详细解答

STEP II 2002，第 12 题

概率与统计 · 正态近似与泊松近似

> **STEP II 2002，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/02-s2-q12

关键思路

这里有三层随机性，一层套一层：一枚硬币之内的正面次数、一天之内的硬币、一星期之内的日子。每一层都是一个二项计数，其“成功”概率来自下一层。公式的构造方法是：最底层用正态分布（normal distribution）近似（并作连续性校正（continuity correction）），中间层先用泊松分布（Poisson distribution）近似，再用正态分布近似，而最顶层则保留为精确的二项分布（binomial distribution）。一旦看出这三层结构，余下的工作就是辨认公式中每个式子出自哪一个近似，然后再问：对于题目给定的数值，每个近似是否可靠。

解答

某一天的一枚硬币

固定某一天和某一枚硬币。设 X 为该硬币 M 次投掷中正面的次数。各次投掷互相独立，每次掷出正面的概率为 p，所以

$$X \sim \mathrm{B}(M, p), \qquad \mathrm{E}(X) = Mp, \qquad \mathrm{Var}(X) = Mp(1-p).$$

当 M 很大时，我们以正态变量 $X' \sim \mathrm{N}(Mp,\, Mp(1-p))$ 近似 X。由于 X 取整数值，“正面少于 m 次”即 $X \le m-1$，作连续性校正后就变成 $X' < m - \frac{1}{2}$。所以

$$r = \mathrm{P}(X < m) = \mathrm{P}(X \le m-1) \approx \Phi\left(\frac{m - \frac{1}{2} - Mp}{\sqrt{Mp(1-p)}}\right) = \Phi\left(\frac{2m - 1 - 2Mp}{2\sqrt{Mp(1-p)}}\right).$$

与题目比较，这说明 $h \approx Lr$。

某一天的所有硬币

设 Y 为 L 枚硬币中，当天正面少于 m 次的硬币数目。各枚硬币的投掷互相独立，而且每枚硬币的表现相同，所以每枚硬币各自独立地以概率 r “成功”，于是

$$Y \sim \mathrm{B}(L,r), \qquad \mathrm{E}(Y) = Lr = h, \qquad \mathrm{Var}(Y) = Lr(1-r) = h(1-r).$$

想一想 q 的公式中，分母是 $\sqrt{h}$。为甚么标准差（standard deviation）是 $\sqrt{h}$ 而不是 $\sqrt{h(1-r)}$?

这条公式是为 r 很小的情况而设的：对单独一枚硬币来说，正面少于 m 次是罕见事件，而 L 则很大。在这种情况下 $1-r \approx 1$，所以方差（variance）近似等于平均值（mean），这正是泊松近似 $Y \approx \mathrm{Po}(h)$ 适用的情形。如果 h 也相当大，$\mathrm{Po}(h)$ 又可以用 $\mathrm{N}(h,h)$ 近似。（等价的做法是：直接以正态分布近似 $\mathrm{B}(L,r)$，再把 $1-r$ 换成 1。）

因此取 $Y' \sim \mathrm{N}(h,h)$。“多于 l 枚硬币”即 $Y \geq l+1$，作连续性校正后就变成 $Y' > l+\frac{1}{2}$。于是

$$q = \mathrm{P}(Y > l) \approx 1 - \Phi\left(\frac{l+\frac{1}{2}-h}{\sqrt{h}}\right) = \Phi\left(\frac{h-l-\frac{1}{2}}{\sqrt{h}}\right) = \Phi\left(\frac{2h-2l-1}{2\sqrt{h}}\right),$$

当中用到标准正态分布的对称性 $1-\Phi(z) = \Phi(-z)$。这就是题目给出的 q。

所有日子

如果某一天有多于 l 枚硬币的正面次数少于 m，就称这一天为好日子。每一天成为好日子的概率同样是 q，而不同日子涉及不同的投掷，所以各天互相独立。因此 K 天之中好日子的数目服从 $\mathrm{B}(K,q)$，而

$$\mathrm{P}(\text{恰好有 } k \text{ 个好日子}) = \binom{K}{k} q^k (1-q)^{K-k},$$

其中 q 按上文近似。这就是所要求的结果。

数值例子

这里 $Mp = 60$，$Mp(1-p) = 24$，所以

$$\frac{2m-1-2Mp}{2\sqrt{Mp(1-p)}} = \frac{96-1-120}{2\sqrt{24}} = \frac{-25}{9.798} \approx -2.55,$$

由此得 $r \approx \Phi(-2.55) \approx 0.0054$ 及 $h = 500r \approx 2.68$。然后

$$\frac{2h-2l-1}{2\sqrt{h}} \approx \frac{5.36-9}{3.27} \approx -1.11, \qquad q \approx \Phi(-1.11) \approx 0.133,$$

而近似公式给出 $\binom{7}{2} q^2(1-q)^5 \approx 21 \times 0.0178 \times 0.490 \approx 0.18$。

想一想 每一个近似都是标准做法。当中哪些是在其可靠条件以外使用的?

