풀이

STEP III 2002, 문제 13

확률과 통계 • 지수분포 확률변수의 최솟값

STEP III 2002, 13번 문제

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

step.maths.org/questions/02-s3-q13

핵심 아이디어

"둘 중 적어도 하나가 x보다 작다"는 "둘 다 x 이상이다"의 여사건(complement)이고, 두 번째 사건의 확률은 독립성(independence)에 의해 곱으로 분해된다. 지수분포(exponential distribution)의 꼬리 확률은 $\mathrm{P}(X_i \geq x) = \mathrm{e}^{-\lambda_i x}$로 아주 간단하므로, 그 곱도 다시 지수분포의 꼬리 확률이 되며, 최솟값에 관한 증명은 이것이 전부이다. 버스 문제는 이 결과를 바로 적용한 것이다. 다음 버스까지 기다리는 시간은 지수분포를 따르는 두 배차 간격 중 작은 쪽이다.

풀이

둘 중 하나 또는 둘 다 x보다 작을 확률

먼저 지수분포를 따르는 확률변수 하나의 꼬리 확률을 구한다. $x \geq 0$일 때

$$\mathrm{P}(X_i \geq x) = \int_x^\infty \lambda_i \mathrm{e}^{-\lambda_i t}\,\mathrm{d}t = \left[-\mathrm{e}^{-\lambda_i t}\right]_x^\infty = \mathrm{e}^{-\lambda_i x}.$$

$X_1 < x$ 또는 $X_2 < x$(또는 둘 다)인 사건을 E라 하자. 그 여사건은 $X_1 \geq x$와 $X_2 \geq x$가 둘 다 성립하는 사건이다. X_1과 X_2는 독립이므로

$$\mathrm{P}(X_1 \geq x \text{ 이고} X_2 \geq x) = \mathrm{P}(X_1 \geq x)\,\mathrm{P}(X_2 \geq x) = \mathrm{e}^{-\lambda_1 x}\mathrm{e}^{-\lambda_2 x},$$

따라서 $x \geq 0$일 때

$$\boxed{\mathrm{P}(E) = 1 - \mathrm{e}^{-(\lambda_1+\lambda_2)x}.}$$

이다. ($x < 0$이면 두 확률변수 모두 음수가 될 수 없으므로 확률은 0이다.)

? **생각해 보기** 그냥 $\mathrm{P}(X_1 < x)$와 $\mathrm{P}(X_2 < x)$를 더하면 왜 안 될까?

두 사건이 겹치기 때문이다. 둘이 동시에 x보다 작을 수 있으므로, 더하면 그 경우를 두 번 세게 된다. 포함배제의 원리(inclusion–exclusion)를 쓰면 $(1-\mathrm{e}^{-\lambda_1 x})+(1-\mathrm{e}^{-\lambda_2 x})-(1-\mathrm{e}^{-\lambda_1 x})(1-\mathrm{e}^{-\lambda_2 x})$가 되고, 이를 정리하면 같은 답이 나오지만 여사건을 이용하는 편이 더 빠르다.

최솟값은 지수분포를 따른다

$X=\min(X_1, X_2)$라 하자. 두 수 중 작은 쪽이 x보다 작은 것은 둘 중 적어도 하나가 x보다 작을 때와 정확히 같으므로, $\{X<x\}$는 위의 사건 E이다. 따라서 X의 분포함수(distribution function)는

$$F(x)=\mathrm{P}(X\le x)=1-\mathrm{e}^{-(\lambda_1+\lambda_2)x} \qquad (x\ge 0),$$

이고, $x<0$일 때 $F(x)=0$이다. (연속확률변수에서는 $\mathrm{P}(X<x)=\mathrm{P}(X\le x)$이다.) 미분하면 X의 확률밀도함수(density)는

$$f(x)=F'(x)=(\lambda_1+\lambda_2)\,\mathrm{e}^{-(\lambda_1+\lambda_2)x} \qquad (0\le x<\infty),$$

이고, 이는 모수(parameter)가 $\lambda_1+\lambda_2$인 지수분포의 확률밀도함수이다. 따라서 요구한 대로 X는 지수분포를 따른다.

