詳細解答

STEP II 1996，第 14 題

概率與統計 • 矩與中心極限定理

STEP II 1996，第 14 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/96-s2-q14

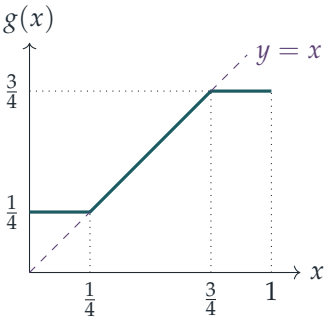
關鍵思路

Y 就是把 X 的兩端「截平」：所有低於 $\frac{1}{4}$ 的概率都堆在 $\frac{1}{4}$ 這一個值上，所有高於 $\frac{3}{4}$ 的概率都堆在 $\frac{3}{4}$ 上。所以 Y 是一個混合型隨機變量（mixed random variable），有兩個點質量（point mass）和中間一段連續部分，而任何期望值（expectation）$\mathrm{E}(g(Y))$ 也同樣分成三部分。最後一部分涉及大量獨立複本之和，這就要用上中心極限定理（central limit theorem）。為此我們只需要 Y 的平均值和方差（variance），而前面幾部分已經把這項工作做好了。

解答

Y 的矩

由於 $X \sim \mathrm{U}[0,1]$，我們有 $\mathrm{P}(X \le \frac{1}{4}) = \mathrm{P}(X \ge \frac{3}{4}) = \frac{1}{4}$，而在 $[\frac{1}{4}, \frac{3}{4}]$ 上 X 的概率密度（density）是 1。寫成 $Y = g(X)$，其中 g 是下圖所示的截平函數（clipping function），便得 $\mathrm{E}(Y^n) = \int_0^1 g(x)^n \,\mathrm{d}x$。



把積分在 $\frac{1}{4}$ 和 $\frac{3}{4}$ 處分開，

$$\begin{aligned}\mathrm{E}(Y^n) &= \left(\tfrac{1}{4}\right)^n \cdot \tfrac{1}{4} + \int_{1/4}^{3/4} x^n \,\mathrm{d}x + \left(\tfrac{3}{4}\right)^n \cdot \tfrac{1}{4} \\ &= \frac{1}{4^{n+1}} + \frac{3^{n+1}-1}{(n+1)4^{n+1}} + \frac{3^n}{4^{n+1}} \\ &= \frac{(n+1)(1+3^n)+3^{n+1}-1}{(n+1)4^{n+1}}.\end{aligned}$$

分子是 $n+1+(n+1)3^n+3\cdot 3^n-1 = n+(n+4)3^n$，所以

$$\boxed{\mathrm{E}(Y^n) = \frac{n+(n+4)\,3^n}{(n+1)\,4^{n+1}}} \qquad (n \geq 1).$$

與 X 比較

對於均勻分佈（uniform distribution）的變量，$\mathrm{E}(X^n) = \int_0^1 x^n \,\mathrm{d}x = \dfrac{1}{n+1}$。

取 $n=1$，公式給出 $\mathrm{E}(Y) = \dfrac{1+5\cdot 3}{2\cdot 16} = \dfrac{1}{2} = \mathrm{E}(X)$，正如所求。（這並不令人意外：截平把兩端的質量向內移動了相同的幅度，而且關於 $\frac{1}{2}$ 對稱。）

取 $n=2$，$\mathrm{E}(Y^2) = \dfrac{2+6\cdot 9}{3\cdot 64} = \dfrac{56}{192} = \dfrac{7}{24}$，而 $\mathrm{E}(X^2) = \frac{1}{3} = \frac{8}{24}$。因此

$$\mathrm{E}(X^2) - \mathrm{E}(Y^2) = \frac{8}{24} - \frac{7}{24} = \frac{1}{24},$$

正如所求。

對於一般的 $n \geq 2$，把兩個矩都寫成以 $(n+1)4^{n+1}$ 為公分母的形式：

$$\mathrm{E}(X^n) = \frac{4^{n+1}}{(n+1)4^{n+1}}.$$

所以我們要證明 $4^{n+1} > n+(n+4)3^n$。利用 $4^{n+1} = 4(3+1)^n$ 和二項式定理（binomial theorem），並把最高的兩項分開寫出，

