詳細解答

STEP II 2000，第 14 題

概率與統計 • 中位數的分佈

STEP II 2000，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/00-s2-q14

本題在 *Beyond STEP* 第 13.4 節平均數還是中位數？中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch13/mean-or-median.html

關鍵思路

中位數（median）的概率密度函數（probability density function）是形如 $y^r(1-y)^s$ 的多項式，所以題目給出的積分已包辦所有工作：先用一次求歸一化常數（normalising constant），再把 y 的冪次分別提高一和提高二，求出首兩個矩（moments）。至於最後一部分，Y 和 $\bar{X}$ 都以 $\frac{1}{2}$ 為中心，而對正態分佈（normal distribution）來說，落在距離中心某固定距離之內的概率只取決於標準差（standard deviation）。離散程度較小的那個變量勝出。

解答

常數 k

密度函數的積分必須等於 1。在題目給出的結果中取 $r = s = n$，

$$1 = \int_0^1 ky^n(1-y)^n \,\mathrm{d}y = k\,\frac{n!\,n!}{(2n+1)!}, \qquad \text{所以} \qquad k = \frac{(2n+1)!}{(n!)^2},$$

正如所求。

平均值

密度函數關於 $y = \frac{1}{2}$ 對稱（把 y 換成 $1-y$，函數不變），所以 $\mathrm{E}(Y) = \frac{1}{2}$。要用積分來確認這一點，取 $r = n+1$、$s = n$：

$$\mathrm{E}(Y) = k\int_0^1 y^{n+1}(1-y)^n \,\mathrm{d}y = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+1)!\,n!}{(2n+2)!} = \frac{n+1}{2n+2} = \frac{1}{2}.$$

方差

現在取 $r = n+2$、$s = n$：

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

因此

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}.$$

比較中位數與平均值

每個 X_i 都在 $[0,1]$ 上服從均勻分佈（uniform distribution），所以 $\mathrm{E}(X_i) = \frac{1}{2}$，而 $\mathrm{Var}(X_i) = \frac{1}{12}$。由於各 X_i 互相獨立（independent），

$$\mathrm{E}(\bar{X}) = \tfrac{1}{2}, \qquad \mathrm{Var}(\bar{X}) = \frac{1}{(2n+1)^2}\cdot(2n+1)\cdot\frac{1}{12} = \frac{1}{12(2n+1)}.$$

> **想一想** 兩個變量都以 $\frac{1}{2}$ 為中心。哪一個數值決定了誰較有可能落在距離中心 $d/\sqrt{n}$ 之內？

答案是標準差。當 n 很大時，我們可以取 $Y \sim \mathrm{N}(\frac{1}{2}, \sigma_Y^2)$ 和 $\bar{X} \sim \mathrm{N}(\frac{1}{2}, \sigma_{\bar{X}}^2)$。若 $W \sim \mathrm{N}(\frac{1}{2}, \sigma^2)$ 且 $c > 0$，則以 Φ 表示標準正態分佈函數（standard normal distribution function），有

$$\mathrm{P}\big(|W - \tfrac{1}{2}| < c\big) = \mathrm{P}\Big(|Z| < \frac{c}{\sigma}\Big) = 2\Phi\Big(\frac{c}{\sigma}\Big) - 1,$$

其中 $Z \sim \mathrm{N}(0,1)$。這是 σ 的嚴格遞減函數，因為 Φ 嚴格遞增。所以取 $c = d/\sqrt{n}$ 時，所求的不等式成立，當且僅當 $\sigma_Y > \sigma_{\bar{X}}$，即 $\mathrm{Var}(Y) > \mathrm{Var}(\bar{X})$：

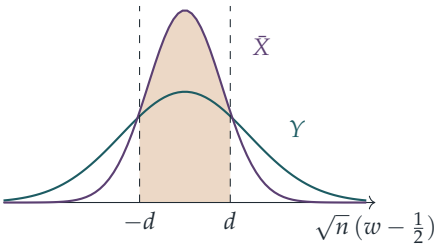
$$\begin{aligned}\frac{1}{4(2n+3)} > \frac{1}{12(2n+1)} &\iff 12(2n+1) > 4(2n+3)\\ &\iff 24n+12 > 8n+12\\ &\iff n > 0.\end{aligned}$$

這對每個 $n \geq 1$ 都成立。所以只要 n 大得足以令正態近似（normal approximation）有效，則不論正數 d 是甚麼，Y 的概率都嚴格小於 $\bar{X}$ 的概率。證明至此完成。

要看看差距有多大，可以把變量標準化：

$$\frac{d/\sqrt{n}}{\sigma_Y} = 2d\sqrt{\frac{2n+3}{n}} \to 2\sqrt{2}\,d, \qquad \frac{d/\sqrt{n}}{\sigma_X} = d\sqrt{\frac{12(2n+1)}{n}} \to 2\sqrt{6}\,d.$$

漸近而言，平均值的標準差是中位數的 $\sqrt{3}$ 分之一。



圖中以 $\sqrt{n}\,(w-\frac{1}{2})$ 為標度，畫出兩個極限密度函數。較窄的曲線（平均值）在任何一條中央帶內都有較大的概率。

驗算

- $n=0$：一個均勻變量的「中位數」就是 X_1 本身，而我們的公式給出 $k=1$ 和 $\mathrm{Var}(Y)=\frac{1}{12}$，正是均勻分佈的方差，與預期相符。
- 當 $n=0$ 時兩個方差都等於 $\frac{1}{12}$，而當 $n\geq 1$ 時兩者便分開，這與上面的不等式 $n>0$ 吻合。
- 以 $n=50$（即 101 個均勻變量）進行模擬，得到 $\mathrm{Var}(Y)\approx 0.00243$，而 $1/412\approx 0.00243$。取 $d=0.3$ 時，模擬得到 $\mathrm{P}(|Y-\frac{1}{2}|<d/\sqrt{n})\approx 0.607$ 和 $\mathrm{P}(|\bar{X}-\frac{1}{2}|<d/\sqrt{n})\approx 0.859$，與正態分佈給出的數值 0.611 和 0.860 相當接近。

常見錯誤

- 寫成 $\mathrm{Var}(\bar{X})=\frac{1}{12}$，或者只除以 $2n+1$ 一次，而不是先在分母得到 $(2n+1)^2$，再乘以 $2n+1$。
- 弄錯方向：方差較大，落在中心附近的概率反而較小。
- 忘記了計算 $\mathrm{E}(Y^2)$ 時用的是 $r=n+2$，而不是 $r=2n$。

STEP 之外

中位數的密度函數是一個貝塔分佈（Beta distribution）$\mathrm{Beta}(n+1,n+1)$，而題目給出的積分就是貝塔函數（Beta function）。你剛才所作的比較，正是*漸近相對效率*（asymptotic

relative efficiency）的概念：對均勻分佈的數據，用樣本平均值（sample mean）估計中心，其方差只有中位數的三分之一；但對厚尾（heavy-tailed）的數據，中位數卻可以勝出。順序統計量（order statistics）和貝塔函數在 Blitzstein 與 Hwang 合著的 *Introduction to Probability*（概率論入門）中有很好的講解。

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。