詳細解答

STEP III 1998，第 14 題

概率與統計 • 無偏估計量

STEP III 1998，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/98-s3-q14

本題在 *Beyond STEP* 第 13.3 節德國坦克問題中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch13/german-tanks.html

關鍵思路

Z_2 是若干獨立（independent）部分之和，所以它的平均值和方差（variance）可直接由單一個均勻分佈（uniform distribution）變量的平均值和方差得出。Z_1 取決於最大值，而處理最大值最自然的方法，是從它的分佈函數（distribution function）入手：各個 X_i 中最大的一個不超過 m，當且僅當每一個 X_i 都不超過 m。這樣便化成若干概率的乘積，再求導就得到密度函數（density）。

解答

Z_2 的平均值

每個 X_i 在 $[0, N]$ 上的密度為 $1/N$，所以

$$\mathrm{E}(X_i) = \frac{N}{2}, \qquad \mathrm{E}(X_i^2) = \frac{1}{N}\int_0^N x^2\,\mathrm{d}x = \frac{N^2}{3}, \qquad \mathrm{Var}(X_i) = \frac{N^2}{3} - \frac{N^2}{4} = \frac{N^2}{12}.$$

由期望值的線性性質（linearity of expectation），

$$\mathrm{E}(Z_2) = \frac{2}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{2}{n}\cdot n\cdot\frac{N}{2} = N.$$

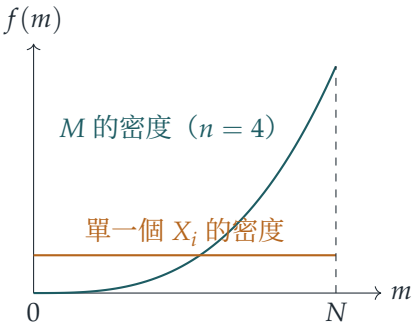
最大值的分佈

設 $M = \max\{X_1, \ldots, X_n\}$。對於 $0 \le m \le N$，事件 $M \le m$ 發生，當且僅當對每個 i 都有 $X_i \le m$；而各個 X_i 互相獨立，所以

$$\mathrm{P}(M \le m) = \prod_{i=1}^{n} \mathrm{P}(X_i \le m) = \left(\frac{m}{N}\right)^n.$$

求導後，得 M 的密度為

$$f(m) = \frac{n\, m^{n-1}}{N^n}, \qquad 0 \le m \le N.$$



最大值集中在 N 附近，所以只需乘上一個很小的修正因子，便可成為無偏（unbiased）估計量。它的矩（moments）為

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n\, m^{n-1}}{N^n}\, \mathrm{d}m = \frac{n}{N^n} \cdot \frac{N^{n+1}}{n+1} = \frac{n}{n+1} N,$$

$$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n\, m^{n-1}}{N^n}\, \mathrm{d}m = \frac{n}{n+2} N^2.$$

Z_1 的平均值

$$\mathrm{E}(Z_1) = \frac{n+1}{n}\,\mathrm{E}(M) = \frac{n+1}{n} \cdot \frac{n}{n+1} N = N.$$

因此兩個估計量（estimator）的平均值都是 N，正如題目所求。

方差

對於 Z_1，首先

$$\mathrm{Var}(M) = \frac{n}{n+2} N^2 - \frac{n^2}{(n+1)^2} N^2 = \frac{n\left[(n+1)^2 - n(n+2)\right]}{(n+2)(n+1)^2} N^2 = \frac{n}{(n+2)(n+1)^2} N^2,$$

因為 $(n+1)^2 - n(n+2) = 1$。乘以 $\left(\frac{n+1}{n}\right)^2$ 後，

$$\boxed{\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2} \cdot \frac{n}{(n+2)(n+1)^2} N^2 = \frac{N^2}{n(n+2)}.}$$

對於 Z_2，各個 X_i 互相獨立，所以它們的方差可以相加：

$$\boxed{\operatorname{Var}(Z_2) = \frac{4}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{4}{n^2}\cdot n\cdot\frac{N^2}{12} = \frac{N^2}{3n}.}$$

想一想 哪一個較小？是否永遠都是同一個較小？

兩者之比為

$$\frac{\operatorname{Var}(Z_1)}{\operatorname{Var}(Z_2)} = \frac{3n}{n(n+2)} = \frac{3}{n+2}.$$

當 $n = 1$ 時，這個比等於 1（此時 $Z_1 = Z_2 = 2X_1$，兩個估計量其實是同一個）；而對每個 $n \geq 2$，它都小於 1。

應選哪一個

兩者都是無偏的，所以我們比較它們的離散程度。我們選擇 Z_1：

- 當 $n \geq 2$ 時，它的方差嚴格較小，而且 n 越大，優勢越明顯：$\operatorname{Var}(Z_1)$ 以 $1/n^2$ 的速度下降，而 $\operatorname{Var}(Z_2)$ 只以 $1/n$ 的速度下降。例如當 $n = 10$ 時，Z_1 的標準差（standard deviation）約為 $0.09N$，而 Z_2 的則約為 $0.18N$。
- Z_1 必定不小於所見到的最大識別號碼，而 N 亦必須如此。Z_2 卻可能比我們實際觀察到的某個號碼還要小（例如讀數為 $1, 2, 30$ 時，得 $Z_2 = 22$），給出一個我們明知不可能的估計值。

驗算

- 當 $n = 1$ 時，兩條公式都給出 $N^2/3$；這是必然的，因為 $Z_1 = Z_2 = 2X_1$，而 $\operatorname{Var}(2X_1) = 4\cdot N^2/12$。
- 以 $N = 10$ 進行 200 000 次模擬試驗，兩個估計量的樣本平均值與 10 相差都不超過 0.01；當 $n = 4$ 時，樣本方差分別為 4.16 和 8.33，而理論值為 $\frac{100}{24} \approx 4.17$ 和 $\frac{100}{12} \approx 8.33$。

常見錯誤

- 試圖直接處理最大值的密度。應先求 $\mathrm{P}(M \leq m)$；由於互相獨立，它只是一個簡單的乘積。
- 忘記把常數平方：$\operatorname{Var}(cY) = c^2\operatorname{Var}(Y)$，所以兩個因子分別是 $\left(\frac{n+1}{n}\right)^2$ 和 $\frac{4}{n^2}$。

STEP 之外

這就是經典的「德國坦克問題」（German tank problem）：第二次世界大戰期間，盟軍的統計學家利用坦克的序號估計德國的坦克產量，結果準確得驚人。在這裏，最大值是一個*充分統計量*（sufficient statistic）：一旦知道了它，樣本的其餘部分就再不能提供關於 N 的任何額外資訊。萊曼–謝費定理（Lehmann–Scheffé theorem）進一步指出，在所有無偏估計量之中，Z_1 的方差最小。充分性和有效估計量（efficient estimator）是大學一年級統計學的課題；若想認識順序統計量（order statistics），Blitzstein 與 Hwang 合著的 *Introduction to Probability*（概率論入門）是很好的起點。

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。