详细解答

STEP I 2008，第 12 题

概率与统计 • 离散均匀分布变量的最大值

STEP I 2008，第 12 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/08-s1-q12

关键思路

X 其实就是两个分数中较大的一个。若直接求“较大者等于 r”的概率，便要分几种情况来数；但“较大者不超过 r”却很容易处理：这恰好在两个分数都不超过 r 时发生。因此我们先求累积分布（cumulative distribution）$\mathrm{P}(X \leq r)$，再取差。有了这个累积分布，中位数（median）也几乎立即可得。

解答

设 X_1 和 X_2 为两个独立（independent）的分数，并设 $X = \max(X_1, X_2)$（当 $X_1 = X_2$ 时，定义中的两种情况给出相同的值，所以 X 是明确定义的）。

第 (i) 部分

对于 $r = 0, 1, \ldots, N$，事件 $\{X \leq r\}$ 就是事件 $\{X_1 \leq r$ 且 $X_2 \leq r\}$。这两个事件各自的概率都是 r/N，而且由于 X_1 和 X_2 独立，它们亦互相独立，所以

$$\mathrm{P}(X \leq r) = \frac{r}{N} \cdot \frac{r}{N} = \frac{r^2}{N^2}.$$

这在 $r = 0$ 时同样成立，所得的值为 0。因此，对于 $r = 1, 2, \ldots, N$，

$$\mathrm{P}(X = r) = \mathrm{P}(X \leq r) - \mathrm{P}(X \leq r - 1) = \frac{r^2 - (r-1)^2}{N^2} = \frac{2r - 1}{N^2},$$

这正是所要证明的结果。（简单核对一下：$\sum_{r=1}^{N}(2r - 1) = N^2$，所以这些概率之和为 1。）

第 (ii) 部分

利用题目给出的两个求和公式，

$$\begin{aligned}\mu = \sum_{r=1}^{N} r\,\frac{2r-1}{N^2} &= \frac{1}{N^2}\left(2\sum_{r=1}^{N} r^2 - \sum_{r=1}^{N} r\right)\\ &= \frac{1}{N^2}\left(\frac{N(N+1)(2N+1)}{3} - \frac{N(N+1)}{2}\right)\\ &= \frac{N+1}{N}\cdot\frac{2(2N+1)-3}{6} = \frac{(N+1)(4N-1)}{6N}.\end{aligned}$$

当 $N = 100$ 时，

$$\mu = \frac{101\times 399}{600} = \frac{40\,299}{600} = 67.165,$$

与题目所求相符。

第 (iii) 部分

由第 (i) 部分，$\mathrm{P}(X \le m) = m^2/N^2$，而 $\mathrm{P}(X \ge m) = 1 - \mathrm{P}(X \le m-1) = 1 - (m-1)^2/N^2$。中位数须满足的两个条件于是变成

$$\mathrm{P}(X \le m) \ge \tfrac{1}{2} \iff m^2 \ge \tfrac{1}{2}N^2 \iff m \ge \frac{N}{\sqrt{2}},$$

$$\mathrm{P}(X \ge m) \ge \tfrac{1}{2} \iff (m-1)^2 \le \tfrac{1}{2}N^2 \iff m-1 \le \frac{N}{\sqrt{2}},$$

这里取了正平方根；由于 $m \ge 1$，m 和 $m-1$ 都是非负数，所以这一步是成立的。因此 m 必须是满足

$$\frac{N}{\sqrt{2}} \le m \le \frac{N}{\sqrt{2}} + 1.$$

的整数。

> ? **想一想** 长度为 1 的区间可以包含两个整数。会不会有两个中位数?

只有当区间的两个端点都是整数时才会，而这需要 $N/\sqrt{2}$ 是整数。但若 $N/\sqrt{2} = k$，便有 $\sqrt{2} = N/k$，是一个有理数（rational number），这是不可能的。所以两个不等式都是严格不等式，而长度为 1 的开区间（open interval）恰好包含一个整数，于是

$$m = \left\lceil \frac{N}{\sqrt{2}} \right\rceil,$$

即大于 $N/\sqrt{2}$ 的最小整数。

当 $N = 100$ 时，$N/\sqrt{2} = 50\sqrt{2} \approx 70.71$，所以 $m = 71$。直接验证：$71^2 = 5041 \ge 5000 = \frac{1}{2}\times 100^2$，而 $70^2 = 4900 \le 5000$，所以两个条件都成立。

第 (iv) 部分

由第 (iii) 部分，$m = N/\sqrt{2} + \theta$，其中 $0 < \theta < 1$，所以

$$\frac{m}{N} = \frac{1}{\sqrt{2}} + \frac{\theta}{N} \to \frac{1}{\sqrt{2}} \quad \text{当} N \to \infty.$$

由第 (ii) 部分，

$$\frac{\mu}{N} = \frac{(N+1)(4N-1)}{6N^2} = \frac{1}{6}\left(1 + \frac{1}{N}\right)\left(4 - \frac{1}{N}\right) \to \frac{4}{6} = \frac{2}{3}.$$

因此，当 N 很大时，

$$\frac{\mu}{m} = \frac{\mu/N}{m/N} \approx \frac{2/3}{1/\sqrt{2}} = \frac{2\sqrt{2}}{3}.$$

验算

- 取 $N = 2$：四对等可能的结果中，$X = 1$ 出现一次，$X = 2$ 出现三次，与第 (i) 部分的 $\frac{1}{4}$ 和 $\frac{3}{4}$ 相符；而 $\mu = \frac{7}{4}$，与 $\frac{3\times 7}{12}$ 相符。
- 对 $N \le 40$，用电脑列出全部 N^2 对结果，可确认第 (i) 和第 (ii) 部分的结果完全正确；对 $N \le 2000$，把每个整数逐一代入中位数的定义检验，每次都恰好找到一个中位数，并且等于 $\lceil N/\sqrt{2} \rceil$。
- 当 $N = 10^5$ 时，$\mu/m \approx 0.942812$，而 $2\sqrt{2}/3 \approx 0.942809$。

常见错误

- 把 $\mathrm{P}(X = r)$ 算成 $2 \times \frac{r}{N} \cdot \frac{1}{N}$，忘记了这样会把 $X_1 = X_2 = r$ 的情况数了两次；正确的数目是 $2(r-1) + 1 = 2r - 1$。
- 写成 $\mathrm{P}(X \ge m) = 1 - m^2/N^2$：这其实是 $\mathrm{P}(X > m)$，而不是 $\mathrm{P}(X \ge m)$。
- 在 $N = 100$ 时答 $m = 70$，原因是把 70.71 四舍五入至最接近的整数，而不是向上取整。

STEP 之外

X 是*顺序统计量*（order statistic）最简单的例子：样本的最大值。当 $N \to \infty$ 时，X/N 趋近一个在 $[0,1]$ 上的连续随机变量（continuous random variable），其密度函数为 $2x$，平均值是 $\frac{2}{3}$，中位数是 $1/\sqrt{2}$，这正是第 (iv) 部分出现该比值的原因。较大样本的顺序统计量会引出贝塔分布（beta distribution）。Blitzstein 与 Hwang 合著的 *Introduction to Probability*（概率论入门）以浅白易明的方式讲解顺序统计量。

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。