詳細解答

STEP III 2025，第 11 題

概率與統計 • 最大值的分佈

STEP III 2025，第 11 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 OCR 提供下載的 2025 年 STEP 試卷（zip 檔案）內的 STEP III 試卷：

www.ocr.org.uk/Images/750486-2025-past-papers.zip

本題在 *Beyond STEP* 第 12.25 節眾多之中的最大值中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch12/extremes.html

關鍵思路

當且僅當每一個 X_i 都不超過 y 時，X_i 之中最大者才不超過 y；而獨立性（independence）把這件事化為一個乘積。此後，整道題就是從 G 讀出分位數（quantile），再看看當 n 很大時它們會怎樣變化。真正起作用的只有一個近似：$\alpha^{1/n} = \mathrm{e}^{(\ln\alpha)/n} \approx 1 + \frac{\ln\alpha}{n}$，因為當 n 很大時，$(\ln\alpha)/n$ 很小。

解答

第 (i) 部分

對於 $x \geq 0$，

$$F(x) = \int_0^x \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = 1 - \mathrm{e}^{-\lambda x}.$$

由於 $Y = \max(X_1, \ldots, X_n)$，事件 $\{Y \leq y\}$ 與「對每一個 i 都有 $X_i \leq y$」這個事件相同。各 X_i 互相獨立，所以對於 $y \geq 0$，

$$G(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(X_1 \leq y) \cdots \mathrm{P}(X_n \leq y) = F(y)^n = \left(1 - \mathrm{e}^{-\lambda y}\right)^n,$$

即為所求。

第 (ii) 部分

Y 是連續隨機變量（continuous random variable），所以 $\mathrm{P}(Y < y) = \mathrm{P}(Y \le y) = G(y)$。因此 $L(\alpha)$ 須滿足的條件是 $G(L(\alpha)) = \alpha$：

$$\begin{aligned}\left(1 - \mathrm{e}^{-\lambda L}\right)^n = \alpha &\iff 1 - \mathrm{e}^{-\lambda L} = \alpha^{1/n} \qquad \text{（取正的 } n \text{ 次方根）} \\ &\iff \mathrm{e}^{-\lambda L} = 1 - \alpha^{1/n} \\ &\iff L(\alpha) = -\frac{1}{\lambda}\ln\left(1 - \alpha^{1/n}\right).\end{aligned}$$

這個對數（logarithm）有定義，因為由 $0 < \alpha < 1$ 可得 $0 < \alpha^{1/n} < 1$。

至於 $U(\alpha)$，條件 $\mathrm{P}(Y > U) = \alpha$ 即 $G(U) = 1 - \alpha$，所以把 α 換成 $1 - \alpha$ 再重複同樣的步驟，便得

$$U(\alpha) = -\frac{1}{\lambda}\ln\left(1 - (1-\alpha)^{1/n}\right).$$

由於 $\alpha \le \frac{1}{2} \le 1 - \alpha$，而 G 是遞增函數，所以 $L(\alpha) \le U(\alpha)$，且只有當 $\alpha = \frac{1}{2}$ 時才取等號。

第 (iii) 部分

寫 $\alpha^{1/n} = \mathrm{e}^{(\ln\alpha)/n}$。當 α 固定而 n 很大時，$t = (\ln\alpha)/n$ 很小，所以由 $\mathrm{e}^t \approx 1 + t$ 得

$$1 - \alpha^{1/n} \approx -\frac{\ln\alpha}{n} = \frac{1}{n}\ln\left(\frac{1}{\alpha}\right).$$

留意這是正數，因為 $\alpha < 1$。代入第 (ii) 部分的結果，

$$\lambda L(\alpha) = -\ln\left(1 - \alpha^{1/n}\right) \approx -\ln\left(\frac{1}{n}\ln\frac{1}{\alpha}\right) = \ln(n) - \ln\left(\ln\frac{1}{\alpha}\right).$$

> ? **想一想** 對一個近似值取對數，是否安全？$1 - \alpha^{1/n}$ 本身非常小，所以 $1 - \alpha^{1/n}$ 的誤差相對於它本身可能很大。

關鍵在於**相對誤差**（relative error）。設 $t = (\ln\alpha)/n$，則 $1 - \mathrm{e}^t = -t\,(1 + \frac{t}{2} + \cdots)$，所以準確值與近似值之比是 $1 + O(1/n)$。取對數後，$\lambda L(\alpha)$ 的誤差是 $\ln(1 + O(1/n))$，它趨於 0。所以這個近似不僅在 n 很大時很好；兩邊之差確實趨於零。

