詳細解答

STEP II 2023，第 12 題

概率與統計 • 最大值的分佈

STEP II 2023，第 12 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 2023 年 STEP 試卷（zip 檔案）內的 STEP II 試卷：

www.ocr.org.uk/Images/707102-2023-past-papers.zip

關鍵思路

若干個數的最大值不超過 t，當且僅當其中每一個都不超過 t，因此 Y 的累積分佈函數（cumulative distribution function）是單個 X_i 的累積分佈函數的冪。其餘一切都源於使用累積分佈函數而不是密度函數：中位數（median）可直接解方程 $\mathrm{P}(Y \le m) = \frac{1}{2}$ 得出，而均值（mean）則可透過分部積分（integration by parts）簡潔地求得，使積分中出現的是分佈函數而非密度函數。

解答

第 (i) 部分

記 $F(t) = \mathrm{P}(X_1 \le t)$。當 $0 \le t \le \pi$ 時，

$$F(t) = \int_0^t \tfrac{1}{2}\sin x \,\mathrm{d}x = \tfrac{1}{2}(1 - \cos t).$$

事件 $\{Y \le t\}$ 發生，當且僅當 $X_1 \le t$、$X_2 \le t$、…、$X_n \le t$ 全部發生：若最大者不超過 t，則其餘各個亦不超過 t，反之亦然。各 X_i 獨立同分佈（independent and identically distributed），故這個交集的概率等於各個概率之積：

$$\mathrm{P}(Y \le t) = \mathrm{P}(X_1 \le t)\,\mathrm{P}(X_2 \le t)\cdots\mathrm{P}(X_n \le t) = [\mathrm{P}(X_1 \le t)]^n.$$

因此，以 G 記 Y 的累積分佈函數，

$$G(t) = \frac{(1-\cos t)^n}{2^n} \qquad (0 \le t \le \pi),$$

且當 $t<0$ 時 $G(t)=0$，當 $t>\pi$ 時 $G(t)=1$。用鏈式法則（chain rule）求導，得 Y 的概率密度函數（probability density function）為

$$g(t)=\frac{n\,(1-\cos t)^{n-1}\sin t}{2^n}\quad \text{當}0\le t\le\pi,$$

其他情況下為零。

第 (ii) 部分

中位數 $m=m(n)$ 滿足 $G(m)=\frac{1}{2}$：

$$\left(\frac{1-\cos m}{2}\right)^n=\frac{1}{2}\iff\frac{1-\cos m}{2}=2^{-1/n}$$
$$\iff\cos m=1-2^{1-1/n},$$

由於 $(1-\cos m)/2\ge 0$，這裏取正實數 n 次方根。由於 m 必須位於 $[0,\pi]$ 內，而餘弦函數在此區間上是一一對應（one-to-one）的，

$$m(n)=\arccos(1-2^{1-1/n}).$$

當 n 增大時，$1/n$ 減小，故 $2^{1-1/n}$ 增大（趨向 2），而 $1-2^{1-1/n}$ 減小（趨向 -1）。由於 arccos 是遞減函數，$m(n)$ 隨 n 遞增，且當 $n\to\infty$ 時 $m(n)\to\arccos(-1)=\pi$。（當 $n=1$ 時得 $m(1)=\arccos 0=\pi/2$，這正符合 $\frac{1}{2}\sin x$ 關於 $\pi/2$ 的對稱性。）這是合理的：取越來越多個值的最大值，結果會被推向取值範圍的頂端。

第 (iii) 部分

按定義，$\mu(n)=\int_0^\pi t\,g(t)\,\mathrm{d}t$。由於 $g=G'$，取 $u=t$ 及 $\mathrm{d}v=g(t)\,\mathrm{d}t$、$v=G(t)$ 進行分部積分：

$$\mu(n)=\Big[t\,G(t)\Big]_0^\pi-\int_0^\pi G(t)\,\mathrm{d}t$$
$$=\pi\cdot 1-0-\frac{1}{2^n}\int_0^\pi(1-\cos x)^n\,\mathrm{d}x,$$

其中用到 $G(\pi)=(\frac{1+1}{2})^n=1$。這就是所求的公式。

想一想 為甚麼要用分部積分，而不是嘗試直接計算 $\int t\,(1-\cos t)^{n-1}\sin t\,\mathrm{d}t$？

目標公式含有 $(1-\cos x)^n$，即 $2^nG(x)$，是分佈函數而非密度函數。分部積分正是把 $t\,g(t)$ 換成 $G(t)$ 的一步，同時亦消去了麻煩的因子 t。

(a) 設 $h(x)=\frac{1}{2}(1-\cos x)$，則在 $[0,\pi]$ 上 $0\le h(x)\le 1$，且當 $0<x<\pi$ 時 $0<h(x)<1$。公式變為 $\mu(n)=\pi-\int_0^\pi h(x)^n\,\mathrm{d}x$，故

$$\mu(n+1)-\mu(n)=\int_0^\pi\left(h(x)^n-h(x)^{n+1}\right)\mathrm{d}x=\int_0^\pi h(x)^n(1-h(x))\,\mathrm{d}x.$$

