풀이

STEP III 1999, 문제 13

확률과 통계 • 연속 사전분포를 이용한 베이즈 추론

STEP III 1999, 13번 문제

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

step.maths.org/questions/99-s3-q13

핵심 아이디어

이 문제에는 무작위성이 두 겹으로 들어 있다. 먼저 내 몫의 크기 X가 정해지고, 그다음에 건포도가 어디에 놓이는지가 정해진다. $X = x$라는 **조건을 주면**(condition) 각 건포도는 다른 건포도와 독립적으로 확률 x로 내 몫에 들어가므로, 건포도의 개수는 이항분포(binomial distribution)를 따른다. 그다음 (i)은 전체 기댓값의 법칙(law of total expectation)이고, (ii)는 연속 사전분포(prior)를 갖는 베이즈 정리(Bayes' theorem)이다.

풀이

먼저 A를 구한다. 확률밀도함수(density)는 적분값이 1이어야 하므로

$$\int_0^1 Ax\,\mathrm{d}x = \frac{A}{2} = 1 \quad \Longrightarrow \quad A = 2,$$

이고, 따라서 X의 확률밀도함수는 $[0,1]$에서 $\mathrm{f}(x) = 2x$이다.

내 몫에 들어 있는 건포도의 개수를 N이라 하자. "무작위로 놓인다"는 말은 자연스러운 방식으로 모형화한다. 즉 각 건포도의 위치는 케이크 전체에 균등하게 분포하며, 다른 건포도의 위치와도, 내가 케이크를 얼마나 가져가는지와도 독립이다. 그러므로 내가 케이크의 비율 x만큼을 가져가면 각 건포도는 다른 건포도와 독립적으로 확률 x로 내 몫에 놓이고, 따라서

$$N \mid (X = x) \sim \mathrm{B}(4, x).$$

소문항 (i)

$X = x$가 주어지면 이항분포의 평균은 $\mathrm{E}(N \mid X = x) = 4x$이다. 이것을 X에 대하여 평균하면(전체 기댓값의 법칙)

$$\begin{aligned}\mathrm{E}(N) &= \int_0^1 \mathrm{E}(N \mid X = x)\, \mathrm{f}(x)\, \mathrm{d}x = \int_0^1 4x \cdot 2x \, \mathrm{d}x \\ &= 8 \int_0^1 x^2 \, \mathrm{d}x = \frac{8}{3}.\end{aligned}$$

같은 말이지만, $\mathrm{E}(X) = \frac{2}{3}$이므로 $\mathrm{E}(N) = 4\,\mathrm{E}(X)$이다. 내 몫에 들어 있는 건포도 개수의 기댓값은 $\frac{8}{3}$이다.

소문항 (ii)

구하려는 것은 $\mathrm{P}(X > \frac{1}{2} \mid N = 4)$이다. $X = x$가 주어지면 건포도 네 개가 모두 내 몫에 있을 확률은 x^4이다.

> **생각해 보기** X가 연속이어서 $\mathrm{P}(X = x) = 0$일 때 베이즈 정리를 어떻게 쓸 수 있을까?

베이즈 정리의 합을 확률밀도함수에 대한 적분으로 바꾼다. 가능한 몫의 크기 가운데 작은 구간 $[x, x + \mathrm{d}x]$의 확률은 $2x\,\mathrm{d}x$이고, 그 구간이 주어졌을 때 건포도 네 개가 모두 들어 있을 확률은 x^4이다. 모든 구간에 대하여 더하면

$$\mathrm{P}(N = 4) = \int_0^1 x^4 \cdot 2x \, \mathrm{d}x = 2 \int_0^1 x^5 \, \mathrm{d}x = \frac{2}{6} = \frac{1}{3}.$$

같은 방법으로 $x > \frac{1}{2}$인 구간만 남기면

$$\mathrm{P}(X > \tfrac{1}{2} \text{ 이고} N = 4) = \int_{1/2}^1 2x^5 \, \mathrm{d}x = \frac{1}{3}\left(1 - \frac{1}{64}\right) = \frac{1}{3} \cdot \frac{63}{64}.$$

따라서

$$\mathrm{P}(X > \tfrac{1}{2} \mid N = 4) = \frac{\frac{1}{3} \cdot \frac{63}{64}}{\frac{1}{3}} = \frac{63}{64}.$$

다른 방법으로 볼 수도 있다. $N = 4$가 주어졌을 때 X의 조건부 확률밀도함수는 $x^4 \cdot 2x$에 비례하므로 $[0, 1]$에서 $6x^5$이다. 그 분포함수(distribution function)는 x^6이고, $1 - (\frac{1}{2})^6 = \frac{63}{64}$이다.

검산

- (ii)의 답은 1에 아주 가까운데, 이는 납득할 만하다. 건포도를 모두 찾았다는 것은 내가 큰 조각을 가져갔다는 강력한 증거이기 때문이다. 사전확률 $\mathrm{P}(X > \frac{1}{2}) = 1 - \frac{1}{4} = \frac{3}{4}$가 $\frac{63}{64} \approx 0.984$로 올라갔다.

- $\frac{8}{3}$은 0과 4 사이에 있고, 마땅히 그래야 하듯 2보다 크다. 확률밀도함수 $2x$는 큰 몫에 더 큰 비중을 주기 때문이다.
- 10^6번 시행한 몬테카를로 시뮬레이션(Monte Carlo simulation)(U가 균등분포를 따를 때 확률밀도함수가 $2x$인 $X = \sqrt{U}$를 택한 뒤 건포도 네 개를 놓는다)에서 $\mathrm{E}(N) \approx 2.667$, $\mathrm{P}(N = 4) \approx 0.333$, 조건부 확률 ≈ 0.9846을 얻었는데, 이는 $\frac{63}{64} = 0.984375$와 잘 맞는다.

흔한 실수

- A를 구하는 것을 잊거나 $A = 1$로 두는 실수. 그러면 f는 확률밀도함수가 아니게 되고, (i)의 기댓값이 틀리게 나온다(다만 (ii)에서는 A가 약분된다).
- $\mathrm{P}(X > \frac{1}{2} \mid N = 4)$ 대신 $\mathrm{P}(N = 4 \mid X > \frac{1}{2})$를 계산하는 실수. 조건이 반대 방향이다.
- 적분 대신 $(\frac{1}{2})^4$를 쓰는 실수. 몫의 크기가 무작위이므로 값 하나를 대입할 것이 아니라 x^4를 확률밀도함수에 대하여 평균해야 한다.

STEP 너머

이것은 베이즈 추론(Bayesian inference)의 축소판이다. 확률밀도함수 $2x$는 알려지지 않은 비율에 대한 **사전분포**이고, 이항분포는 **가능도**(likelihood)이며, $6x^5$는 **사후분포**(posterior)이다. 사전분포와 사후분포는 모두 베타분포(beta distribution)인데, 이는 이항 자료에 대한 켤레 사전분포(conjugate prior)의 한 예이며, 같은 계산이 라플라스(Laplace)의 계승 규칙(rule of succession)의 바탕을 이룬다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 베타-이항 이야기와 연속 확률변수에 대한 조건부 확률을 매우 명료하게 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.