详细解答

STEP III 1998，第 14 题

概率与统计 · 无偏估计量

STEP III 1998，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/98-s3-q14

本题在 *Beyond STEP* 第 13.3 节德国坦克问题中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/german-tanks.html

关键思路

Z_2 是若干独立（independent）部分之和，所以它的平均值和方差（variance）可直接由单一个均匀分布（uniform distribution）变量的平均值和方差得出。Z_1 取决于最大值，而处理最大值最自然的方法，是从它的分布函数（distribution function）入手：各个 X_i 中最大的一个不超过 m，当且仅当每一个 X_i 都不超过 m。这样便化成若干概率的乘积，再求导就得到密度函数（density）。

解答

Z_2 的平均值

每个 X_i 在 $[0, N]$ 上的密度为 $1/N$，所以

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

由期望值的线性性质（linearity of expectation），

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

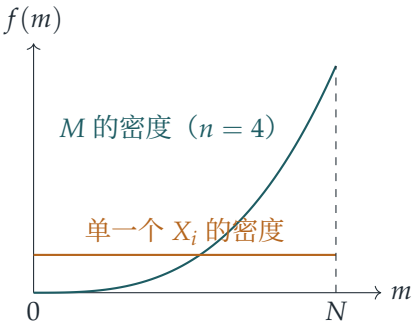
最大值的分布

设 $M = \max\{X_1, \ldots, X_n\}$。对于 $0 \le m \le N$，事件 $M \le m$ 发生，当且仅当对每个 i 都有 $X_i \le m$；而各个 X_i 互相独立，所以

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

求导后，得 M 的密度为

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



最大值集中在 N 附近，所以只需乘上一个很小的修正因子，便可成为无偏（unbiased）估计量。它的矩（moments）为

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

Z_1 的平均值

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

因此两个估计量（estimator）的平均值都是 N，正如题目所求。

方差

对于 Z_1，首先

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

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

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

对于 Z_2，各个 X_i 互相独立，所以它们的方差可以相加：

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

想一想 哪一个较小？是否永远都是同一个较小？

两者之比为

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

当 $n = 1$ 时，这个比等于 1（此时 $Z_1 = Z_2 = 2X_1$，两个估计量其实是同一个）；而对每个 $n \geq 2$，它都小于 1。

应选哪一个

两者都是无偏的，所以我们比较它们的离散程度。我们选择 Z_1：

- 当 $n \geq 2$ 时，它的方差严格较小，而且 n 越大，优势越明显：$\operatorname{Var}(Z_1)$ 以 $1/n^2$ 的速度下降，而 $\operatorname{Var}(Z_2)$ 只以 $1/n$ 的速度下降。例如当 $n = 10$ 时，Z_1 的标准差（standard deviation）约为 $0.09N$，而 Z_2 的则约为 $0.18N$。
- Z_1 必定不小于所见到的最大识别号码，而 N 亦必须如此。Z_2 却可能比我们实际观察到的某个号码还要小（例如读数为 $1, 2, 30$ 时，得 $Z_2 = 22$），给出一个我们明知不可能的估计值。

验算

- 当 $n = 1$ 时，两条公式都给出 $N^2/3$；这是必然的，因为 $Z_1 = Z_2 = 2X_1$，而 $\operatorname{Var}(2X_1) = 4\cdot N^2/12$。
- 以 $N = 10$ 进行 200 000 次模拟试验，两个估计量的样本平均值与 10 相差都不超过 0.01；当 $n = 4$ 时，样本方差分别为 4.16 和 8.33，而理论值为 $\frac{100}{24} \approx 4.17$ 和 $\frac{100}{12} \approx 8.33$。

常见错误

- 试图直接处理最大值的密度。应先求 $\mathrm{P}(M \leq m)$；由于互相独立，它只是一个简单的乘积。
- 忘记把常数平方：$\operatorname{Var}(cY) = c^2\operatorname{Var}(Y)$，所以两个因子分别是 $\left(\frac{n+1}{n}\right)^2$ 和 $\frac{4}{n^2}$。

STEP 之外

这就是经典的“德国坦克问题”（German tank problem）：第二次世界大战期间，盟军的统计学家利用坦克的序号估计德国的坦克产量，结果准确得惊人。在这里，最大值是一个*充分统计量*（sufficient statistic）：一旦知道了它，样本的其余部分就再不能提供关于 N 的任何额外信息。莱曼–谢费定理（Lehmann–Scheffé theorem）进一步指出，在所有无偏估计量之中，Z_1 的方差最小。充分性和有效估计量（efficient estimator）是大学一年级统计学的课题；若想认识顺序统计量（order statistics），Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）是很好的起点。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。