풀이

STEP II 1993, 문제 16

확률과 통계 • 푸아송 과정

> **STEP II 1993, 16번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/93-s2-q16

핵심 아이디어

이 문제에는 서로 독립인 두 가지 무작위성의 원천이 있다. 하나는 기하분포(geometric distribution)를 따르는 대기 시간을 주는 버스이고, 다른 하나는 포아송 과정(Poisson process)을 이루는 승객이다. 첫 버스가 언제 오는지를 조건(condition)으로 두자. 그것을 알고 나면 기다리는 사람 수는 알려진 시간 길이 동안의 포아송 개수일 뿐이다. 소문항 (iii)에서는 전확률 법칙(law of total probability)을 쓰는데, 진짜 문제는 그 결과로 나오는 무한급수(infinite series)의 값을 표와 계산기만으로 유효숫자(significant figures) 두 자리까지 어떻게 구하느냐이다.

풀이

시간은 8:00 이후 경과한 분 단위로 잰다. 첫 버스가 도착하는 시각을 T라 하고, $(0, t]$ 동안 도착한 승객 수를 X_t라 하자. 승객 과정은 분당 비율이 4인 포아송 과정이므로 $X_t \sim \mathrm{Po}(4t)$이다. 버스는 $t = 1, 2, 3, \ldots$에만 올 수 있고("8시 직후"라는 말은 8:00에 올 수도 있는 버스가 이미 왔다 갔다는 뜻이다), 승객과는 독립이다.

소문항 (i)

첫 버스가 8시 n분에 도착하는 것은 정확히 $1, 2, \ldots, n-1$분에는 버스가 오지 않고 n분에 버스가 올 때이다. 이 사건들은 서로 독립이므로

$$\mathrm{P}(T = n) = \left(\tfrac{3}{4}\right)^{n-1} \tfrac{1}{4}, \qquad n = 1, 2, 3, \ldots$$

이다. 따라서 T는 모수가 $\frac{1}{4}$인 기하분포를 따른다. (검산으로, 이 확률들의 합은 $\frac{1}{4}/(1 - \frac{3}{4}) = 1$이다.)

소문항 (ii)

8:05에 기다리고 있는 사람들은 정확히 $(0,5]$ 동안 도착한 사람들이다. 8:05 전에는 버스가 오지 않았으므로 떠난 사람은 아무도 없다. 따라서 $T=5$일 때 X_5의 조건부 분포를 구하면 된다.

> ? **생각해 보기** 첫 버스가 8:05에 왔다는 사실을 알면 승객 수의 분포가 달라지는가?

그렇지 않다. 사건 $T=5$는 버스에만 의존하고, 버스는 승객과 독립이다. 따라서 $T=5$를 조건으로 두어도 X_5의 분포는 변하지 않으며, $X_5 \sim \mathrm{Po}(4\times 5)=\mathrm{Po}(20)$이다. 그러므로

$$\mathrm{P}(m\text{명이 기다림} \mid T=5)=\frac{\mathrm{e}^{-20}\,20^m}{m!}, \qquad m=0,1,2,\ldots$$

이다.

소문항 (iii)

버스가 모든 사람을 태우지 못하는 것은 정확히 25명보다 많은 사람이 기다릴 때, 즉 $X_T \geq 26$일 때이다. 이 사건을 F라 하자. T를 조건으로 두고 (ii)에서처럼 독립성을 쓰면

$$\mathrm{P}(F)=\sum_{n=1}^{\infty}\mathrm{P}(T=n)\,\mathrm{P}(X_n\geq 26)=\sum_{n=1}^{\infty}\left(\tfrac{3}{4}\right)^{n-1}\tfrac{1}{4}\,\mathrm{P}(\mathrm{Po}(4n)\geq 26).$$

이다. 이제 이 무한급수를 유한한 양의 계산으로 바꾸되, 오차가 0.005보다 충분히 작도록 하는 계획이 필요하다. 유효숫자 두 자리를 원하는데, 답은 0.1 정도의 크기로 나오게 된다.

1단계: 여사건(complement)을 다루어 꼬리 부분을 쉽게 만든다. $w_n=\left(\tfrac{3}{4}\right)^{n-1}\tfrac{1}{4}$, $q_n=\mathrm{P}(\mathrm{Po}(4n)\leq 25)$라 쓰자. $\sum w_n=1$이므로

$$\mathrm{P}(F)=1-\sum_{n=1}^{\infty}w_n\,q_n.$$

이다. 이 형태가 더 나은 까닭은 n이 클 때 q_n이 아주 작기 때문이다. $n\geq 11$이면 평균 $4n\geq 44$가 25보다 훨씬 크다. n이 커질수록 q_n은 감소하므로(시간이 길수록 승객이 많아진다)

$$\sum_{n\geq 11}w_nq_n\leq q_{11}\sum_{n\geq 11}w_n=q_{11}\left(\tfrac{3}{4}\right)^{10}.$$

이다. 정규근사(normal approximation)를 쓰면 $q_{11}\approx\Phi((25.5-44)/\sqrt{44})=\Phi(-2.79)\approx 0.003$이므로 이 꼬리 부분은 0.0002보다 작아 버려도 된다. 그러면 열 개의 항이 남는다.