逐层检视这三层。

- **顶层**。$\mathrm{B}(7,q)$ 这一步是精确的（在 q 已知的前提下），所以它本身不会带来误差。但由于 $k=2$，它会把 q 的误差平方。
- **底层**。$M=100$，$Mp=60$，$M(1-p)=40$，完全处于通常使用正态分布近似二项分布的范围之内。但我们是在深入尾部 2.55 个标准差的位置使用它，用来求一个约为 0.005 的概率。就绝对误差而言，正态近似在分布的主体部分表现良好；但在尾部，它的相对误差可以很大，尤其因为 $p\neq\frac{1}{2}$ 令二项分布偏斜。r 的任何相对误差都会直接传到 $h=Lr$，而 q 对 h 十分敏感。
- **中间层**。以 $\mathrm{Po}(h)$ 取代 $\mathrm{B}(500,r)$ 效果极佳，因为 r 极小而 L 很大。但接着在 $h\approx 2.7$ 的情况下以正态分布近似 $\mathrm{Po}(h)$，却是没有根据的：平均值小于 3 的泊松分布明显偏斜，而常用的经验法则要求平均值至少约为 10（有人说 15 或 20）。

所以答案是否定的：我们不会预期这个近似准确。有疑问的步骤是对 $\mathrm{B}(100,0.6)$ 尾部的正态近似，以及对平均值很小的泊松分布的正态近似。

事实上，r 的精确值是 $\mathrm{P}(X\leq 47)\approx 0.00576$，比正态估计值高出约 7%，所以 h 应约为 2.88；而精确的 $\mathrm{P}(Y>4)$ 则约为 0.165，而不是 0.133。最终的概率约为 0.23，而不是 0.18。（就这组特定的数值而言，对 $\mathrm{Po}(2.68)$ 的正态近似所给出的尾部概率，碰巧与泊松分布本身几乎相同，即 0.133 对 0.134，所以大部分的偏差来自底层。但你事前不可能指望有这种运气。）

验算

- 连续性校正的方向正确：“少于 m” 即 $X\leq m-1$，所以我们在 $m-\frac{1}{2}$ 处截止；“多于 l” 即 $Y\geq l+1$，所以我们由 $l+\frac{1}{2}$ 开始。题目给出的公式中正好含有 $2m-1$ 和 $2l+1$。
- 数值上，$\Phi(-2.5516)=0.005362$，$h=2.681$，$q=\Phi(-1.1108)=0.1333$，由此得 0.1825。精确的二项求和给出 $r=0.005761$，$q=0.1645$（已由模拟 2×10^4 天加以确认，所得结果为 0.164），最终答案为 0.2314。

常见错误

- 用了 m 而不是 $m-\frac{1}{2}$（或用了 l 而不是 $l+\frac{1}{2}$）：这样公式就对不上。
- 用了 $\sqrt{h(1-r)}$，然后无法得出 $\sqrt{h}$。题目的公式略去了因子 $1-r$，因为它实际上是在使用泊松近似，而你应该指出这一点。

- 只凭“$Mp > 5$ 及 $M(1-p) > 5$”来判断准确度。这条法则针对的是分布的中心部分，而不是一个 0.005 的尾部概率。

STEP 之外

这道题目所关乎的，是误差如何在一连串近似之中传播，以及尾部的绝对误差与相对误差之间的分别。在大学里，中心极限定理（central limit theorem）会由 Berry–Esseen 定理进一步加强，该定理给出正态近似的绝对误差的上界。尾部概率则需要另一些工具，例如切尔诺夫界（Chernoff bounds）和大偏差理论（large-deviation theory），它们描述尾部概率缩小得有多快。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门）清楚地讲解了泊松近似、正态近似以及切尔诺夫型不等式；Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论导引）则是一本很好的辅助读物。

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。