풀이

STEP III 1992, 문제 16

확률과 통계 • 푸아송 솎아내기

STEP III 1992, 16번 문제

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

step.maths.org/questions/92-s3-q16

핵심 아이디어

오탈자(misprint)의 총개수가 정해지면, 각 오탈자는 내가 찾아내거나 놓치는 하나의 독립 시행(independent trial)이 된다. 따라서 소문항 (i)은 이항 확률(binomial probability)이다. 소문항 (ii)는 조건부 확률(conditional probability)인데, 분모인 "내가 정확히 m개의 오탈자를 찾아낼 확률"을 구하려면 놓친 오탈자의 개수로 가능한 모든 경우에 대해 더해야 한다. 이 합은 지수급수(exponential series)가 되고, 바로 이 덕분에 모든 것이 깔끔한 푸아송(Poisson) 꼴로 정리된다.

풀이

이 풀이 전체에서 $q = 1 - p$로 쓴다. 신문 한 호에 있는 오탈자의 총개수를 N, 내가 찾아낸 개수를 S, 내가 찾지 못한 개수를 $U = N - S$라 하자. 문제에서 다음이 주어졌다.

$$\mathrm{P}(N = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

소문항 (i)

$N = m + n$이 주어지면, $m + n$개의 오탈자 각각은 다른 것들과 독립적으로 확률 p로 발견된다. 그러므로 $N = m + n$이라는 조건 아래에서 발견된 개수는 이항분포(binomial distribution) $\mathrm{B}(m + n, p)$를 따른다. 발견된 오탈자 m개와 놓친 오탈자 n개를 특정하게 고르는 각 경우의 확률은 $p^m q^n$이고, 이런 선택은 $\binom{m+n}{m}$가지이므로

$$\mathrm{P}(S = m \mid N = m + n) = \binom{m+n}{m} p^m (1-p)^n.$$

소문항 (ii)

구하려는 것은 $\mathrm{P}(U = n \mid S = m)$이다. 사건 $\{S = m,\ U = n\}$은 사건 $\{N = m+n,\ S = m\}$과 같으므로, 소문항 (i)에 의해

$$\begin{aligned}\mathrm{P}(S = m,\ U = n) &= \mathrm{P}(N = m+n)\,\mathrm{P}(S = m \mid N = m+n)\\ &= \frac{\mathrm{e}^{-\lambda}\lambda^{m+n}}{(m+n)!}\cdot\frac{(m+n)!}{m!\,n!}\,p^m q^n\\ &= \mathrm{e}^{-\lambda}\,\frac{(\lambda p)^m}{m!}\cdot\frac{(\lambda q)^n}{n!}.\end{aligned}$$

인수 $(m+n)!$이 약분되는데, 이것이 이 문제가 풀리는 핵심이다.

생각해 보기 오탈자가 몇 개였는지 모를 때 $\mathrm{P}(S = m)$은 어떻게 구할까?

내가 정확히 m개를 찾아냈다면, 놓친 개수는 $0, 1, 2, \ldots$ 중 어느 것이든 될 수 있고, 이 경우들은 서로 배반(mutually exclusive)이다. 따라서 결합확률(joint probability)을 U의 모든 값 j에 대해 더한다.

$$\begin{aligned}\mathrm{P}(S = m) &= \sum_{j=0}^{\infty}\mathrm{e}^{-\lambda}\,\frac{(\lambda p)^m}{m!}\cdot\frac{(\lambda q)^j}{j!} = \mathrm{e}^{-\lambda}\,\frac{(\lambda p)^m}{m!}\,\mathrm{e}^{\lambda q}\\ &= \frac{(\lambda p)^m}{m!}\,\mathrm{e}^{-\lambda p},\end{aligned}$$

여기서 $\sum_{j\ge 0} x^j/j! = \mathrm{e}^x$와 $-\lambda + \lambda q = -\lambda p$를 사용했다. (즉 내가 찾아낸 오탈자의 개수 자체도 평균이 λp인 푸아송 분포를 따른다.) 이 값은 양수이므로 이것으로 나눌 수 있다. 따라서

$$\begin{aligned}\mathrm{P}(U = n \mid S = m) &= \frac{\mathrm{P}(S = m,\ U = n)}{\mathrm{P}(S = m)} = \frac{\mathrm{e}^{-\lambda}\,\dfrac{(\lambda p)^m}{m!}\cdot\dfrac{(\lambda q)^n}{n!}}{\mathrm{e}^{-\lambda p}\,\dfrac{(\lambda p)^m}{m!}}\\ &= \frac{(\lambda q)^n}{n!}\,\mathrm{e}^{-\lambda+\lambda p} = \frac{(1-p)^n\lambda^n}{n!}\,\mathrm{e}^{-(1-p)\lambda},\end{aligned}$$

이 되어 원하는 결과를 얻는다.

답은 m에 전혀 의존하지 않는다. 즉 내가 오탈자를 몇 개 찾아냈는지 알아도, 몇 개를 놓쳤는지에 대해서는 아무것도 알 수 없다. m이 무엇이든 놓친 개수는 평균이 $(1-p)\lambda$인 푸아송 분포를 따른다.

검산

- 최종 답을 $n = 0, 1, 2, \ldots$에 대해 모두 더하면 $\mathrm{e}^{(1-p)\lambda}\mathrm{e}^{-(1-p)\lambda} = 1$이 되어, 조건부 분포가 만족해야 할 성질을 만족한다.

- $p \to 1$이면 나는 모든 오탈자를 찾아내고, 답은 $n = 0$일 때 1로, $n \geq 1$일 때 0으로 수렴한다. 맞다.
- $\lambda = 3$, $p = 0.6$, $m = 2$, $n = 1$일 때 공식은 $1.2\,\mathrm{e}^{-1.2} \approx 0.3614$를 준다. 신문 400 000부를 시뮬레이션하여 정확히 두 개의 오탈자가 발견된 경우만 남겼더니 0.360이 나왔다.

흔한 실수

- 소문항 (i)에서 이항계수 $\binom{m+n}{m}$를 빠뜨리는 것: 발견된 m개의 오탈자는 $m+n$개 중 어느 m개라도 될 수 있다.
- 놓친 오탈자의 개수로 가능한 모든 경우에 대해 더하지 않고, $\mathrm{P}(S = m)$을 $\mathrm{P}(N = m)$으로 보거나 소문항 (i)의 답으로 보는 것.

STEP 너머

이것은 푸아송 분포의 **솎아내기**(thinning, 또는 분할(splitting)) 성질이다. 개수가 푸아송(λ) 분포를 따르는 대상들을 각각 확률 p로 남긴다면, 남은 개수와 버려진 개수는 평균이 각각 λp와 $\lambda(1-p)$인 **독립인** 푸아송 확률변수이다. 우리가 구한 결합확률은 정확히 이 두 조각으로 인수분해된다. 솎아내기는 푸아송 과정(Poisson process) 이론의 핵심 도구로, 하나의 무작위 사건의 흐름을 서로 독립인 하위 흐름들로 나눌 수 있게 해 준다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)에서 이 내용을 다루며, 이를 한 줄로 증명하는 생성함수(generating function) 방법도 함께 소개한다.

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.