$$4^{n+1} = 4\cdot 3^n + 4n\cdot 3^{n-1} + 4\sum_{k=0}^{n-2}\binom{n}{k}3^k.$$

減去 $n+(n+4)3^n = 4\cdot 3^n + 3n\cdot 3^{n-1} + n$，得

$$4^{n+1} - (n+(n+4)3^n) = n\,3^{n-1} - n + 4\sum_{k=0}^{n-2}\binom{n}{k}3^k.$$

? **想一想** 為甚麼這個差是正數？

當 $n \geq 2$ 時，$3^{n-1} \geq 3$，所以 $n\,3^{n-1} - n \geq 2n > 0$。那個和也是正數，因為它包含 $k = 0$ 的一項，其值為 1，而且每一項都是非負的。因此這個差嚴格為正，於是

$$\mathrm{E}(X^n) > \mathrm{E}(Y^n) \qquad \text{對所有} n \geq 2.$$

（當 $n = 1$ 時，這個差是 $3^0 - 1 + 0 = 0$，與 $\mathrm{E}(X) = \mathrm{E}(Y)$ 吻合。）

240 000 個複本之和

由上一部分，

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 = \frac{7}{24} - \frac{1}{4} = \frac{1}{24}.$$

設 $S = Y_1 + \cdots + Y_N$，其中 $N = 240\,000$。由於各 Y_i 互相獨立（independent），期望值和方差都可以相加：

$$\mathrm{E}(S) = \frac{N}{2} = 120\,000, \qquad \mathrm{Var}(S) = \frac{N}{24} = 10\,000,$$

所以 S 的標準差（standard deviation）是 100。各 Y_i 獨立同分佈（independent and identically distributed），方差有限，而 N 非常大，所以根據中心極限定理，S 近似服從 $\mathrm{N}(120\,000,\, 100^2)$。因此

$$\mathrm{P}(S < K) \approx \Phi\Big(\frac{K - 120\,000}{100}\Big) = \frac{3}{4} \quad \Longrightarrow \quad \frac{K - 120\,000}{100} = \Phi^{-1}(0.75) \approx 0.6745,$$

所以

$$K \approx 120\,000 + 67.45 \approx 120\,067.$$

這裏不需要連續性校正（continuity correction）：Y 並非只取整數值，而且標準差為 100，任何這類校正對這個準確度根本沒有影響。

驗算

- $\mathrm{E}(Y^n)$ 的公式所給的值介乎 $(\frac{1}{4})^n$ 與 $(\frac{3}{4})^n$ 之間，這是必然的，因為 $\frac{1}{4} \leq Y \leq \frac{3}{4}$。當 $n = 3$ 時，公式給出 $\frac{3+7\cdot 27}{4\cdot 256} = \frac{192}{1024} = \frac{3}{16} = 0.1875$，而用中點法則（midpoint rule）計算 $g(x)^3$ 的積分，結果吻合至小數點後十位。
- $\mathrm{Var}(Y) = \frac{1}{24} < \frac{1}{12} = \mathrm{Var}(X)$：截平把數值拉向平均值，所以離散程度理應下降。
- 模擬 20 000 個由 2400 個 Y 的複本相加而成的和，得出標準化後的上四分位數（upper quartile）約為 0.678，接近 0.6745，這支持了正態近似（normal approximation）。

常見錯誤

- 把 Y 當作在 $[\frac{1}{4}, \frac{3}{4}]$ 上均勻分佈，因而忘記了兩端各有大小為 $\frac{1}{4}$ 的點質量。在該區間上均勻分佈的變量，密度是 2，不是 1。

- 用了 $\operatorname{Var}(S) = N^2 \operatorname{Var}(Y)$ 而不是 $N \operatorname{Var}(Y)$。獨立性使方差可以相加，所以標準差按 $\sqrt{N}$ 的方式增長。
- 在最後一步弄錯正負號。$\mathrm{P}(S < K) = \frac{3}{4} > \frac{1}{2}$，所以 K 必定高於平均值。

STEP 之外

Y 是 X 經截平（又稱「縮尾」，Winsorised）後的版本，它的分佈既不是離散的，也不是連續的。測度論概率（measure-theoretic probability）通過分佈函數（distribution function）統一處理這類分佈，每個期望值都變成單一個積分 $\int g \, \mathrm{d}F$。在平均值相等的情況下，不等式 $\mathrm{E}(X^n) > \mathrm{E}(Y^n)$ 是凸序（convex ordering）的一個例子：X 是 Y 的「保均值擴散」（mean-preserving spread），所以對每個凸函數 ϕ 都有 $\mathrm{E}(\phi(X)) \geq \mathrm{E}(\phi(Y))$。最後用到的中心極限定理，其證明見 Grimmett and Welsh, *Probability: An Introduction*（一本概率論入門教科書）。

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。