详细解答

STEP III 1987，第 16 题

概率与统计 · 顺序统计量与泊松分布

STEP III 1987，第 16 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/87-s3-q16

关键思路

在第 (i) 部分，事件 “$A \le k$ 且 $B \ge 1-k$” 直接处理起来颇为棘手，但它的补集（complement）却是两个简单事件的并集：“所有 X_i 都大于 k” 和 “所有 X_i 都小于 $1-k$”。余下的工作交给容斥原理（inclusion–exclusion）便可。在第 (ii) 部分，得奖人数服从二项分布（binomial distribution），其中 n 极大而 p 极小，这正是泊松分布（Poisson distribution）大派用场的地方。因此我们先求出恰有一位或两位得奖者的泊松概率的最大值，然后再确保近似所带来的误差不会把答案推过 60%。

解答

第 (i) 部分

设 E 为事件 $A > k$，F 为事件 $B < 1-k$。那么 $A > k$ 表示每个 X_i 都大于 k，而 $B < 1-k$ 表示每个 X_i 都小于 $1-k$。由独立性（independence），

$$\mathrm{P}(E) = (1-k)^n, \qquad \mathrm{P}(F) = (1-k)^n, \qquad \mathrm{P}(E \cap F) = (1-2k)^n,$$

其中最后一式成立，是因为 $E \cap F$ 表示每个 X_i 都落在 $(k, 1-k)$ 之内，而这个区间的长度为 $1-2k > 0$。

我们要求的事件是 $E \cup F$ 的补集，所以由容斥原理得

$$p_n = 1 - \mathrm{P}(E \cup F) = 1 - 2(1-k)^n + (1-2k)^n.$$

由于 $0 < 1-2k < 1-k < 1$，两个幂都趋向 0，因此

$$p_n \to 1 \quad 当 n \to \infty.$$

*评注。*无论把 k 取得多小，只要观测值够多，我们便几乎可以肯定：有某个观测值与 0 相距不超过 k，又有另一个观测值与 1 相距不超过 k。换言之，样本最小值收敛于 0，样本最大值收敛于 1（均指依概率收敛），所以样本全距（sample range）$B - A$ 趋向 1，即整个区间的长度。这正是为甚么（举例说）A 和 B 是未知均匀分布（uniform distribution）两个端点的自然估计量（estimator）。

第 (ii) 部分

设 $N = 4\,000\,000$，并设 X 为得奖人数。各读者是否得奖互相独立，每人得奖的概率为 p，所以 $X \sim \mathrm{B}(N, p)$。记 $q = 1 - p$。你能保住工作，当且仅当 $X = 1$ 或 $X = 2$，其概率为

$$K(p) = Npq^{N-1} + \binom{N}{2} p^2 q^{N-2}.$$

*从泊松分布的角度看。*由于 N 极大而 p 很小，X 近似服从平均值为 $\lambda = Np$ 的泊松分布，所以

$$K(p) \approx h(\lambda), \qquad \text{其中} h(\lambda) = \mathrm{e}^{-\lambda}\Big(\lambda + \frac{\lambda^2}{2}\Big).$$

求导得

$$h'(\lambda) = \mathrm{e}^{-\lambda}\Big(1 + \lambda - \lambda - \frac{\lambda^2}{2}\Big) = \mathrm{e}^{-\lambda}\Big(1 - \frac{\lambda^2}{2}\Big),$$

它在 $0 < \lambda < \sqrt{2}$ 时为正，在 $\lambda > \sqrt{2}$ 时为负。所以 h 在 $\lambda = \sqrt{2}$ 处取得最大值：

$$h(\sqrt{2}) = \mathrm{e}^{-\sqrt{2}}(\sqrt{2} + 1) \approx 0.2431 \times 2.4142 \approx 0.5869.$$

你的最佳选择是 $p = \sqrt{2}/N \approx 3.5 \times 10^{-7}$，但即使如此，你保住工作的机会也只有约 58.7%。这就是题目所要求的推理。

> **想一想** 泊松分布只是一个近似。真正的二项概率会不会悄悄越过 0.6?

不会。我们可以直接为 $K(p)$ 找出上界。由于 $q \le 1$ 及 $\binom{N}{2} \le \frac{1}{2}N^2$，

$$K(p) = q^{N-2}\Big(Npq + \binom{N}{2}p^2\Big) \le q^{N-2}\Big(\lambda + \frac{\lambda^2}{2}\Big).$$

对所有 p 都有 $1 - p \le \mathrm{e}^{-p}$（直线 $1 - p$ 是凸曲线 e^{-p} 在 $p = 0$ 处的切线），所以 $q^{N-2} \le \mathrm{e}^{-(N-2)p} = \mathrm{e}^{-\mu}$，其中 $\mu = (N-2)p$。设 $c = N/(N-2)$，则 $\lambda = c\mu$，且 $c > 1$。于是

$$K(p) \le \mathrm{e}^{-\mu}\Big(c\mu + \frac{c^2\mu^2}{2}\Big) \le c^2\,\mathrm{e}^{-\mu}\Big(\mu + \frac{\mu^2}{2}\Big) = c^2 h(\mu) \le c^2 h(\sqrt{2}).$$

这里 $c^2 = (1 - \frac{2}{N})^{-2} < 1.000002$，所以无论 p 取甚么值，

$$K(p) < 1.000002 \times 0.58694 < 0.587 < 0.6.$$

因此，无论你怎样选择 p，保住工作的机会都少于 60%，这正是所要证明的。

验算

- 在第 (i) 部分，取 $n = 1$ 得 $p_1 = 1 - 2(1-k) + (1-2k) = 0$。这是对的：当 $k < \frac{1}{2}$ 时，单独一个观测值不可能既 $\leq k$ 又 $\geq 1-k$。取 $n = 2$ 则得 $p_2 = 2k^2$，即两点分别落在两端的窄带内（次序不拘）的概率。
- 以 $n = 10$、$k = 0.3$ 进行模拟，结果为 0.944，而公式给出 0.9436。
- 在 $\lambda = Np$ 的细密网格上计算精确的二项概率 $K(p)$，得出最大值为 0.58694，在 $\lambda \approx 1.414$ 处取得，与泊松值在小数点后五位吻合。

常见错误

- 写成 $p_n = \mathrm{P}(A \leq k)\,\mathrm{P}(B \geq 1-k)$。同一个样本的最小值和最大值并不独立。
- 直接对 p 求二项概率的最大值。这样做是可行的，但代数运算相当繁琐；改用 $\lambda = Np$ 来处理，问题便化为一行就能完成的微积分题。
- 做到泊松近似便停笔，没有留意到 0.587 与 0.6 相当接近，值得补上一些论证。

STEP 之外

第 (i) 部分让我们初尝*次序统计量*（order statistics）和*依概率收敛*（convergence in probability）的滋味：样本的最小值和最大值会收敛于支集（support）的两端。第 (ii) 部分是泊松的“稀有事件定律”（law of rare events），而我们所用的上界，正是以下一般事实的一个简单例子：当 p 很小时，二项分布与泊松分布在全变差（total variation）的意义下非常接近。这两个想法在 Grimmett 与 Welsh 的 *Probability: An Introduction*，以及 Blitzstein 与 Hwang 的 *Introduction to Probability* 中都有详细阐述。

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。