풀이

STEP III 2019, 문제 11

확률과 통계 • 푸아송 솎아내기

STEP III 2019, 11번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2019년 STEP 시험지 묶음(zip 파일) 가운데 STEP III 시험지에 있다:

www.ocr.org.uk/Images/696590-2019-past-papers.zip

핵심 아이디어

모든 것은 도착한 손님의 수, 또는 모래를 가져간 손님의 수에 대해 조건을 걸고(conditioning) 푸아송 급수를 더하는 데서 출발한다. 소문항 (i)은 푸아송 분포(Poisson distribution)의 고전적인 "솎아내기(thinning)"이다. 이항 확률과 푸아송 확률을 곱하면, 관찰되지 않은 전체 수에 대한 합이 하나의 지수함수로 정리된다. 모래를 가져가는 손님의 수가 푸아송 분포를 따른다는 것을 알고 나면, 소문항 (ii)와 (iii)에서는 $(1-k)^Y$의 기댓값(expectation)만 구하면 되고, 이것 역시 지수 급수 하나로 계산된다.

풀이

하루 동안 오는 손님의 수를 N이라 하면 $N \sim \mathrm{Po}(\lambda)$이고, 그중 모래를 가져가는 손님의 수를 Y라 하자. 손님들은 서로 독립적으로, 그리고 N과도 독립적으로 결정한다고 가정한다. 따라서 $N = n$이 주어지면 $Y \sim \mathrm{B}(n, p)$이다.

소문항 (i)

$y = 0, 1, 2, \ldots$에 대하여 N에 조건을 걸면($n \geq y$인 항만 기여한다) 다음을 얻는다.

$$\begin{aligned}
\mathrm{P}(Y = y) &= \sum_{n=y}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!} \binom{n}{y} p^y (1-p)^{n-y} \\
&= \mathrm{e}^{-\lambda} \frac{(p\lambda)^y}{y!} \sum_{n=y}^{\infty} \frac{(\lambda(1-p))^{n-y}}{(n-y)!} \qquad \text{왜냐하면} \frac{\lambda^n}{n!} \cdot \frac{n!}{y!\,(n-y)!} = \frac{\lambda^y \lambda^{n-y}}{y!\,(n-y)!} \\
&= \mathrm{e}^{-\lambda} \frac{(p\lambda)^y}{y!} \mathrm{e}^{\lambda(1-p)} = \mathrm{e}^{-p\lambda} \frac{(p\lambda)^y}{y!},
\end{aligned}$$

여기서 $m = n - y$로 놓고 $\sum_{m\geq 0} x^m/m! = \mathrm{e}^x$를 이용하였다. 이것은 $\mathrm{Po}(p\lambda)$의 확률이므로, 요구한 대로 $Y \sim \mathrm{Po}(p\lambda)$이다.

소문항 (ii)

모래를 가져가는 손님은 각각 그때 남아 있던 양의 $1-k$만큼을 남긴다. 따라서 모래를 가져가는 손님 Y명이 지나간 뒤 남은 질량은 $S(1-k)^Y$이고, 가져간 질량은 다음과 같다.

$$T = S(1-(1-k)^Y).$$

소문항 (i)에 의하여

$$\mathrm{E}[(1-k)^Y] = \sum_{y=0}^{\infty} (1-k)^y \,\mathrm{e}^{-p\lambda}\frac{(p\lambda)^y}{y!} = \mathrm{e}^{-p\lambda}\,\mathrm{e}^{p\lambda(1-k)} = \mathrm{e}^{-kp\lambda}.$$

그러므로 기댓값의 선형성(linearity of expectation)에 의하여

$$\mathrm{E}(T) = S(1-\mathrm{E}[(1-k)^Y]) = (1-\mathrm{e}^{-kp\lambda})S,$$

이고, 이것이 구하는 결과이다.

소문항 (iii)

> **생각해 보기** "모래의 k만큼의 비율"을 특정한 모래알 하나에 대한 확률로 어떻게 바꿀 수 있을까?

모래알은 매우 많고 어느 하나도 특별히 구별되지 않으므로, 누군가 남은 모래의 k만큼을 가져갈 때마다 아직 자루에 남아 있는 모래알은 이전에 일어난 일과 독립적으로 확률 k로 선택된다. (같은 말로, 황금 모래알이 어떤 몫 안에 들어 있을 확률은 그 몫이 전체 질량에서 차지하는 비율과 같다.)

점원이 황금 모래알을 가져가는 사건을 G라 하자. $Y = y$가 주어지면, 모래알이 y명의 손님을 모두 견디고 남을 확률은 $(1-k)^y$이고, 그다음 점원이 그것을 가져갈 확률은 k이다. 따라서

$$\mathrm{P}(G \mid Y = y) = k(1-k)^y,$$

이고, 소문항 (ii)에서 구한 기댓값을 이용하면

$$\mathrm{P}(G) = k\,\mathrm{E}[(1-k)^Y] = k\,\mathrm{e}^{-kp\lambda}.$$

이것은 소문항 (ii)와 맞아떨어진다. 황금 모래알이 하루를 견디고 남을 확률은 $\mathrm{e}^{-kp\lambda}$이고, 이는 정확히 남아 있는 모래 비율의 기댓값이다.

