詳細解答

STEP II 2002，第 12 題

概率與統計 • 正態近似與泊松近似

STEP II 2002，第 12 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/02-s2-q12

關鍵思路

這裏有三層隨機性，一層套一層：一枚硬幣之內的正面次數、一天之內的硬幣、一星期之內的日子。每一層都是一個二項計數，其「成功」概率來自下一層。公式的構造方法是：最底層用正態分佈（normal distribution）近似（並作連續性校正（continuity correction）），中間層先用泊松分佈（Poisson distribution）近似，再用正態分佈近似，而最頂層則保留為精確的二項分佈（binomial distribution）。一旦看出這三層結構，餘下的工作就是辨認公式中每個式子出自哪一個近似，然後再問：對於題目給定的數值，每個近似是否可靠。

解答

某一天的一枚硬幣

固定某一天和某一枚硬幣。設 X 為該硬幣 M 次投擲中正面的次數。各次投擲互相獨立，每次擲出正面的概率為 p，所以

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

當 M 很大時，我們以正態變量 $X' \sim \mathrm{N}(Mp,\, Mp(1-p))$ 近似 X。由於 X 取整數值，「正面少於 m 次」即 $X \le m-1$，作連續性校正後就變成 $X' < m - \frac{1}{2}$。所以

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

與題目比較，這說明 $h \approx Lr$。

某一天的所有硬幣

設 Y 為 L 枚硬幣中，當天正面少於 m 次的硬幣數目。各枚硬幣的投擲互相獨立，而且每枚硬幣的表現相同，所以每枚硬幣各自獨立地以概率 r「成功」，於是

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

想一想 q 的公式中，分母是 $\sqrt{h}$。為甚麼標準差（standard deviation）是 $\sqrt{h}$ 而不是 $\sqrt{h(1-r)}$？

這條公式是為 r 很小的情況而設的：對單獨一枚硬幣來說，正面少於 m 次是罕見事件，而 L 則很大。在這種情況下 $1-r \approx 1$，所以方差（variance）近似等於平均值（mean），這正是泊松近似 $Y \approx \mathrm{Po}(h)$ 適用的情形。如果 h 也相當大，$\mathrm{Po}(h)$ 又可以用 $\mathrm{N}(h,h)$ 近似。（等價的做法是：直接以正態分佈近似 $\mathrm{B}(L,r)$，再把 $1-r$ 換成 1。）

因此取 $Y' \sim \mathrm{N}(h,h)$。「多於 l 枚硬幣」即 $Y \geq l+1$，作連續性校正後就變成 $Y' > l + \frac{1}{2}$。於是

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

當中用到標準正態分佈的對稱性 $1-\Phi(z) = \Phi(-z)$。這就是題目給出的 q。

所有日子

如果某一天有多於 l 枚硬幣的正面次數少於 m，就稱這一天為**好日子**。每一天成為好日子的概率同樣是 q，而不同日子涉及不同的投擲，所以各天互相獨立。因此 K 天之中好日子的數目服從 $\mathrm{B}(K,q)$，而

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

其中 q 按上文近似。這就是所要求的結果。

數值例子

這裏 $Mp = 60$，$Mp(1-p) = 24$，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 $h = 500r \approx 2.68$。然後

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

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

想一想 每一個近似都是標準做法。當中哪些是在其可靠條件以外使用的？

逐層檢視這三層。

- **頂層**。$\mathrm{B}(7,q)$ 這一步是精確的（在 q 已知的前提下），所以它本身不會帶來誤差。但由於 $k=2$，它會把 q 的誤差平方。
- **底層**。$M=100$，$Mp=60$，$M(1-p)=40$，完全處於通常使用正態分佈近似二項分佈的範圍之內。但我們是在深入尾部 2.55 個標準差的位置使用它，用來求一個約為 0.005 的概率。就絕對誤差而言，正態近似在分佈的主體部分表現良好；但在尾部，它的相對誤差可以很大，尤其因為 $p\neq\frac{1}{2}$ 令二項分佈偏斜。r 的任何相對誤差都會直接傳到 $h=Lr$，而 q 對 h 十分敏感。
- **中間層**。以 $\mathrm{Po}(h)$ 取代 $\mathrm{B}(500,r)$ 效果極佳，因為 r 極小而 L 很大。但接着在 $h\approx 2.7$ 的情況下以正態分佈近似 $\mathrm{Po}(h)$，卻是沒有根據的：平均值小於 3 的泊松分佈明顯偏斜，而常用的經驗法則要求平均值至少約為 10（有人說 15 或 20）。

所以答案是否定的：我們不會預期這個近似準確。有疑問的步驟是對 $\mathrm{B}(100,0.6)$ 尾部的正態近似，以及對平均值很小的泊松分佈的正態近似。

事實上，r 的精確值是 $\mathrm{P}(X\leq 47)\approx 0.00576$，比正態估計值高出約 7%，所以 h 應約為 2.88；而精確的 $\mathrm{P}(Y>4)$ 則約為 0.165，而不是 0.133。最終的概率約為 0.23，而不是 0.18。（就這組特定的數值而言，對 $\mathrm{Po}(2.68)$ 的正態近似所給出的尾部概率，碰巧與泊松分佈本身幾乎相同，即 0.133 對 0.134，所以大部分的偏差來自底層。但你事前不可能指望有這種運氣。）

驗算

- 連續性校正的方向正確：「少於 m」即 $X\leq m-1$，所以我們在 $m-\frac{1}{2}$ 處截止；「多於 l」即 $Y\geq l+1$，所以我們由 $l+\frac{1}{2}$ 開始。題目給出的公式中正好含有 $2m-1$ 和 $2l+1$。
- 數值上，$\Phi(-2.5516)=0.005362$，$h=2.681$，$q=\Phi(-1.1108)=0.1333$，由此得 0.1825。精確的二項求和給出 $r=0.005761$，$q=0.1645$（已由模擬 2×10^4 天加以確認，所得結果為 0.164），最終答案為 0.2314。

常見錯誤

- 用了 m 而不是 $m-\frac{1}{2}$（或用了 l 而不是 $l+\frac{1}{2}$）：這樣公式就對不上。
- 用了 $\sqrt{h(1-r)}$，然後無法得出 $\sqrt{h}$。題目的公式略去了因子 $1-r$，因為它實際上是

在使用泊松近似，而你應該指出這一點。

- 只憑「$Mp > 5$ 及 $M(1-p) > 5$」來判斷準確度。這條法則針對的是分佈的中心部分，而不是一個 0.005 的尾部概率。

STEP 之外

這道題目所關乎的，是誤差如何在一連串近似之中傳播，以及尾部的絕對誤差與相對誤差之間的分別。在大學裏，中心極限定理（central limit theorem）會由 Berry–Esseen 定理進一步加強，該定理給出正態近似的絕對誤差的上界。尾部概率則需要另一些工具，例如切爾諾夫界（Chernoff bounds）和大偏差理論（large-deviation theory），它們描述尾部概率縮小得有多快。Blitzstein 與 Hwang 的 *Introduction to Probability*（概率論入門）清楚地講解了泊松近似、正態近似以及切爾諾夫型不等式；Grimmett 與 Welsh 的 *Probability: An Introduction*（概率論導引）則是一本很好的輔助讀物。

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。