第 (iv) 部分

Y 的中位數（median）m 滿足 $G(m) = \frac{1}{2}$，所以 $m = L(\frac{1}{2})$（它同時也是 $U(\frac{1}{2})$）。由第 (iii) 部分，

$$\lambda m \approx \ln n - \ln(\ln 2).$$

由於 $\ln(\ln 2)$ 是一個固定的數（約為 -0.37），而 $\ln n \to \infty$，所以當 n 增大時，中位數趨於無限大。

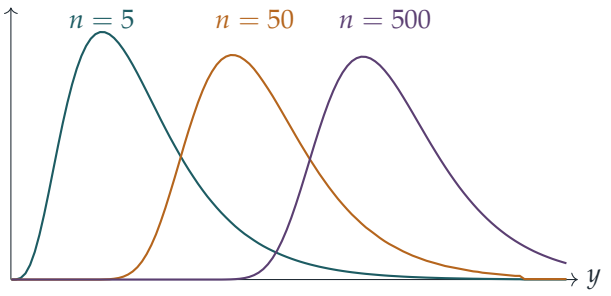
至於 $U(\alpha)$，在第 (iii) 部分中把 α 換成 $1-\alpha$：

$$\lambda U(\alpha) \approx \ln n - \ln\left(\ln \frac{1}{1-\alpha}\right).$$

兩式相減，$\ln n$ 項互相抵消：

$$U(\alpha) - L(\alpha) \approx \frac{1}{\lambda}\left[\ln\left(\ln \frac{1}{\alpha}\right) - \ln\left(\ln \frac{1}{1-\alpha}\right)\right].$$

由於兩個近似的誤差都趨於零（正如上面方框所述），區間寬度趨於這個值，它取決於 α 和 λ，但與 n 無關。所以當 n 增大時，Y 的分佈像 $\frac{1}{\lambda}\ln n$ 那樣往無限大漂移，但並不會擴散開來。



上圖顯示 $\lambda = 1$ 時 Y 的概率密度（density）：n 每增大十倍，曲線便向右移約 $\ln 10 \approx 2.3$，而曲線的形狀則趨於穩定。

第 (v) 部分

取 $\alpha = 0.05$。那麼

$$\mathrm{P}\big(L(\alpha) \le Y \le U(\alpha)\big) = 1 - \mathrm{P}(Y < L) - \mathrm{P}(Y > U) = 1 - 2(0.05) = 0.9,$$

所以無論 n 是多少，Y 落在區間 $[L(0.05), U(0.05)]$ 內的概率恰好是 0.9。當 n 很大時，由第 (iv) 部分，這個區間的寬度約為

$$\frac{1}{\lambda}\left[\ln(\ln 20) - \ln\left(\ln \frac{1}{0.95}\right)\right].$$

現在運用題目給出的兩個事實：

- $\mathrm{e}^3 \approx 20$，所以 $\ln 20 \approx 3$，而 $\ln(\ln 20) \approx \ln 3 \approx 1.1$；
- 利用 $\ln(1+t) \approx t$（取 $t = -0.05$），得 $\ln \frac{1}{0.95} = -\ln(1-0.05) \approx 0.05 = \frac{1}{20}$；所以 $\ln\left(\ln \frac{1}{0.95}\right) \approx \ln \frac{1}{20} = -\ln 20 \approx -3$。

因此寬度約為

$$\frac{1}{\lambda}(\ln 3 + 3) \approx \frac{4.1}{\lambda} \approx 4\lambda^{-1},$$

即為所求。

驗算

- 在第 (ii) 部分中，把 $L(\alpha)$ 代回 G，得 $\left(1-(1-\alpha^{1/n})\right)^n = \alpha$，完全吻合。
- 當 $\alpha = 0.05$ 而 $n = 10^4$ 時，$\lambda L(\alpha)$ 的準確值是 8.11330，而近似式 $\ln n - \ln\ln 20$ 給出 8.11315；當 $n = 10$ 時，兩者分別是 1.35 和 1.21，可見「足夠大」這個條件確實重要。
- 第 (v) 部分中寬度的準確極限值是 $4.067/\lambda$，與 $4/\lambda$ 相當接近。以電腦模擬五個指數分佈隨機變量的最大值，結果亦與 $G(y)$ 吻合至三位小數。

常見錯誤

- 寫成 $G(y) = 1-(\mathrm{e}^{-\lambda y})^n$：這是最小值的累積分佈函數（cumulative distribution function），而不是最大值的。
- 把 $\mathrm{P}(Y > U) = \alpha$ 寫成 $G(U) = \alpha$，而不是 $G(U) = 1-\alpha$。
- 在第 (v) 部分中選取 $\alpha = 0.1$：區間兩端的尾部各應留下 0.05，而不是 0.1。

STEP 之外

這是初窺極值理論（extreme value theory）。第 (iii) 部分其實是在說明 $\lambda Y - \ln n$ 有一個極限分佈（limiting distribution），且 $\mathrm{P}(\lambda Y - \ln n \le x) \to \exp(-\mathrm{e}^{-x})$：這就是甘貝爾分佈（Gumbel distribution），可用來為破紀錄的洪水、一年中最強的陣風以及最大額的保險索償建立模型。一條了不起的定理指出，經適當縮放的最大值只可能收斂到三個分佈族之一。Blitzstein 和 Hwang 的 *Introduction to Probability*（《概率導論》）討論了順序統計量（order statistics），是入門的好起點。

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。