2단계: 처음 몇 항. $n = 1$과 $n = 2$(평균 4와 8)에서는 승객이 26명 이상일 가능성이 매우 낮아서, 필요한 정확도를 훨씬 넘어서까지 $q_1 = q_2 = 1$이다. 따라서 이 항들은 $w_1 + w_2 = \frac{1}{4} + \frac{3}{16}$만큼 기여한다.

3단계: 표에서 읽는 중간 항. $n = 3, 4, 5$(평균 12, 16, 20)에서는 $q_n = \mathrm{P}(\mathrm{Po}(4n) \le 25)$를 누적 포아송 분포표(cumulative Poisson tables)에서 바로 읽는다. 보통 이 표는 평균 20까지 나와 있다. 표에 평균 24와 28도 있다면 $n = 6, 7$에도 그것을 쓴다.

4단계: 더 큰 평균은 정규근사로. 10까지의 나머지 n(평균 40까지)에서는 연속성 수정(continuity correction)을 하여 $\mathrm{Po}(\mu) \approx \mathrm{N}(\mu, \mu)$를 쓴다.

$$q_n \approx \Phi\left(\frac{25.5 - 4n}{\sqrt{4n}}\right),$$

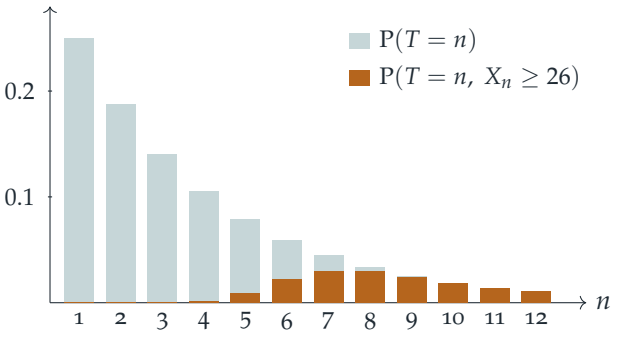
여기서 Φ는 정규분포표에서 읽는다. 이 평균들은 근사가 좋을 만큼 충분히 크다. 더 중요한 점은 가중치 w_n이 이미 작다는 것이다($w_8 \approx 0.033$, $w_{10} \approx 0.019$). 따라서 q_n에 천분의 몇 정도의 오차가 있어도 $\mathrm{P}(F)$는 0.001보다 적게 변한다.

5단계: 합치기. 계산기로 $w_1, \ldots, w_{10}$을 구하고

$$\mathrm{P}(F) \approx 1 - \sum_{n=1}^{10} w_n q_n,$$

을 계산한 다음 유효숫자 두 자리로 반올림한다. 버리거나 근사한 각 부분은 많아야 0.001 정도이므로, 합계가 반올림 경계에 아주 가깝게 떨어지지 않는 한 반올림은 안전하다. 만약 경계에 가깝다면, 이 방법은 어느 항을 더 정밀하게 계산해야 하는지 정확히 보여 준다.

문제에서 이것을 실제로 계산하라고 요구하지는 않지만, 계산해 보면 $\mathrm{P}(F) \approx 0.19$를 얻는다. 아래 그림은 이 확률이 어디서 나오는지 보여 준다. 대부분은 첫 버스가 8:06에서 8:10 사이에 오는 경우에서 나오는데, 이때 기대 대기 인원 24에서 40명은 정원에 가깝거나 정원을 넘는다.



검산

- (i)에서 확률들의 합은 1이고, 당연히 $\mathrm{P}(T = 1) = \frac{1}{4}$이다.
- (ii)에서 기다리는 사람 수의 평균은 20이며, 이는 5분 동안 분당 4명이다.
- (iii)에서 (컴퓨터로 계산한) 정확한 합은 0.1904이다. 2×10^5번의 아침을 몬테카를로 시뮬레이션(Monte Carlo simulation)하면 0.189를 얻는다. 위의 표와 정규근사를 쓴 방법은 0.1912를 준다. 세 값 모두 반올림하면 0.19이다.

흔한 실수

- 경계를 하나 틀리는 것. 버스는 26명 이상이 기다릴 때만 모두를 태우지 못하므로, 구해야 할 확률은 $1 - \mathrm{P}(\mathrm{Po}(4n) \leq 24)$가 아니라 $1 - \mathrm{P}(\mathrm{Po}(4n) \leq 25)$이다.
- 대기 인원을 $\mathrm{Po}(4\,\mathrm{E}[T]) = \mathrm{Po}(16)$으로 취급하는 것. 대기 시간이 확률적이므로, 평균 시간을 대입할 것이 아니라 포아송 확률들을 T에 대해 평균해야 한다.
- 왜 잘라도 되는지 말하지 않고 급수를 무한한 채로 두는 것. 버린 꼬리 부분이 너무 작아서 유효숫자 둘째 자리에 영향을 주지 않음을 보여야 한다.

STEP 너머

기다리는 사람 수 X_T는 평균이 확률적인 포아송 개수, 즉 포아송 분포들의 **혼합**(*mixture*)이다. T가 기하분포를 따르므로 X_T는 깔끔한 닫힌 형태의 생성함수(generating function)를 갖게 되며, 이와 같은 조건부 논증은 대기행렬 이론(queueing theory)과, 처음 배우는 확률론 강의에서 만나는 복합 분포(compound distribution)와 혼합 분포의 기초이다. 급수를 어디서 자르고 오차를 어떻게 한정하느냐 하는 문제는 수치해석(numerical analysis)의 출발점이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 포아송 과정과 생성함수를 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.