详细解答

STEP I 1989，第 14 题

概率与统计 · 条件分布

STEP I 1989，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/89-s1-q14

关键思路

一切都建基于一组概率：种子落在每一条阔一里格的地带内的机会。种子当且仅当落在园林或农地时才会长成树，所以树的数目服从二项分布（binomial distribution）。树生长之处的距离 Y，其实就是在“落在陆地上”这个条件下的 L。它的密度函数（density）就是把 f 只保留陆地地带的部分，再按比例放大，使总面积等于 1。最后一部分是假设检验（hypothesis test）：对一个二项计数而言，农地上有 23 棵树是否多得出奇？

解答

各地带的概率

L 的累积分布函数（cumulative distribution function）在 $0 \le x \le 4$ 时为

$$\mathrm{F}(x) = \int_0^x \left(\tfrac{1}{2} - \tfrac{1}{8}t\right)\mathrm{d}t = \frac{x}{2} - \frac{x^2}{16},$$

所以 $\mathrm{F}(1) = \frac{7}{16}$、$\mathrm{F}(2) = \frac{3}{4}$、$\mathrm{F}(3) = \frac{15}{16}$ 及 $\mathrm{F}(4) = 1$。因此

$$\mathrm{P}(\text{园林}) = \mathrm{P}(0 \le L \le 1) = \tfrac{7}{16}, \qquad \mathrm{P}(\text{湖}) = \tfrac{3}{4} - \tfrac{7}{16} = \tfrac{5}{16},$$
$$\mathrm{P}(\text{农地}) = \mathrm{P}(2 \le L \le 3) = \tfrac{15}{16} - \tfrac{3}{4} = \tfrac{3}{16}, \qquad \mathrm{P}(\text{海}) = \tfrac{1}{16}.$$

（由于 L 是连续的，各边界的概率为零，所以边界属于哪一条地带并无影响。）一颗种子长成树的概率为

$$p = \tfrac{7}{16} + \tfrac{3}{16} = \tfrac{5}{8}.$$

设 T 为一年内长出的树的数目。由于 96 颗种子各自独立地被风吹走，每颗长成树的概率为 $\frac{5}{8}$，故 $T \sim \mathrm{B}(96, \frac{5}{8})$，而每年新长出的发烧树数目的平均值为

$$\mathrm{E}(T) = 96 \times \tfrac{5}{8} = 60.$$

第 (i) 部分

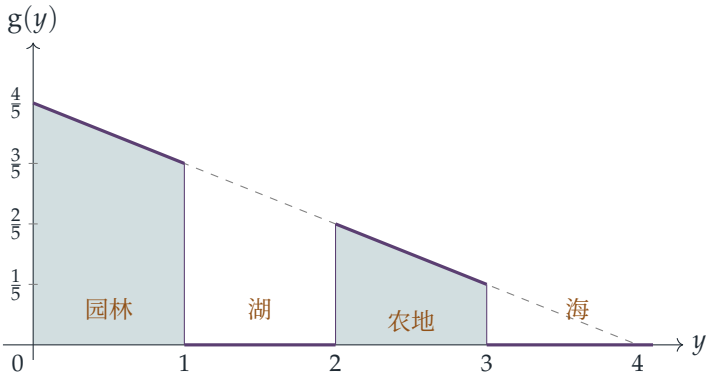
Y 的分布，就是已知种子落在陆地上时 L 的分布。以 $A = [0,1] \cup [2,3]$ 表示陆地。对于 $y \in A$，

$$\mathrm{P}(y \le Y \le y + \delta y) = \frac{\mathrm{P}(y \le L \le y + \delta y)}{\mathrm{P}(L \in A)} \approx \frac{\mathrm{f}(y)\,\delta y}{5/8},$$

而 Y 永远不会取 A 以外的值。所以 Y 的密度函数为

$$\mathrm{g}(y) = \begin{cases} \frac{8}{5}(\frac{1}{2} - \frac{1}{8}y) = \dfrac{4-y}{5} & \text{当} 0 \le y \le 1 \text{ 或} 2 \le y \le 3, \\ 0 & \text{其他情况。} \end{cases}$$

验算一下，它的面积是 $\frac{8}{5}(\frac{7}{16} + \frac{3}{16}) = 1$。图像是同一条下降直线上的两段：在园林上由 $\frac{4}{5}$ 降至 $\frac{3}{5}$，在湖上为零，然后在农地上由 $\frac{2}{5}$ 降至 $\frac{1}{5}$。