在 $0 < x < \pi$ 上，被積函數連續且嚴格為正，因為在該處 $h(x) > 0$ 且 $1 - h(x) > 0$。所以積分嚴格為正，對每個 $n \geq 1$ 都有 $\mu(n+1) > \mu(n)$：$\mu(n)$ 隨 n 遞增。

(b) 取 $n = 2$，展開並利用 $\cos^2 x = \frac{1}{2}(1 + \cos 2x)$：

$$\int_0^{\pi} (1 - \cos x)^2 \, \mathrm{d}x = \int_0^{\pi} \left(1 - 2\cos x + \tfrac{1}{2} + \tfrac{1}{2}\cos 2x\right) \mathrm{d}x = \tfrac{3}{2}\pi,$$

因為 $\cos x$ 和 $\cos 2x$ 在 $[0, \pi]$ 上的積分均為零。因此

$$\mu(2) = \pi - \tfrac{1}{4} \cdot \tfrac{3}{2}\pi = \frac{5\pi}{8}.$$

由第 (ii) 部分，$m(2) = \arccos(1 - \sqrt{2})$。$5\pi/8$ 與 $m(2)$ 均位於 $[0, \pi]$ 內，而餘弦函數在此區間上嚴格遞減，故

$$\mu(2) < m(2) \iff \cos \tfrac{5\pi}{8} > \cos m(2) = 1 - \sqrt{2}.$$

> **想一想** 怎樣求得 $\cos \frac{5\pi}{8}$ 的準確值？

寫成 $\cos \frac{5\pi}{8} = \cos(\frac{\pi}{2} + \frac{\pi}{8}) = -\sin \frac{\pi}{8}$，並利用半角公式（half-angle formula）$\sin^2 \frac{\pi}{8} = \frac{1}{2}(1 - \cos \frac{\pi}{4}) = \frac{2 - \sqrt{2}}{4}$。由於 $\sin \frac{\pi}{8} > 0$，

$$\cos \tfrac{5\pi}{8} = -\tfrac{1}{2}\sqrt{2 - \sqrt{2}}.$$

所以我們需要證明 $-\frac{1}{2}\sqrt{2 - \sqrt{2}} > 1 - \sqrt{2}$，即 $\sqrt{2} - 1 > \frac{1}{2}\sqrt{2 - \sqrt{2}}$。兩邊均為正數，故此式等價於兩邊平方後的不等式，再乘以 4：

$$\begin{aligned} 4(3 - 2\sqrt{2}) > 2 - \sqrt{2} &\iff 10 > 7\sqrt{2} \\ &\iff 100 > 98, \end{aligned}$$

其中最後一步是把兩個正數平方。此式成立，而每一步都是等價變換，故 $\cos \frac{5\pi}{8} > 1 - \sqrt{2}$，因此 $\mu(2) < m(2)$。

驗算

- 當 $n = 1$ 時，$Y = X_1$ 關於 $\pi/2$ 對稱，而確實 $\mu(1) = \pi - \frac{1}{2}\int_0^{\pi}(1 - \cos x)\,\mathrm{d}x = \pi/2 = m(1)$。
- 數值上，$\mu(2) = 5\pi/8 \approx 1.9635$，$m(2) = \arccos(1 - \sqrt{2}) \approx 1.9979$，所以均值略低於中位數。對 n 由 1 至 7 的每個值各以電腦模擬 10^5 個最大值，結果與兩條公式的誤差均在 0.01 以內（例如 $\mu(3) \approx 2.1598$，$m(3) \approx 2.1986$），而且兩個數列都遞增。
- 當 $n = 2$ 時，Y 的密度函數為 $\frac{1}{2}(1 - \cos t)\sin t$，屬左偏（skewed to the left）分佈（有一條伸向 0 的長尾），而左偏分佈的均值通常低於中位數。

常見錯誤

- 忘記論證 $\mathrm{P}(Y \le t) = [F(t)]^n$：既需要「全部都不超過 t」這個等價關係，也需要獨立性。
- 做了分部積分，卻忘記邊界項 $\left[tG(t)\right]_0^{\pi} = \pi$，而開首的 π 正是由此而來。
- 透過餘弦比較角的大小時，沒有指出餘弦函數在 $[0, \pi]$ 上是遞減的，結果把不等號方向弄反。

STEP 之外

樣本的最大值是順序統計量（order statistic）的一個例子，而第 (i) 部分的技巧可以求出任何樣本最大值的分佈。恆等式 $\mathrm{E}[Y] = \pi - \int_0^{\pi} G(t)\,\mathrm{d}t$ 是非負隨機變量的一般尾部積分公式（tail-integral formula）$\mathrm{E}[Y] = \int_0^{\infty} \mathrm{P}(Y > t)\,\mathrm{d}t$ 的特例，你在大學概率課程的初期就會遇到。這兩個概念在 Grimmett 與 Welsh 的 *Probability: An Introduction*（《概率導論》）中均有深入討論。

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。