詳細解答

STEP II 2002，第 14 題

概率與統計 • 混合分佈與貝葉斯定理

STEP II 2002，第 14 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-q14

關鍵思路

這是一個分兩階段進行的試驗：先選出一個區域，然後該區域內發生意外。因此應以區域為條件來考慮。全期望定律（law of total expectation）給出平均值，全概率定律（law of total probability）給出分佈，最後由貝葉斯定理（Bayes' theorem）把問題反過來問。題目只給出平均值，所以我們以泊松分佈（Poisson distribution）作為某區域內意外宗數的模型；而要證明的公式帶有 $\mathrm{e}^{-\lambda}\lambda^k/k!$ 的形式，也印證了這個模型。

解答

設定

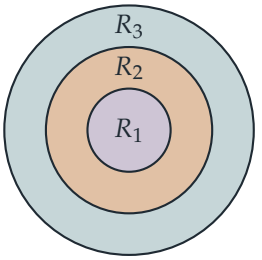
設 S 為所選區域的編號，$\lambda_n = 2 - \frac{n}{N}$ 為 R_n 內意外宗數的平均值。區域 R_n 是一個圓環（annulus），內外半徑分別為 $n-1$ 和 n，所以它的面積是

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

整個島的面積是 πN^2。（驗算一下：$\sum_{n=1}^{N}(2n-1) = N^2$。）因此

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

在 $S=n$ 的條件下，我們把 X 看成平均值為 λ_n 的泊松變量。由於對每個 n 都有 $\lambda_n \geq 1 > 0$，這些泊松分佈全都有意義。



區域面積：$R_n = \pi(2n-1)$

長期平均值

根據大數定律（law of large numbers），長遠而言，每天記錄到的平均宗數就是 $\mathrm{E}(X)$。以區域為條件，可得

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

第一個和等於 1，因為各區域被選中的概率加起來是 1。括號內我們用了題目給出的 $\sum n^2$ 公式，以及 $\sum n = \frac{1}{2}N(N+1)$。抽出 $\frac{1}{6}N(N+1)$ 後，餘下 $2(2N+1)-3 = 4N-1$，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 $N\cdot N$，分給兩個因子。這就是所要求的結果。

X 的分佈

根據全概率定律，對 $k = 0, 1, 2, \ldots$，

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

現在把每一部分用 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 $1/k!$ 都與 n 無關，所以可以提到求和符號之外：

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

正如所求。

$N = 3$、兩宗意外的情況

我們要求 $\mathrm{P}(S = 2 \mid X = 2)$。根據貝葉斯定理，

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

想一想 我們需要把 $\mathrm{P}(X = 2)$ 完整地計算出來嗎？

不需要。分子正是 $\mathrm{P}(X = 2)$ 的求和式中 $n = 2$ 的那一項。因此公因子 $\mathrm{e}^{-2}N^{-k-2}/k!$ 可以約去，所求概率就是 $\sum(2n-1)(2N-n)^k\mathrm{e}^{n/N}$ 中 $n = 2$ 的一項除以整個和。

取 $N = 3$ 和 $k = 2$，各項 $(2n-1)(6-n)^2\mathrm{e}^{n/3}$ 分別是

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

於是，把分子和分母同除以 $\mathrm{e}^{2/3}$，得

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

這就是所要求的結果。

驗算

- 當 $N = 1$ 時只有一個區域，平均值為 1，而公式給出 $2 - \frac{1}{6} \cdot 2 \cdot 3 = 1$。當 $N \to \infty$ 時，平均值趨向 $2 - \frac{4}{6} = \frac{4}{3}$。這與 $\int_0^1 2x(2-x)\,\mathrm{d}x = \frac{4}{3}$ 吻合，其中均勻選取的點位於相對半徑 x 處的密度為 $2x$。
- 當 $N = 3$ 時，平均值是 $\frac{32}{27} \approx 1.185$。模擬 4×10^5 天（在圓盤內均勻地選一點，然後抽取一個泊松計數），得出的平均值是 1.182。模擬同時得出 $\mathrm{P}(X = 2) \approx 0.209$（準確值為 0.2094），而在發生兩宗意外的日子當中，落在 R_2 的比例為 0.372。最後答案的數值是 0.3729。
- 把分佈公式對 k 求和，在我們試過的每個 N 之下，總概率都是 1。

常見錯誤

- 以相等的概率 $1/N$ 選取區域，而不是按面積 $2n-1$ 的比例選取。
- 把各個泊松分佈的平均值取平均，然後聲稱 X 是以該值為平均值的泊松變量。泊松分佈的混合並不是泊松分佈。只有期望值才能這樣簡單地通過取平均而得出。

STEP 之外

變量 X 服從一個混合分佈（mixture distribution）：區域充當隱變量（hidden variable），而最後一部分就是對它作貝葉斯推斷（Bayesian inference）。上面第二個錯誤可以用全方差定律（law of total variance）來解釋：$\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X \mid S)) + \mathrm{Var}(\mathrm{E}(X \mid S))$，它說明混合分佈總比平均值相同的單一泊松分佈更為分散。Grimmett 與 Welsh 所著的 *Probability: An Introduction*（概率論入門）對條件期望（conditional expectation）和這類混合分佈有很好的講解。

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。