虚线是 $\frac{8}{5}\mathrm{f}$，画出来以作比较；Y 的密度函数只在陆地上才与它重合。

至于平均值，$\int y(4-y)\,\mathrm{d}y = 2y^2 - \frac{1}{3}y^3$，所以

$$\begin{aligned} \mathrm{E}(Y) &= \frac{1}{5}\left[2y^2 - \tfrac{1}{3}y^3\right]_0^1 + \frac{1}{5}\left[2y^2 - \tfrac{1}{3}y^3\right]_2^3 \\ &= \frac{1}{5}\left(\frac{5}{3}\right) + \frac{1}{5}\left(9 - \frac{16}{3}\right) = \frac{1}{5}\cdot\frac{5}{3} + \frac{1}{5}\cdot\frac{11}{3} = \frac{16}{15}. \end{aligned}$$

所以新树平均生长在城堡以南 $\frac{16}{15}$ 里格处。这个位置刚过了园林，落在湖中，但这并无不妥：平均值不一定是变量可以取到的值。

第 (ii) 部分

设 X 为一年内在农地长出的树的数目。若有依从传统，则放出的种子有 96 颗，每颗独立地以概率 $\frac{3}{16}$ 落在农地，所以

$$X \sim \mathrm{B}\left(96, \tfrac{3}{16}\right), \qquad \mathrm{E}(X) = 18, \qquad \mathrm{Var}(X) = 96 \cdot \tfrac{3}{16} \cdot \tfrac{13}{16} = \tfrac{117}{8} \approx 14.6,$$

由此得标准差（standard deviation）约为 3.82。

想一想 我们究竟在检验甚么？朝哪个方向检验？

零假设（null hypothesis）是传统得到依从：种子有 96 颗，所以 $X \sim \mathrm{B}\left(96, \frac{3}{16}\right)$。农地上的树比平常多，最自然的怀疑是放出的种子多于 96 颗，所以我们采用单尾检验（one-tailed test），问在零假设下出现 23 棵或以上的可能性有多大。

由于 $np = 18$ 和 $n(1-p) = 78$ 都够大，我们采用正态近似（normal approximation）$X \approx \mathrm{N}(18, 14.625)$，并作连续性修正（continuity correction）：

$$\mathrm{P}(X \geq 23) \approx \mathrm{P}\left(Z \geq \frac{22.5 - 18}{3.824}\right) = \mathrm{P}(Z \geq 1.18) \approx 0.12.$$

（二项分布的精确值是 0.121。）所以在正常情况下，农地上有 23 棵或以上的树，大约每八年便有一年出现。这远远未达 5% 水平（甚至 10% 水平）的显著性。这个观察结果与传统得到依从的说法相当吻合，单凭这项证据，国王的怀疑并无充分理由。

作为比较，即使在 5% 水平下，数目也要达到至少 25（精确值 $\mathrm{P}(X \geq 25) \approx 0.049$），国王才算有合理的根据。假如他同时掌握园林的树木数目，便可以检验树的总数 $T \sim \mathrm{B}\left(96, \frac{5}{8}\right)$，这对于判断放出了多少颗种子，会是好得多的证据。

验算

- 四个地带的概率 $\frac{7}{16}, \frac{5}{16}, \frac{3}{16}, \frac{1}{16}$ 相加得 1，而 Y 的密度函数的总面积为 1。
- 用中点法则作数值积分得 $\mathrm{E}(Y) = 1.06667$，与 $\frac{16}{15}$ 吻合。模拟 400 000 颗种子（取 $L = 4 - 4\sqrt{1-U}$，其中 U 在 $[0,1]$ 上均匀分布），结果有 62.5% 落在陆地上，而这些种子的平均落点距离为 1.069。
- 把二项概率逐项精确相加，得 $\mathrm{P}(X \geq 23) = 0.1213$，与正态估计值 0.120 相近。

常见错误

- 直接把 f 当作 Y 的密度函数。它在陆地上的面积只有 $\frac{5}{8}$，所以必须除以 $\frac{5}{8}$。
- 计算了 $\mathrm{E}(L;\, L \in A) = \int_A y\,\mathrm{f}(y)\,\mathrm{d}y = \frac{2}{3}$，却忘记除以 $\mathrm{P}(L \in A)$。

- 检验 $\mathrm{P}(X = 23)$ 而非尾部概率 $\mathrm{P}(X \geq 23)$。X 取任何一个个别数值的可能性本来就不大。

STEP 之外

第 (i) 部分是*条件分布*（conditional distribution）的一个例子：把随机变量限制在某个事件上，再重新归一化。密度函数形式的贝叶斯定理（Bayes′ theorem），以及统计学中随处可见的截尾分布（truncated distribution），背后都是同一种运算。第 (ii) 部分让你初尝假设检验和 p 值，大学一年级的课程会用奈曼－皮尔逊（Neyman–Pearson）框架把这些概念严格化。Blitzstein 与 Hwang 的 *Introduction to Probability*（概率论入门教材）讲解条件分布时，附有大量像本题这样带故事情境的例题。

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。