详细解答

STEP II 2002，第 14 题

概率与统计 · 混合分布与贝叶斯定理

STEP II 2002，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/02-s2-q14

关键思路

这是一个分两阶段进行的试验：先选出一个区域，然后该区域内发生意外。因此应以区域为条件来考虑。全期望定律（law of total expectation）给出平均值，全概率定律（law of total probability）给出分布，最后由贝叶斯定理（Bayes' theorem）把问题反过来问。题目只给出平均值，所以我们以泊松分布（Poisson distribution）作为某区域内意外宗数的模型；而要证明的公式带有 $\mathrm{e}^{-\lambda}\lambda^k/k!$ 的形式，也印证了这个模型。

解答

设置

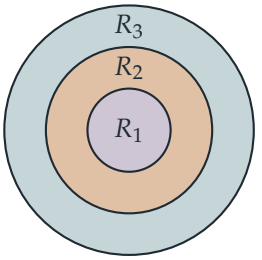
设 S 为所选区域的编号，$\lambda_n = 2 - \frac{n}{N}$ 为 R_n 内意外宗数的平均值。区域 R_n 是一个圆环（annulus），内外半径分别为 $n-1$ 和 n，所以它的面积是

$$\pi n^2 - \pi(n-1)^2 = \pi(2n-1).$$

整个岛的面积是 πN^2。（验算一下：$\sum_{n=1}^{N}(2n-1) = N^2$。）因此

$$\mathrm{P}(S=n) = \frac{2n-1}{N^2}, \qquad n = 1, \ldots, N.$$

在 $S = n$ 的条件下，我们把 X 看成平均值为 λ_n 的泊松变量。由于对每个 n 都有 $\lambda_n \geq 1 > 0$，这些泊松分布全都有意义。



区域面积：$R_n = \pi(2n-1)$

长期平均值

根据大数定律（law of large numbers），长远而言，每天记录到的平均宗数就是 $\mathrm{E}(X)$。以区域为条件，可得

$$\begin{aligned}\mathrm{E}(X) &= \sum_{n=1}^{N} \mathrm{P}(S=n)\,\mathrm{E}(X \mid S=n) = \sum_{n=1}^{N} \frac{2n-1}{N^2}\Big(2-\frac{n}{N}\Big)\\ &= 2\sum_{n=1}^{N}\frac{2n-1}{N^2} \;-\; \frac{1}{N^3}\sum_{n=1}^{N}(2n^2-n)\\ &= 2-\frac{1}{N^3}\Big(\frac{N(N+1)(2N+1)}{3}-\frac{N(N+1)}{2}\Big).\end{aligned}$$

第一个和等于 1，因为各区域被选中的概率加起来是 1。括号内我们用了题目给出的 $\sum n^2$ 公式，以及 $\sum n = \frac{1}{2}N(N+1)$。抽出 $\frac{1}{6}N(N+1)$ 后，余下 $2(2N+1)-3 = 4N-1$，所以

$$\mathrm{E}(X) = 2-\frac{(N+1)(4N-1)}{6N^2} = 2-\frac{1}{6}\Big(1+\frac{1}{N}\Big)\Big(4-\frac{1}{N}\Big),$$

其中把 N^2 拆成 $N\cdot N$，分给两个因子。这就是所要求的结果。

X 的分布

根据全概率定律，对 $k = 0, 1, 2, \ldots$，

$$\mathrm{P}(X=k) = \sum_{n=1}^{N}\mathrm{P}(S=n)\,\mathrm{P}(X=k \mid S=n) = \sum_{n=1}^{N}\frac{2n-1}{N^2}\cdot\frac{\mathrm{e}^{-\lambda_n}\lambda_n^k}{k!}.$$

现在把每一部分用 N 表示：

$$\lambda_n^k = \Big(\frac{2N-n}{N}\Big)^k = N^{-k}(2N-n)^k, \qquad \mathrm{e}^{-\lambda_n} = \mathrm{e}^{-2+n/N} = \mathrm{e}^{-2}\,\mathrm{e}^{n/N}.$$

