풀이

STEP III 2003, 문제 12

확률과 통계 • 지수분포를 따르는 대기 시간

STEP III 2003, 12번 문제

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

step.maths.org/questions/03-s3-q12

핵심 아이디어

임의의 구간에서 방해 횟수가 포아송 분포(Poisson distribution)를 따른다면, "다음 방해가 t분보다 더 뒤에 일어난다"는 사건은 "다음 t분 동안 방해가 **전혀** 없다"는 사건과 같다. 이 사실이 이산 분포를 연속 분포로 바꾸어 주며, 대기 시간은 지수분포(exponential distribution)를 따른다. 그다음에는 연속한 방해 사이의 간격들이 서로 독립(independent)이므로, 긴 간격을 찾는 일은 독립 시행의 열이 된다. 각 시행은 "성공"(30분 이상의 간격) 아니면 "실패"이다. 나머지는 기하분포(geometric distribution)와 조건부 기댓값(conditional expectation)이 해결한다.

풀이

T의 분포

방해는 평균 10분마다 일어나므로 비율은 분당 $\frac{1}{10}$이고, t분 동안의 방해 횟수는 평균이 $\frac{t}{10}$인 포아송 분포를 따른다.

방금 일어난 방해 시점에서 시계를 시작하자. 다음 방해까지의 시간 T가 t 이상인 것은 다음 t분 동안 방해가 없을 때와 정확히 같으므로

$$\mathrm{P}(T \geq t) = \mathrm{P}(\text{방해 없는 시간 구간의 길이} t) = \mathrm{e}^{-t/10}.$$

따라서 $t \geq 0$에 대하여 작업 시간이 t분보다 짧을 확률은

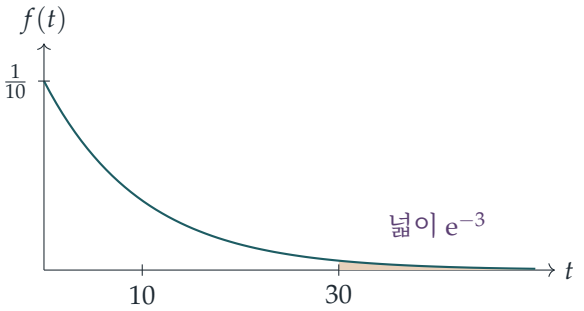
$$\mathrm{P}(T < t) = 1 - \mathrm{e}^{-t/10}.$$

이것이 T의 누적분포함수(cumulative distribution function) $F(t)$이다($t < 0$이면 $F(t) =$

0). 미분하면

$$f(t) = F'(t) = \tfrac{1}{10}\,\mathrm{e}^{-t/10} \quad (t \geq 0), \qquad f(t) = 0 \quad (t < 0).$$

그러므로 T는 평균이 10분인 지수분포를 따른다.



방해 횟수의 기댓값

이 과정은 기억이 없으므로(겹치지 않는 구간에서의 포아송 횟수들은 서로 독립이다), 방해가 일어날 때마다 다음 방해까지의 시간은 이전의 모든 것과 독립적으로 T와 같은 분포를 가지고 새로 시작된다. 따라서 연속한 간격 $T_1, T_2, \ldots$는 서로 독립이고 각각 T와 같은 분포를 따른다.

간격이 30분 이상이면 그 간격을 **긴** 간격이라 하자. 각 간격은 다른 간격과 독립적으로 확률

$$p = \mathrm{P}(T \geq 30) = \mathrm{e}^{-30/10} = \mathrm{e}^{-3},$$

으로 긴 간격이 된다. 첫 번째 긴 간격이 오기 전까지의 방해 횟수를 N이라 하자(방금 일어난 방해는 세지 않는다). 짧은 간격은 모두 방해로 끝나고 첫 번째 긴 간격이 바로 방해받지 않는 30분이므로, N은 첫 번째 긴 간격 이전의 짧은 간격의 개수이다. 따라서

$$\mathrm{P}(N = n) = (1-p)^n p, \qquad n = 0, 1, 2, \ldots$$

? **생각해 보기** 1이 아니라 0에서 시작하는 이 기하분포의 평균은 어떻게 구할까?

기하급수(geometric series) $\sum_{n\geq 0} x^n = \frac{1}{1-x}$를 미분하면 $|x| < 1$일 때 $\sum_{n\geq 1} n x^{n-1} = \frac{1}{(1-x)^2}$을 얻는다. $x = 1-p$로 두면

$$\mathrm{E}(N) = \sum_{n=1}^{\infty} n(1-p)^n p = p(1-p)\cdot\frac{1}{p^2} = \frac{1-p}{p}.$$

(다른 방법으로, $N+1$은 첫 번째 성공을 포함하여 그때까지의 시행 횟수이고, 그 평균은 $\frac{1}{p}$이다.) $p = \mathrm{e}^{-3}$이므로

$$\mathrm{E}(N) = \frac{1-\mathrm{e}^{-3}}{\mathrm{e}^{-3}} = \mathrm{e}^3 - 1,$$

이 되어 요구된 결과를 얻는다.

