풀이

STEP I 견본 문제지, 문제 16

확률과 통계 • 연속확률변수

STEP I 견본 문제지, 16번 문제

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

step.maths.org/questions/spec-s1-q16

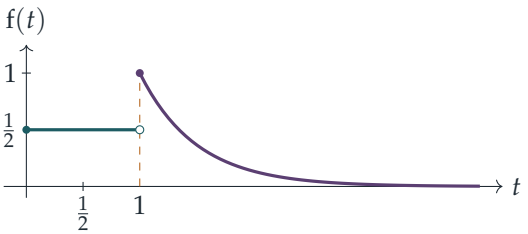
핵심 아이디어

미지수 k는 전체 확률이 1이라는 사실에서 정해지고, 나머지는 모두 밀도함수(density)의 두 조각을 따로 적분하면 얻어진다. 비용을 구할 때의 요령은 연속(continuous) 확률변수 T를 이산(discrete) 확률변수로 바꾸는 것이다. 동전의 개수는 완전히 지나간 30초 구간의 개수에만 의존하므로, T가 각 30초 구간에 들어갈 확률을 구하면 된다. 그다음은 등비급수(geometric series) 꼴의 급수가 해결해 준다.

풀이

k와 평균 구하기

밀도함수는 $[0,1)$에서 높이가 $\frac{1}{2}$인 평평한 조각과 그 뒤에 이어지는 지수함수 꼴의 꼬리(tail)로 이루어져 있으며, $t=1$에서 불연속적으로 뛴다.



첫 번째 조각이 확률 $\frac{1}{2}$을 차지하므로, 꼬리가 나머지 $\frac{1}{2}$을 차지해야 한다.

$$k\int_1^\infty \mathrm{e}^{-2t}\,\mathrm{d}t = k\left[-\tfrac{1}{2}\mathrm{e}^{-2t}\right]_1^\infty = \tfrac{1}{2}k\mathrm{e}^{-2} = \tfrac{1}{2}, \qquad \text{따라서} \qquad k = \mathrm{e}^2.$$

(따라서 $t \geq 1$에서 밀도함수는 e^{2-2t}이고, 이는 그림에서처럼 $t = 1$에서 1이다.)

평균(mean)을 구하기 위해 $u = t$, $\mathrm{d}v = \mathrm{e}^{-2t}\,\mathrm{d}t$로 두고 부분적분(integration by parts)을 하면

$$\int_1^\infty t\,\mathrm{e}^{-2t}\,\mathrm{d}t = \left[-\tfrac{1}{2}t\,\mathrm{e}^{-2t}\right]_1^\infty + \tfrac{1}{2}\int_1^\infty \mathrm{e}^{-2t}\,\mathrm{d}t = \tfrac{1}{2}\mathrm{e}^{-2} + \tfrac{1}{4}\mathrm{e}^{-2} = \tfrac{3}{4}\mathrm{e}^{-2},$$

이다. 여기서 $t \to \infty$일 때 $t\,\mathrm{e}^{-2t} \to 0$임을 이용했다. 그러므로

$$\mathrm{E}(T) = \int_0^1 \tfrac{1}{2}t\,\mathrm{d}t + \mathrm{e}^2 \cdot \tfrac{3}{4}\mathrm{e}^{-2} = \tfrac{1}{4} + \tfrac{3}{4} = 1,$$

이 되어, 문제에서 요구한 대로 통화 시간의 기댓값(expected value)은 1분이다.

누적분포함수

$\mathrm{F}(t) = \mathrm{P}(T \leq t)$라 하자. $t < 0$이면 0이고, $0 \leq t < 1$이면 $\int_0^t \tfrac{1}{2}\,\mathrm{d}s = \tfrac{1}{2}t$이며, $t \geq 1$이면

$$\mathrm{F}(t) = \tfrac{1}{2} + \int_1^t \mathrm{e}^{2-2s}\,\mathrm{d}s = \tfrac{1}{2} + \tfrac{1}{2}(1 - \mathrm{e}^{2-2t}) = 1 - \tfrac{1}{2}\mathrm{e}^{2-2t}.$$

이다. 따라서 누적분포함수(cumulative distribution function)는

$$\mathrm{F}(t) = \begin{cases} 0 & t < 0, \\ \tfrac{1}{2}t & 0 \leq t < 1, \\ 1 - \tfrac{1}{2}\mathrm{e}^{2-2t} & t \geq 1. \end{cases}$$

이다. f는 $t = 1$에서 뛰지만 F는 $t = 1$에서 연속이고(두 식 모두 $\tfrac{1}{2}$을 준다), $t \to \infty$일 때 $\mathrm{F}(t) \to 1$임에 주목하자. 앞으로 사용할 형태는 다음과 같다. $t \geq 1$이면 $\mathrm{P}(T > t) = \tfrac{1}{2}\mathrm{e}^{2-2t}$이다.

비용의 기댓값

통화를 시작할 때 동전 하나를 넣고, 30초가 완전히 지날 때마다 하나씩 더 넣는다. 따라서 통화가 T분 동안 이어지면 완전히 지난 30초 구간의 개수는 $\lfloor 2T \rfloor$이고

$$C = 10(1 + \lfloor 2T \rfloor).$$

이다. (통화가 정확히 30초 단위의 순간에 끝날 때 동전을 넣는지 여부는 상관없다. T는 연속 확률변수이므로 그런 일이 일어날 확률은 0이다.)

? 생각해 보기 연속인 T에서 C의 분포로 어떻게 넘어갈 수 있을까?

사건 $C = 10(n+1)$은 정확히 사건 $\tfrac{n}{2} \leq T < \tfrac{n+1}{2}$와 같으므로, 그 확률을 F에서 바로 읽어 낼 수 있다.

