풀이

STEP I 1995, 문제 13

확률과 통계 • 푸아송 솎아내기

> **STEP I 1995, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/95-s1-q13

핵심 아이디어

소문항 (i)은 푸아송 분포(Poisson distribution)의 **솎아내기**(*thinning*) 성질이다. 실제 암세포의 수를 조건으로 두면, 표시된 세포의 수는 이항분포(binomial distribution)를 따른다. 알 수 없는 실제 개수에 대해 합하면 지수급수(exponential series)로 깔끔하게 정리된다. 소문항 (ii)에서 "두 번째 표시가 슬라이드 k에 있다"는 사건은 처음 $k-1$개의 슬라이드에 있는 표시의 총 개수(0개 또는 1개)에 의해 결정되며, 소문항 (i)과 독립인 푸아송 확률변수의 가법성(additivity)을 함께 쓰면 그 총합이 정확히 어떤 분포를 따르는지 알 수 있다.

풀이

이 풀이 전체에서 $q = 1-p$, $\lambda = \mu p$로 둔다. 서로 다른 슬라이드의 암세포 수는 독립(independent)이라고 가정하고("무작위"라는 말의 자연스러운 해석이다), 과학자의 발견들은 서로 독립이며 세포 수와도 독립이라고 가정한다.

소문항 (i)

한 슬라이드에서 암세포의 수를 N, 과학자가 표시하는 세포의 수를 M이라 하면 $N \sim \mathrm{Po}(\mu)$이다. $N = n$일 때 n개의 세포 각각은 독립적으로 확률 p로 발견되므로 $M \mid N = n \sim \mathrm{B}(n, p)$이다. $m = 0, 1, 2, \ldots$에 대하여 전확률 법칙(law of total probability)에 의해 ($n \geq m$인 항만 기여한다)

$$\begin{aligned} \mathrm{P}(M = m) &= \sum_{n=m}^{\infty} \mathrm{e}^{-\mu} \frac{\mu^n}{n!} \binom{n}{m} p^m q^{n-m} = \mathrm{e}^{-\mu} \frac{(\mu p)^m}{m!} \sum_{n=m}^{\infty} \frac{(\mu q)^{n-m}}{(n-m)!} \\ &= \mathrm{e}^{-\mu} \frac{(\mu p)^m}{m!} \mathrm{e}^{\mu q} = \mathrm{e}^{-\mu p} \frac{(\mu p)^m}{m!}, \end{aligned}$$

이다. 여기서 $\binom{n}{m}/n! = 1/(m!\,(n-m)!)$을 쓰고, 이어서 $j = n - m$으로 두어 지수급수를 쓰고, 마지막으로 $-\mu + \mu q = -\mu p$를 썼다. 따라서 요구한 대로 $M \sim \mathrm{Po}(\mu p)$이다.

소문항 (ii)

슬라이드 i에서 표시된 세포의 수를 M_i라 하자. 소문항 (i)과 슬라이드 사이의 독립성에 의해 $M_1, M_2, \ldots$는 독립인 $\mathrm{Po}(\lambda)$ 확률변수이다. 처음 $k-1$개의 슬라이드에서 표시된 수를 $S = M_1 + \cdots + M_{k-1}$이라 하자. 독립인 푸아송 확률변수들의 합은 평균을 더한 푸아송 분포를 따르므로 $S \sim \mathrm{Po}((k-1)\lambda)$이고, S는 M_k와 독립이다.

> **생각해 보기** 두 번째 표시가 슬라이드 k에 오려면 정확히 무슨 일이 일어나야 하는가?

슬라이드 k 이전에는 표시가 많아야 하나여야 하고, 슬라이드 k까지 마쳤을 때는 적어도 둘이어야 한다. 따라서 서로 배반인 두 경우가 있다.



$S = 0$이고 $M_k \geq 2$, 또는 $S = 1$이고 $M_k \geq 1$

($k = 1$이면 앞선 슬라이드가 없으므로 자동으로 $S = 0$이고, 아래 식들은 여전히 성립한다.) S와 M_k의 독립성을 쓰면

$$\mathrm{P}(S = 0,\ M_k \geq 2) = \mathrm{e}^{-(k-1)\lambda}\left(1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda}\right),$$
$$\mathrm{P}(S = 1,\ M_k \geq 1) = (k-1)\lambda\, \mathrm{e}^{-(k-1)\lambda}\left(1 - \mathrm{e}^{-\lambda}\right).$$

이다. 둘을 더하면

$$Q = \mathrm{e}^{-(k-1)\lambda}\left\{\left(1 + (k-1)\lambda\right)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda \mathrm{e}^{-\lambda}\right\}.$$

이다. 요구된 꼴에 맞추기 위해 $1 + (k-1)\lambda = (1 + k\lambda) - \lambda$로 쓰면

$$\begin{aligned}\left(1 + (k-1)\lambda\right)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda \mathrm{e}^{-\lambda} &= (1 + k\lambda)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda + \lambda \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} \\ &= (1 + k\lambda)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda.\end{aligned}$$

이다. $\lambda = \mu p$를 대입하면

$$Q = \mathrm{e}^{-\mu p(k-1)}\left\{(1 + k\mu p)\left(1 - \mathrm{e}^{-\mu p}\right) - \mu p\right\},$$

이고, 이것이 주어진 결과이다.

검산

- $k = 1$이면 식은 $1 - \mathrm{e}^{-\lambda} - \lambda\mathrm{e}^{-\lambda} = \mathrm{P}(M_1 \geq 2)$가 되어, 마땅히 그래야 하는 값이 나온다.
- $k = 1, 2, \ldots$에 대해 확률을 모두 더하면 1이 되어야 한다(과학자는 결국 확률 1로 두 번째 세포를 표시한다). 여러 λ 값에 대해 수치적으로 더해 보면 이것이 확인된다.
- $\mu = 1.6$, $p = 0.45$(따라서 $\lambda = 0.72$)로 시뮬레이션(simulation)을 하면 200 000번의 시행 중 약 26.0%에서 두 번째 표시가 슬라이드 2에 나왔다. 식으로 구하면 $Q \approx 0.2591$이다.

흔한 실수

- "슬라이드 k 이전에 표시 하나, 슬라이드 k에 적어도 하나"인 경우를 빠뜨리거나, 실수로 $S = 1$, $M_k = 0$인 경우를 포함하는 것.
- "표시가 있는 첫 번째 슬라이드"와 같은 논법으로 슬라이드를 하나씩 따져 나가는 것. 이렇게 하면 복잡해진다. 처음 $k - 1$개 슬라이드에 걸친 총합 S를 다루는 것이 효율적이다.
- 독립성(슬라이드 사이의 독립성, 그리고 S와 M_k 사이의 독립성)을 어디서 쓰는지 밝히지 않는 것.

STEP 너머

솎아내기는 푸아송 과정(Poisson process)의 한 측면이다. 사건들이 푸아송 과정으로 일어나고 각 사건을 독립적으로 확률 p로 남기면, 남은 사건들은 다시 비율이 p배가 된 푸아송 과정을 이룬다. 그러면 소문항 (ii)는 두 번째 사건까지의 대기 시간(waiting time)의 이산적인 버전이며, 연속 시간에서는 이 대기 시간이 감마 분포(gamma distribution)를 따른다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 푸아송 과정과 이러한 대기 시간 분포를 전개한다.

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.