풀이

STEP I 2016, 문제 13

확률과 통계 • 지수분포의 순서통계량

> **STEP I 2016, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/16-s1-q13

핵심 아이디어

첫 번째와 두 번째 도착에 대해서는, 그 일이 **아직 일어나지 않은** 때가 언제인지를 묻는다. 첫 번째 이메일이 시각 t 이후에 도착하는 것은 정확히 n개 모두가 그 이후에 도착할 때이고, 두 번째 이메일이 t 이후에 도착하는 것은 정확히 그때까지 도착한 이메일이 많아야 하나일 때이다. 두 사건 모두 독립성(independence)을 이용하면 쉽게 셀 수 있고, 생존 확률(survival probability)을 미분하면 밀도함수(density)를 얻는다.

풀이

$X_1, \ldots, X_n$을 도착 시각이라 하자. 이들은 서로 독립이고, 각각

$$\mathrm{P}(X_i > t) = \int_t^\infty \lambda \mathrm{e}^{-\lambda s}\,\mathrm{d}s = \mathrm{e}^{-\lambda t}, \qquad \mathrm{P}(X_i \le t) = 1 - \mathrm{e}^{-\lambda t}$$

을 만족한다(단, $t \ge 0$). $k > 0$일 때의 표준 적분 $\int_0^\infty t\,k\mathrm{e}^{-kt}\,\mathrm{d}t = \frac{1}{k}$를 사용할 것인데, 이는 모수(parameter)가 k인 지수분포(exponential distribution)의 평균이다. (부분적분(integration by parts)을 하면 $[-t\mathrm{e}^{-kt}]_0^\infty + \int_0^\infty \mathrm{e}^{-kt}\,\mathrm{d}t = 0 + \frac{1}{k}$이다.)

소문항 (i)

첫 번째 도착이 t보다 늦게 일어날 필요충분조건은 모든 이메일이 t보다 늦게 도착하는 것이다. 독립성에 의해

$$\mathrm{P}(T > t) = \prod_{i=1}^{n} \mathrm{P}(X_i > t) = \left(\mathrm{e}^{-\lambda t}\right)^n = \mathrm{e}^{-n\lambda t}.$$

따라서 누적분포함수(cumulative distribution function)는 $t \geq 0$에 대하여 $F(t) = 1 - \mathrm{e}^{-n\lambda t}$이다. 이를 미분하면 밀도함수

$$f(t) = F'(t) = n\lambda\,\mathrm{e}^{-n\lambda t} \qquad (t \geq 0),$$

를 얻으며, 이것이 구하는 결과이다. 이는 모수가 $n\lambda$인 지수분포의 밀도함수이므로, $k = n\lambda$로 두고 표준 적분을 쓰면

$$\mathrm{E}(T) = \frac{1}{n\lambda}.$$

소문항 (ii)

S를 두 번째 이메일의 도착 시각이라 하자(따라서 $n \geq 2$이어야 한다). 두 번째 도착이 t보다 늦게 일어날 필요충분조건은 시각 t까지 도착한 이메일이 **많아야 하나**인 것이다. 즉 하나도 도착하지 않았거나, 정확히 하나만 도착한 경우이다.

생각해 보기 시각 t까지 정확히 하나의 이메일만 도착하는 경우는 몇 가지인가?

n개의 이메일 중 어느 것이든 일찍 도착한 하나가 될 수 있다. 그 하나는 도착했고(확률 $1 - \mathrm{e}^{-\lambda t}$), 나머지 $n-1$개는 도착하지 않았다(각각 확률 $\mathrm{e}^{-\lambda t}$). 독립성에 의해, 그리고 n가지 경우가 서로 배반(mutually exclusive)이므로

$$\mathrm{P}(S > t) = \mathrm{e}^{-n\lambda t} + n\left(1 - \mathrm{e}^{-\lambda t}\right)\mathrm{e}^{-(n-1)\lambda t}.$$

두 번째 항을 전개하면 $n\mathrm{e}^{-(n-1)\lambda t} - n\mathrm{e}^{-n\lambda t}$이므로

$$\mathrm{P}(S > t) = n\,\mathrm{e}^{-(n-1)\lambda t} - (n-1)\,\mathrm{e}^{-n\lambda t}.$$

간단히 확인해 보면, 이 값은 $t = 0$에서 1이고 $t \to \infty$일 때 0으로 수렴한다.