- $n = 0$: $\mathrm{P}(C = 10) = \mathrm{F}(\frac{1}{2}) = \frac{1}{4}$;
- $n = 1$: $\mathrm{P}(C = 20) = \mathrm{F}(1) - \mathrm{F}(\frac{1}{2}) = \frac{1}{4}$;
- $n \geq 2$: 두 끝점이 모두 1 이상이므로

$$\mathrm{P}\big(C = 10(n+1)\big) = \tfrac{1}{2}\mathrm{e}^{2-n} - \tfrac{1}{2}\mathrm{e}^{2-(n+1)} = \tfrac{1}{2}\big(1 - \mathrm{e}^{-1}\big)\mathrm{e}^{2-n}.$$

$r = \mathrm{e}^{-1}$, $m = n - 2$로 두면 $\mathrm{e}^{2-n} = r^m$이고 $10(n+1) = 10(m+3)$이다. 그러면

$$\begin{aligned}\mathrm{E}(C) &= 10 \cdot \tfrac{1}{4} + 20 \cdot \tfrac{1}{4} + \sum_{m=0}^{\infty} 10(m+3) \cdot \tfrac{1}{2}(1-r)\, r^m \\ &= \tfrac{15}{2} + 5(1-r)\left(\sum_{m=0}^{\infty} m r^m + 3\sum_{m=0}^{\infty} r^m\right).\end{aligned}$$

이다. $0 < r < 1$이므로 등비급수에서 $\sum r^m = \dfrac{1}{1-r}$이고, 이를 항별로 미분(differentiate term by term)한 뒤 r를 곱하면 $\sum m r^m = \dfrac{r}{(1-r)^2}$를 얻는다. 그러므로

$$\mathrm{E}(C) = \tfrac{15}{2} + 5(1-r)\left(\frac{r}{(1-r)^2} + \frac{3}{1-r}\right) = \tfrac{15}{2} + \frac{5r}{1-r} + 15.$$

이다. 마지막으로 $\dfrac{r}{1-r} = \dfrac{\mathrm{e}^{-1}}{1-\mathrm{e}^{-1}} = \dfrac{1}{\mathrm{e}-1}$이므로

$$\mathrm{E}(C) = 22\tfrac{1}{2} + \frac{5}{\mathrm{e}-1},$$

이 되어 요구한 결과를 얻는다. 수치로는 약 25.41펜스이다.

해석

1분에 20펜스라면 평균 길이가 1분인 통화는 평균 20펜스가 들어야 할 것 같지만, 실제로는 $\mathrm{E}(C) \approx 25.4$이다. 이런 차이가 생기는 까닭은 그 남자가 30초마다 **미리** 돈을 내기 때문이다. 동전 하나는 30초를 통째로 사지만, 마지막 동전으로 산 시간은 거의 언제나 다 쓰이지 않는다. 실제로 $\lfloor 2T \rfloor > 2T - 1$이므로 모든 통화에 대해 $C = 10(1 + \lfloor 2T \rfloor) > 20T$이다. 비용은 통화 시간을 다음 30초 단위로 **올림**하므로, $\mathrm{E}(C)$는 $20\,\mathrm{E}(T)$보다 돈을 내고도 쓰지 않은 시간의 기댓값의 20배만큼(여기서는 약 5.4펜스) 더 크다.

검산

- C의 확률을 모두 더하면 1이다: $\frac{1}{4} + \frac{1}{4} + \sum_{m \geq 0} \frac{1}{2}(1-r)r^m = \frac{1}{2} + \frac{1}{2} = 1$.
- 부등식 $20T < C \leq 20T + 10$에서 $20 < \mathrm{E}(C) \leq 30$이고, 25.41은 넉넉히 그 안에 들어간다. 마지막 30초 구간에서 쓰지 않은 시간이 균등분포(uniform distribution)를 따른다면 초과분은 약 5펜스로 예상되는데, 긴 꼬리 때문에 조금 더 커진다.

- 수치적으로 심프슨 공식(Simpson's rule)을 쓰면 전체 확률 1과 평균 1을 얻고, 통화 10^6건을 시뮬레이션하면(절반은 $[0,1)$에서 균등분포, 절반은 1에 비율 2인 지수분포(exponential distribution)를 더한 값) 평균 비용이 25.41펜스로 나와 $22.5 + 5/(\mathrm{e} - 1) = 25.4099$와 일치한다.

흔한 실수

- 처음에 넣는 동전을 잊고 동전의 개수를 $1 + \lfloor 2T \rfloor$가 아니라 $\lfloor 2T \rfloor$로 잡는 실수. 이렇게 하면 답에서 정확히 10펜스가 빠진다.
- $t = \frac{1}{2}$에서 F에 $t \geq 1$일 때의 식을 쓰는 실수. 구간 $[0, \frac{1}{2})$와 $[\frac{1}{2}, 1)$은 균등한 조각으로 다루어야 한다.

STEP 너머

T의 꼬리는 무기억성(memoryless property)을 가진다는 점에 주목하자. $T \geq 1$이라는 조건 아래에서 남은 시간 $T - 1$은 비율 2인 지수분포를 따르며, 바로 이 때문에 $n \geq 2$에 대한 확률들이 등비수열을 이룬다. $N = \lfloor 2T \rfloor$로 두면 꼬리합 공식(tail-sum formula) $\mathrm{E}(N) = \sum_{n \geq 1} \mathrm{P}(N \geq n)$으로 같은 답을 훨씬 더 빨리 얻을 수 있다. 이처럼 이산적 성질과 연속적 성질이 섞인 경우는 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)에서 꼼꼼하게 다루며, Blitzstein과 Hwang의 *Introduction to Probability*(확률 입문)에는 무기억성에 관한 예가 훨씬 더 많이 실려 있다.

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.