풀이

STEP II 2002, 문제 14

확률과 통계 • 혼합분포와 베이즈 정리

STEP II 2002, 14번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/02-s2-q14

핵심 아이디어

이것은 두 단계로 이루어진 실험이다. 먼저 한 구역이 선택되고, 그다음 그 구역에서 사고가 일어난다. 따라서 구역에 대해 조건을 건다(condition). 기댓값은 전체 기댓값의 법칙(law of total expectation)으로, 분포는 전확률 법칙(law of total probability)으로 구하고, 마지막에는 베이즈 정리(Bayes' theorem)로 질문의 방향을 뒤집는다. 한 구역의 사고 건수는 푸아송 분포(Poisson distribution)로 모형화한다. 문제가 평균만 주기 때문이며, $\mathrm{e}^{-\lambda}\lambda^k/k!$ 꼴을 가진 목표 공식이 이 모형이 맞음을 확인해 준다.

풀이

설정

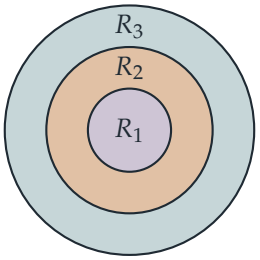
선택된 구역의 번호를 S라 하고, R_n에서의 평균 사고 건수를 $\lambda_n = 2 - \frac{n}{N}$이라 하자. 구역 R_n은 반지름이 $n-1$과 n인 원환(annulus)이므로 그 넓이는

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

이다. 섬 전체의 넓이는 πN^2이다. (검산하면 $\sum_{n=1}^{N}(2n-1) = N^2$이다.) 따라서

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

$S = n$일 때 X를 평균이 λ_n인 푸아송 분포로 모형화한다. 모든 n에 대해 $\lambda_n \geq 1 > 0$이므로 이 푸아송 분포들은 모두 의미가 있다.



넓이 $R_n = \pi(2n-1)$

장기적인 평균

큰 수의 법칙(law of large numbers)에 의해, 장기적으로 하루에 기록되는 평균 건수는 $\mathrm{E}(X)$이다. 구역에 대해 조건을 걸면

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

이다. 선택 확률의 합이 1이므로 첫 번째 합은 1이다. 괄호 안에서는 주어진 $\sum n^2$ 공식과 $\sum n = \frac{1}{2}N(N+1)$을 사용하였다. $\frac{1}{6}N(N+1)$을 묶어 내면 $2(2N+1)-3 = 4N-1$이 남으므로

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

이다. 여기서 N^2을 $N\cdot N$으로 나누어 두 인수에 하나씩 배분하였다. 이것이 구하려던 결과이다.

X의 분포

전확률 법칙에 의해 $k = 0, 1, 2, \ldots$에 대하여

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

이제 각 부분을 N으로 나타내면

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

이다. 인수 e^{-2}, N^{-2}, N^{-k}, $1/k!$은 n에 의존하지 않으므로 합의 밖으로 나온다.

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

이로써 원하는 결과를 얻는다.

$N = 3$이고 사고가 두 건인 경우

$\mathrm{P}(S = 2 \mid X = 2)$를 구하려 한다. 베이즈 정리에 의해

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

? **생각해 보기** $\mathrm{P}(X = 2)$를 끝까지 계산해야 할까?

그럴 필요 없다. 분자는 $\mathrm{P}(X = 2)$에 대한 합의 $n = 2$인 항일 뿐이다. 따라서 공통 인수 $\mathrm{e}^{-2}N^{-k-2}/k!$은 약분되고, 구하는 확률은 $\sum(2n-1)(2N-n)^k\mathrm{e}^{n/N}$의 $n = 2$인 항을 전체 합으로 나눈 값이다.

$N = 3, k = 2$일 때 항 $(2n-1)(6-n)^2\mathrm{e}^{n/3}$은

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

이다. 따라서 분자와 분모를 $\mathrm{e}^{2/3}$으로 나누면

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

이고, 이것이 구하려던 결과이다.

검산

- $N = 1$이면 평균이 1인 구역 하나뿐이고, 공식도 $2 - \frac{1}{6} \cdot 2 \cdot 3 = 1$을 준다. $N \to \infty$이면 평균은 $2 - \frac{4}{6} = \frac{4}{3}$으로 수렴한다. 이는 $\int_0^1 2x(2-x)\,\mathrm{d}x = \frac{4}{3}$과 일치하는데, 여기서 균일하게 고른 점이 상대 반지름 x에 있을 밀도(density)는 $2x$이다.
- $N = 3$일 때 평균은 $\frac{32}{27} \approx 1.185$이다. 4×10^5일에 대한 시뮬레이션(원판에서 균일하게 점을 고른 뒤 푸아송 건수를 추출)의 평균은 1.182였다. 또한 $\mathrm{P}(X = 2) \approx 0.209$(정확한 값 0.2094)를 얻었고, 사고가 두 건인 날 중 R_2에서 보낸 날의 비율은 0.372로 나왔다. 최종 답의 값은 0.3729이다.
- 분포 공식을 k에 대해 합하면, 시험해 본 모든 N에 대해 전체 확률이 1이 된다.

흔한 실수

- 구역을 넓이 $2n-1$에 비례하는 확률이 아니라 같은 확률 $1/N$으로 선택하는 것.
- 푸아송 분포의 **평균**들을 평균한 다음 X가 그 평균을 갖는 푸아송 분포라고 주장하는 것. 푸아송 분포들의 혼합은 푸아송 분포가 아니다. 이렇게 단순하게 평균을 거쳐 그대로 전달되는 것은 기댓값뿐이다.

STEP 너머

변수 X는 **혼합 분포**(*mixture distribution*)를 따른다. 구역은 은닉 변수(hidden variable) 역할을 하고, 마지막 문항은 이에 대한 베이즈 추론(Bayesian inference)을 수행한다. 위의 두 번째 실수는 전체 분산의 법칙(law of total variance) $\mathrm{Var}(X) = \mathrm{E}(\mathrm{Var}(X \mid S)) + \mathrm{Var}(\mathrm{E}(X \mid S))$으로 설명되는데, 이 법칙은 혼합 분포가 같은 평균을 갖는 단일 푸아송 분포보다 항상 더 넓게 퍼져 있음을 보여 준다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 조건부 기댓값(conditional expectation)과 이러한 혼합 분포를 잘 다룬다.

Beyond STEP 풀이: https://beyondstep.keithmcnulty.org/ko/. 풀이 글은 CC0 1.0이다. 문제 자체의 저작권은 © Cambridge University Press & Assessment에 있으며 여기에 싣지 않는다. 영어 원문을 기계 번역하였으며 아직 검토되지 않았다.