$k = 0$인 경우. 이때 $\mathrm{P}(G) = 0$이다. 이는 타당하다. $k = 0$이면 점원을 포함하여 아무도 모래를 가져가지 않으므로 모래알은 자루에 그대로 남는다.

$k \to 1$일 때의 극한. 이때 $\mathrm{P}(G) \to \mathrm{e}^{-p\lambda}$이고, 이는 $\mathrm{P}(Y=0)$이다. k가 1에 가까우면 모래를 가져가는 첫 손님이 거의 모든 모래를 가져가고, 황금 모래알도 거의 확실히 함께 가져간다. 점원은 사실상 어떤 손님도 모래를 가져가지 않은 경우에만 모래알을 얻으며, 그때 (거의) 자루 전체를 가져간다.

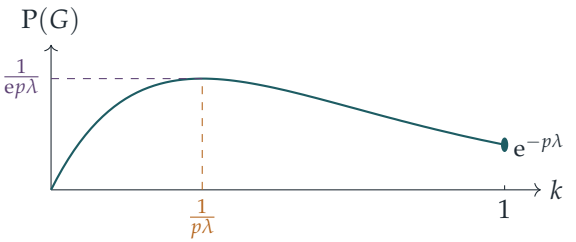
최댓값 구하기. $0 \le k < 1$에 대하여 $\mu = p\lambda$, $f(k) = k\,\mathrm{e}^{-\mu k}$라 하자. 그러면

$$f'(k) = \mathrm{e}^{-\mu k}(1-\mu k),$$

이고, 이는 $k < 1/\mu$이면 양수, $k > 1/\mu$이면 음수이다. $\mu > 1$이므로 점 $k = 1/\mu$는 $(0,1)$ 안에 있다. 따라서 f는 이 점까지 증가하다가 그 뒤로 감소하므로, 이 점에서 $[0,1)$ 위의 최댓값을 갖는다. 그러므로 확률은

$$k = \frac{1}{p\lambda},$$

일 때 최대이고, 최대 확률은 $\dfrac{1}{\mathrm{e}\,p\lambda}$이다. ($p\lambda \le 1$이면 f는 $[0,1)$ 전체에서 증가하므로 최댓값이 없고, $k \to 1$일 때의 상한(supremum) $\mathrm{e}^{-p\lambda}$만 있다. 문제에서 $p\lambda > 1$로 제한한 이유가 이것이다.)



그래프는 $p\lambda = 3$인 경우의 $\mathrm{P}(G) = k\,\mathrm{e}^{-kp\lambda}$를 나타낸다.

검산

- 소문항 (i)은 당연한 시험을 통과한다. $p = 1$이면 $\mathrm{Po}(\lambda)$가 다시 나오고, $p = 0$이면 확실히 $Y = 0$이다.
- 소문항 (ii)에서 $k = 0$이면 가져간 모래가 0이고, $\lambda \to \infty$이면 가져간 질량의 기댓값은 마땅히 그래야 하듯 S로 수렴한다.
- $\lambda = 3$, $p = 0.6$, $k = 0.4$이면 공식은 $\mathrm{P}(G) = 0.4\,\mathrm{e}^{-0.72} \approx 0.1947$을 준다. 200 000일에 대한 컴퓨터 모의실험에서는 0.1953이 나왔다.
- $p\lambda = 3$일 때, 촘촘한 격자에서 $k\,\mathrm{e}^{-3k}$의 가장 큰 값은 $k \approx 0.3333$에서 나타나며 그 값은 $0.1226 \approx 1/(3\mathrm{e})$이다.

흔한 실수

- 각 손님이 **처음** 모래의 k만큼을 가져가는 것처럼 가져간 모래를 kSY로 쓰는 것. 각 손님은 **남은** 모래의 일부를 가져가므로, 가져가는 양은 등비수열(geometric sequence)을 이룬다.
- $\mathrm{E}[(1-k)^Y]$를 $(1-k)^{\mathrm{E}(Y)} = (1-k)^{p\lambda}$로 계산하는 것. 비선형 함수의 기댓값은 기댓값의 함숫값이 아니다.
- $1/(p\lambda)$가 실제로 허용 범위 $0 \leq k < 1$ 안에 있는지 확인하지 않는 것. 바로 이 부분에서 조건 $p\lambda > 1$이 쓰인다.

STEP 너머

소문항 (i)은 푸아송 분포의 솎아내기 성질이다. 소문항 (ii)와 (iii)에 나타나는 기댓값 $\mathrm{E}[s^Y] = \mathrm{e}^{\mu(s-1)}$는 $\mathrm{Po}(\mu)$의 **확률생성함수**(*probability generating function*)를 $s = 1-k$에서 계산한 값이다. 생성함수를 쓰면 솎아내기는 거의 한 줄로 끝난다. $\mathrm{E}[s^N] = \mathrm{e}^{\lambda(s-1)}$이면 $\mathrm{E}[s^Y] = \mathrm{E}[(1-p+ps)^N] = \mathrm{e}^{\lambda p(s-1)}$이다. 푸아송 과정(Poisson process)의 이론에서 이것은 푸아송 과정의 점들에 독립적으로 표시를 하면 다시 푸아송 과정들을 얻는다는 명제가 된다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 생성함수와 이러한 아이디어들을 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.