버스

모수가 λ인 지수분포의 평균(mean)은 (부분적분으로) $\int_0^\infty t\lambda\mathrm{e}^{-\lambda t}\,\mathrm{d}t = 1/\lambda$이다. 시간을 분 단위로 재면, 평균 배차 간격 15분과 30분에서 모수는

$$\lambda_A=\tfrac{1}{15}, \qquad \lambda_B=\tfrac{1}{30}.$$

이다. 각 노선의 버스가 방금 떠났으므로, 다음 A 노선 버스까지의 시간은 A 노선 버스 사이의 온전한 배차 간격 T_A이고, B 노선도 마찬가지로 T_B이다. 이 둘은 서로 독립이며 각각 모수가 λ_A, λ_B인 지수분포를 따른다. 어느 노선을 타도 대학에 갈 수 있으므로 내가 기다리는 시간은 $W=\min(T_A, T_B)$이고, 방금 증명한 결과에 의해 W는 모수가

$$\lambda_A+\lambda_B=\tfrac{1}{15}+\tfrac{1}{30}=\tfrac{1}{10}.$$

인 지수분포를 따른다. 중앙값(median) m은 $\mathrm{P}(W\le m)=\frac{1}{2}$를 만족한다.

$$1-\mathrm{e}^{-m/10}=\tfrac{1}{2} \iff \mathrm{e}^{-m/10}=\tfrac{1}{2} \iff m=10\ln 2.$$

$\ln 2\approx 0.693$이므로 기다리는 시간의 중앙값은 $m\approx 6.93$분이고, 요구한 대로 약 7분이다.

검산

- $\lambda_2 \to 0$이면(B 노선 버스가 거의 오지 않으면) 답은 $1 - \mathrm{e}^{-\lambda_1 x}$가 되어, 마땅히 그래야 하듯 X_1의 분포와 같아진다.
- 기다리는 시간의 평균은 10분으로 두 노선의 평균 15분, 30분보다 짧고, 중앙값 $10\ln 2$는 오른쪽으로 치우친 분포에서 흔히 그렇듯 평균보다 작다.
- 평균이 15와 30인 독립 지수분포 배차 간격 2×10^5쌍을 모의실험하였더니, 최솟값의 중앙값은 약 6.91분으로 $10\ln 2 = 6.931\ldots$와 비슷하게 나왔다.

흔한 실수

- 평균을 모수로 쓰는 실수: 평균이 15인 지수분포의 모수는 15가 아니라 $\lambda = \frac{1}{15}$이다.
- 두 확률을 $\mathrm{P}(X_1 < x) + \mathrm{P}(X_2 < x)$처럼 더하는 실수: 이렇게 하면 중복해서 세게 되고 값이 1을 넘을 수도 있다.
- 두 노선의 중앙값을 평균하거나, A 노선만의 중앙값($15\ln 2 \approx 10.4$)을 택하는 실수: 노선이 두 개이면 기다리는 시간이 줄어든다.

STEP 너머

버스 사이의 배차 간격이 지수분포를 따른다는 것은 버스가 **푸아송 과정**(Poisson process)으로 도착한다는 뜻이며, 여기서 얻은 결과는 서로 독립인 두 푸아송 과정을 합치면 비율이 두 비율의 합인 또 다른 푸아송 과정이 된다는 것을 말해 준다. 무기억성(memoryless property) $\mathrm{P}(X > s + t \mid X > s) = \mathrm{P}(X > t)$와 함께, 이것은 대학 1학년 확률론 강의에서 만나는 대기행렬 이론(queueing theory)과 연속시간 마르코프 연쇄(continuous-time Markov chain)의 토대가 된다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 지수분포를 따르는 확률변수와 푸아송 과정을 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.