짧은 간격의 길이의 기댓값

구하려는 것은 $\mathrm{E}(T \mid T < 30)$이다. $T < 30$이라는 조건 아래에서 T의 조건부 확률밀도함수(probability density function)는 $0 \le t < 30$에서 $f(t)/\mathrm{P}(T < 30)$이므로

$$\mathrm{E}(T \mid T < 30) = \frac{1}{1 - \mathrm{e}^{-3}} \int_0^{30} \frac{t}{10} \mathrm{e}^{-t/10} \, \mathrm{d}t.$$

$u = t$, $\mathrm{d}v = \frac{1}{10}\mathrm{e}^{-t/10} \, \mathrm{d}t$로 두고 부분적분(integration by parts)하면 $v = -\mathrm{e}^{-t/10}$이므로

$$\begin{aligned} \int_0^{30} \frac{t}{10} \mathrm{e}^{-t/10} \, \mathrm{d}t &= \Big[-t \, \mathrm{e}^{-t/10} \Big]_0^{30} + \int_0^{30} \mathrm{e}^{-t/10} \, \mathrm{d}t \\ &= -30 \, \mathrm{e}^{-3} + 10(1 - \mathrm{e}^{-3}) \\ &= 10 - 40 \, \mathrm{e}^{-3}. \end{aligned}$$

따라서 분자와 분모에 e^3을 곱하면

$$\mathrm{E}(T \mid T < 30) = \frac{10 - 40 \, \mathrm{e}^{-3}}{1 - \mathrm{e}^{-3}} = \frac{10 \, \mathrm{e}^3 - 40}{\mathrm{e}^3 - 1} \text{ 분} \approx 8.43 \text{ 분}.$$

이 값이 10보다 조금 작은 것은 자연스럽다. 긴 간격을 버렸기 때문이다.

대기 시간의 기댓값

방해받지 않는 30분이 시작되기 전까지의 대기 시간은 N개의 짧은 간격의 합이다. $N = n$일 때 이것들은 각각 30분보다 짧다는 조건이 붙은 n개의 독립인 간격이므로, $N = n$일 때 대기 시간의 기댓값은 $n \, \mathrm{E}(T \mid T < 30)$이다. N에 대하여 평균을 내면

$$\mathrm{E}(\text{대기 시간}) = \mathrm{E}(N) \, \mathrm{E}(T \mid T < 30) = (\mathrm{e}^3 - 1) \cdot \frac{10 \, \mathrm{e}^3 - 40}{\mathrm{e}^3 - 1} = 10 \, \mathrm{e}^3 - 40 \text{ 분},$$

이고, 이는 약 161분, 즉 2시간 41분이다. (논문을 **끝낼** 때까지의 시간을 원한다면 30분 자체를 더하면 된다: $10 \, \mathrm{e}^3 - 10 \approx 191$분.)

검산

- $\int_0^\infty \frac{1}{10}\mathrm{e}^{-t/10} \, \mathrm{d}t = 1$이므로 f는 실제로 확률밀도함수이고, 그 평균은 10으로 "평균 10분마다"와 일치한다.
- 조건 없이 $\mathrm{E}(T) = 10$은 $\mathrm{P}(T < 30) \, \mathrm{E}(T \mid T < 30) + \mathrm{P}(T \ge 30) \, \mathrm{E}(T \mid T \ge 30)$로 나뉘어야 한다. 비기억성(memorylessness)에 의해 $\mathrm{E}(T \mid T \ge 30) = 40$이고, 실제로 $(10 - 40\mathrm{e}^{-3}) + 40\mathrm{e}^{-3} = 10$이다.
- 이 과정을 200 000번 시뮬레이션한 결과 평균 방해 횟수는 19.1회($\mathrm{e}^3 - 1 \approx 19.09$와 비교), 평균 대기 시간은 161.1분($10\mathrm{e}^3 - 40 \approx 160.86$과 비교)이었다.

흔한 실수

- $\mathrm{E}(N) = \frac{1}{p} = e^3$을 사용하는 것: 이것은 긴 간격까지 세는 것인데, 긴 간격은 30분이 지나기 전에 방해 없이 끝난다.
- 대기 시간을 구하려고 $\mathrm{E}(N)$에 조건 없는 평균 10을 곱하는 것. 대기 시간을 이루는 간격들은 짧다는 것이 알려져 있으므로, 그 평균은 10이 아니라 $\mathrm{E}(T \mid T < 30)$이다.

STEP 너머

방해들은 **포아송 과정**(Poisson process)을 이루며, 간격들이 서로 독립인 지수분포를 따른다는 결과는 그 정의적 성질 가운데 하나이다. 마지막 단계인 $\mathrm{E}(\text{항}N\text{개의 합}) = \mathrm{E}(N) \times \mathrm{E}(\text{항})$은 **왈드 항등식**(Wald's identity)과 밀접하게 관련된 조건부 논증이다. 이 둘은 함께 대학 1, 2학년 확률론 강의에서 배우는 재생 이론(renewal theory)과 대기행렬 모형(queueing models)으로 이어진다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 포아송 과정과 조건부 기댓값을 알맞은 수준에서 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.