S의 밀도함수 g는 $1 - \mathrm{P}(S > t)$의 도함수이다.

$$\begin{aligned} g(t) &= -\frac{\mathrm{d}}{\mathrm{d}t}\left(n\,\mathrm{e}^{-(n-1)\lambda t} - (n-1)\,\mathrm{e}^{-n\lambda t}\right) \\ &= n(n-1)\lambda\,\mathrm{e}^{-(n-1)\lambda t} - n(n-1)\lambda\,\mathrm{e}^{-n\lambda t} \\ &= n(n-1)\lambda\left(\mathrm{e}^{-(n-1)\lambda t} - \mathrm{e}^{-n\lambda t}\right) \qquad (t \geq 0). \end{aligned}$$

기댓값(expectation)을 구하기 위해 g를 지수분포 밀도함수들의 결합으로 나타낸다.

$$g(t) = n \cdot (n-1)\lambda\,\mathrm{e}^{-(n-1)\lambda t} \;-\; (n-1) \cdot n\lambda\,\mathrm{e}^{-n\lambda t}.$$

$k = (n-1)\lambda$와 $k = n\lambda$로 두고 표준 적분을 쓰면($n \geq 2$이므로 두 경우 모두 $k > 0$이어서 사용할 수 있다)

$$\begin{aligned} \mathrm{E}(S) = \int_0^\infty t\,g(t)\,\mathrm{d}t &= \frac{n}{(n-1)\lambda} - \frac{n-1}{n\lambda} \\ &= \frac{1}{\lambda} \cdot \frac{n^2 - (n-1)^2}{n(n-1)} = \frac{1}{\lambda} \cdot \frac{2n-1}{n(n-1)}. \end{aligned}$$

$\dfrac{1}{n-1}+\dfrac{1}{n}=\dfrac{n+(n-1)}{n(n-1)}=\dfrac{2n-1}{n(n-1)}$이므로, 이는

$$\mathrm{E}(S)=\frac{1}{\lambda}\left(\frac{1}{n-1}+\frac{1}{n}\right),$$

이고, 이것이 구하는 결과이다.

? 생각해 보기 답이 왜 간단한 두 항의 합으로 나오는가?

지수분포는 무기억성(memoryless property)을 가진다. 첫 번째 도착은 비율이 $n\lambda$인 지수분포를 따르는 시간 뒤에 일어나며, 그 평균은 $\frac{1}{n\lambda}$이다. 그 뒤 남은 $n-1$개의 이메일은 마치 시계가 방금 초기화된 것처럼 새로 출발하므로, 추가로 기다리는 시간은 비율이 $(n-1)\lambda$이고 평균이 $\frac{1}{(n-1)\lambda}$인 지수분포를 따른다. 두 대기 시간을 더하면 곧바로 답이 나오며, 이는 결과를 기억하는 좋은 방법이기도 하다.

검산

- $n=2, \lambda=1$일 때 두 번째 도착은 독립인 두 지수분포 시간 중 늦은 쪽이다. 그 평균은 $1+\frac{1}{2}=1.5$이어야 하며, 10^5회 시행한 컴퓨터 시뮬레이션에서는 1.501이 나왔다.
- 밀도함수 g를 적분하면 1이다: $\int_0^\infty g = n-(n-1)=1$.
- $n=5, \lambda=1$일 때 공식은 $\mathrm{E}(S)=\frac{1}{4}+\frac{1}{5}=0.45$를 주며, 이는 $\int_0^\infty t\,g(t)\,\mathrm{d}t$를 수치적으로 계산한 값과 일치한다.

흔한 실수

- 정확히 하나의 이메일만 도착한 경우에 인수 n을 빠뜨리는 것. n개 중 어느 것이든 일찍 도착한 하나가 될 수 있다.
- $\mathrm{P}(S>t)$를 **정확히** 하나가 도착했을 확률로 써서, 하나도 도착하지 않은 경우를 빠뜨리는 것.

STEP 너머

T와 S는 지수분포 표본의 처음 두 **순서통계량**(*order statistics*)이다. 위의 무기억성 논증을 확장하면 k번째 도착의 평균이 $\frac{1}{\lambda}\left(\frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{n-k+1}\right)$임을 보일 수 있으며, 이는 포아송 과정(Poisson process)과 대기행렬 이론(queueing theory)의 기초가 된다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 순서통계량과 무기억성을 명쾌하게 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.