲到的日子當中，他乘搭了往 King's Cross 火車的比例，是 $\mathrm{P}(K \mid D)$ 的估計值；而因為 M 很大，我們把觀察所得的 $\frac{3}{5}$ 視為等於這個概率。根據貝葉斯定理，

$$\mathrm{P}(K \mid D) = \frac{\mathrm{P}(D \mid K)\,\mathrm{P}(K)}{\mathrm{P}(D)} = \frac{pA}{pA + (1-p)B}.$$

令此式等於 $\frac{3}{5}$：

$$5pA = 3pA + 3(1-p)B$$
$$2pA + 3pB = 3B,$$

所以估計值是

$$\hat{p} = \frac{3B}{2A + 3B}.$$

驗算

- 以數值計算，$A \approx 0.0620$ 而 $B \approx 0.0228$，所以經 King's Cross 約有 6% 的時候遲到，經 Liverpool Street 則約 2%。對每條路線各模擬 4×10^5 次行程，結果吻合至小數點後三位。
- 代入這些數值，$\hat{p} \approx 0.355$。代回原式，$pA/(pA + (1-p)B) = 0.6$ 準確成立；而以這個 p 模擬整個過程，所得的遲到個案中約有 60% 來自 King's Cross。
- 對最後答案作合理性檢查：假如兩條路線同樣可靠（$A = B$），遲到的情況便只反映他選乘哪班火車，所以當中有 $\frac{3}{5}$ 經 King's Cross，就表示 $p = \frac{3}{5}$。公式給出 $3B/5B = \frac{3}{5}$。由於 King's Cross 是較不可靠的路線，它在遲到的日子中所佔的比例偏高，所以 p 必定小於 $\frac{3}{5}$，而事實正是如此。

常見錯誤

- 把標準差（standard deviation）相加，而不是把方差相加：King's Cross 路線的總時間方差為 $25 + 144 = 169$，所以其標準差是 13，而不是 $5 + 12 = 17$。
- 兩班火車都用同一個 105 分鐘的時限：往 Liverpool Street 的火車遲五分鐘開出。
- 直接令 $p = \frac{3}{5}$。這個 $\frac{3}{5}$ 是 $\mathrm{P}(K \mid D)$，而不是 $\mathrm{P}(K)$；兩者不同，是因為兩條路線並非同樣可靠。

STEP 之外

這裏有兩個想法，會在大學第一個概率課程中反覆出現。獨立正態變量之和仍是正態的這個事實，用矩母函數（moment generating function）來證明最為簡潔；它也是正態分佈族的穩定性（stability）的一個特例，而這種穩定性正是中心極限定理（central limit theorem）的基礎。最後一部分則是統計推論（statistical inference）的縮影：我們觀察到一個結果（遲到），然後運用貝葉斯定理推斷隱藏的原因（選乘哪班火車）。Grimmett 與 Welsh 的 *Probability: An Introduction*（概率入門）涵蓋了這兩個課題，而 Blitzstein 與 Hwang 的 *Introduction to Probability*（概率導論）則有更多正是這種風格的貝葉斯題目。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。