因子 e^{-2}、N^{-2}、N^{-k} 和 $1/k!$ 都与 n 无关，所以可以提到求和符号之外：

$$\mathrm{P}(X=k) = \frac{\mathrm{e}^{-2}N^{-k-2}}{k!}\sum_{n=1}^{N}(2n-1)(2N-n)^k\,\mathrm{e}^{n/N},$$

正如所求。

$N=3$、两宗意外的情况

我们要求 $\mathrm{P}(S=2 \mid X=2)$。根据贝叶斯定理，

$$\mathrm{P}(S=2 \mid X=2)=\frac{\mathrm{P}(S=2)\,\mathrm{P}(X=2 \mid S=2)}{\mathrm{P}(X=2)}.$$

想一想 我们需要把 $\mathrm{P}(X=2)$ 完整地计算出来吗?

不需要。分子正是 $\mathrm{P}(X=2)$ 的求和式中 $n=2$ 的那一项。因此公因子 $\mathrm{e}^{-2}N^{-k-2}/k!$ 可以约去，所求概率就是 $\sum(2n-1)(2N-n)^k\mathrm{e}^{n/N}$ 中 $n=2$ 的一项除以整个和。

取 $N=3$ 和 $k=2$，各项 $(2n-1)(6-n)^2\mathrm{e}^{n/3}$ 分别是

$$n=1:\ 1\cdot 25\,\mathrm{e}^{1/3},\qquad n=2:\ 3\cdot 16\,\mathrm{e}^{2/3}=48\,\mathrm{e}^{2/3},\qquad n=3:\ 5\cdot 9\,\mathrm{e}=45\,\mathrm{e}.$$

于是，把分子和分母同除以 $\mathrm{e}^{2/3}$，得

$$\mathrm{P}(S=2 \mid X=2)=\frac{48\,\mathrm{e}^{2/3}}{25\,\mathrm{e}^{1/3}+48\,\mathrm{e}^{2/3}+45\,\mathrm{e}}=\frac{48}{48+45\,\mathrm{e}^{1/3}+25\,\mathrm{e}^{-1/3}},$$

这就是所要求的结果。

验算

- 当 $N=1$ 时只有一个区域，平均值为 1，而公式给出 $2-\frac{1}{6}\cdot 2\cdot 3=1$。当 $N\to\infty$ 时，平均值趋向 $2-\frac{4}{6}=\frac{4}{3}$。这与 $\int_0^1 2x(2-x)\,\mathrm{d}x=\frac{4}{3}$ 吻合，其中均匀选取的点位于相对半径 x 处的密度为 $2x$。
- 当 $N=3$ 时，平均值是 $\frac{32}{27}\approx 1.185$。模拟 4×10^5 天（在圆盘内均匀地选一点，然后抽取一个泊松计数），得出的平均值是 1.182。模拟同时得出 $\mathrm{P}(X=2)\approx 0.209$（准确值为 0.2094），而在发生两宗意外的日子当中，落在 R_2 的比例为 0.372。最后答案的数值是 0.3729。
- 把分布公式对 k 求和，在我们试过的每个 N 之下，总概率都是 1。

常见错误

- 以相等的概率 $1/N$ 选取区域，而不是按面积 $2n-1$ 的比例选取。
- 把各个泊松分布的*平均值*取平均，然后声称 X 是以该值为平均值的泊松变量。泊松分布的混合并不是泊松分布。只有期望值才能这样简单地通过取平均而得出。

STEP 之外

变量 X 服从一个混合分布（mixture distribution）：区域充当隐变量（hidden variable），而最后一部分就是对它作贝叶斯推断（Bayesian inference）。上面第二个错误可以用全方差定律（law of total variance）来解释：$\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X \mid S)) + \mathrm{Var}(\mathrm{E}(X \mid S))$，它说明混合分布总比平均值相同的单一泊松分布更为分散。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门）对条件期望（conditional expectation）和这类混合分布有很好的讲解。

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。