풀이

STEP I 2006, 문제 13

확률과 통계 • 푸아송 혼합분포

> **STEP I 2006, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/06-s1-q13

핵심 아이디어

다이아몬드가 일정한 평균 비율로 서로 독립적으로 흩어져 있으면 이는 푸아송 과정(Poisson process)이므로, m킬로그램의 초콜릿에 들어 있는 다이아몬드의 개수는 평균이 m인 푸아송 분포(Poisson distribution)를 따른다. 그런데 초콜릿의 양 자체가 확률적이므로, 그 양에 대해 조건부(conditioning)로 생각한다. 즉 양이 고정되어 있을 때의 답을 먼저 구한 뒤, 전확률 법칙(law of total probability)과 그 기댓값 버전을 이용하여 주사위 결과에 대해 평균을 낸다. 그러면 각 문항은 급수(series)의 합을 구하는 문제로 바뀌는데, (i)에서는 유한급수이고 (ii)에서는 등비급수이다.

풀이

내가 찾는 다이아몬드의 개수를 D라 하자. 다이아몬드는 킬로그램당 평균 한 개의 비율로 서로 독립적으로 숨겨져 있으므로, 질량이 $m\,\mathrm{kg}$인 초콜릿 바에 들어 있는 다이아몬드의 개수는 평균이 m인 푸아송 분포를 따른다. 특히 질량이 $m\,\mathrm{kg}$으로 주어졌을 때

$$\mathrm{P}(D = 0) = \mathrm{e}^{-m}, \qquad \mathrm{E}(D) = m.$$

이하에서 $100k$그램은 $0.1k$킬로그램이다.

소문항 (i)

주사위의 눈을 N이라 하면, $n = 1, \ldots, 6$에 대하여 $\mathrm{P}(N = n) = \frac{1}{6}$이다. $N = n$일 때 나는 $0.1n$ kg을 사므로, $D \mid N = n$은 평균이 $0.1n$인 푸아송 분포를 따른다. 전확률 법칙에 의해

$$\begin{aligned}\mathrm{P}(D = 0) &= \sum_{n=1}^{6} \mathrm{P}(N = n)\,\mathrm{P}(D = 0 \mid N = n) = \frac{1}{6}\sum_{n=1}^{6} \mathrm{e}^{-0.1n} \\ &= \frac{1}{6}\left(\mathrm{e}^{-0.1} + \mathrm{e}^{-0.2} + \cdots + \mathrm{e}^{-0.6}\right).\end{aligned}$$

이것은 첫째항이 $\mathrm{e}^{-0.1}$, 공비(common ratio)가 $\mathrm{e}^{-0.1}$이고 항이 여섯 개인 등비급수(geometric series)이므로

$$\mathrm{P}(D = 0) = \frac{\mathrm{e}^{-0.1}}{6}\left(\frac{1 - \mathrm{e}^{-0.6}}{1 - \mathrm{e}^{-0.1}}\right),$$

이고, 이것이 구하려던 결과이다.

기댓값(expectation)도 같은 방식으로 조건부로 구한다.

$$\mathrm{E}(D) = \sum_{n=1}^{6} \mathrm{P}(N = n)\,\mathrm{E}(D \mid N = n) = \frac{1}{6}\sum_{n=1}^{6} 0.1n = \frac{0.1 \times 21}{6} = 0.35.$$

다시 말해 $\mathrm{E}(D) = 0.1\,\mathrm{E}(N) = 0.1 \times 3.5$이며, 이것이 보여야 할 내용이다.

소문항 (ii)

이제 처음으로 6이 나오는 던지기의 번호를 T라 하자. 각 던지기는 독립이고, $T = t$는 6이 아닌 눈이 $t - 1$번 나온 뒤 6이 나온다는 뜻이므로

$$\mathrm{P}(T = t) = \left(\tfrac{5}{6}\right)^{t-1}\tfrac{1}{6}, \qquad t = 1, 2, 3, \ldots$$

$T = t$일 때 나는 $0.1t$ kg을 사므로 $\mathrm{P}(D = 0 \mid T = t) = \mathrm{e}^{-0.1t}$이다. 따라서

$$\mathrm{P}(D = 0) = \sum_{t=1}^{\infty}\left(\tfrac{5}{6}\right)^{t-1}\tfrac{1}{6}\,\mathrm{e}^{-0.1t} = \frac{\mathrm{e}^{-0.1}}{6}\sum_{t=1}^{\infty}\left(\tfrac{5}{6}\mathrm{e}^{-0.1}\right)^{t-1}.$$

? **생각해 보기** 이 급수를 무한히 더해도 되는 이유는 무엇인가?

이것은 공비가 $r = \frac{5}{6}\mathrm{e}^{-0.1}$인 등비급수이고 $0 < r < 1$이므로, 수렴(converge)하며 그 합은 $1/(1 - r)$이다. 따라서

$$\mathrm{P}(D = 0) = \frac{\mathrm{e}^{-0.1}}{6} \cdot \frac{1}{1 - \frac{5}{6}\mathrm{e}^{-0.1}} = \frac{\mathrm{e}^{-0.1}}{6 - 5\mathrm{e}^{-0.1}},$$

이다(분자와 분모에 6을 곱하였다). 이것이 구하려던 결과이다.

다이아몬드 개수의 기댓값은

$$\mathrm{E}(D) = \sum_{t=1}^{\infty} \mathrm{P}(T=t)\,\mathrm{E}(D \mid T=t) = \sum_{t=1}^{\infty} \left(\tfrac{5}{6}\right)^{t-1} \tfrac{1}{6} \times 0.1t = \frac{0.1}{6} \sum_{t=1}^{\infty} t \left(\tfrac{5}{6}\right)^{t-1}.$$

여기서 힌트가 쓰인다. $|x| < 1$일 때 이항 전개(binomial expansion)에 의해

$$(1-x)^{-2} = 1 + 2x + 3x^2 + 4x^3 + \cdots = \sum_{t=1}^{\infty} t x^{t-1}.$$

$x = \frac{5}{6}$이면 합은 $\left(1 - \frac{5}{6}\right)^{-2} = 36$이므로

$$\mathrm{E}(D) = \frac{0.1}{6} \times 36 = 0.6.$$

이번에도 이 값은 $0.1\,\mathrm{E}(T)$인데, 여기서 $\mathrm{E}(T) = 6$은 6이 나올 때까지 필요한 던지기 횟수의 평균으로 잘 알려진 값이다.

검산

- 수치로 계산하면 (i)의 확률은 약 0.7150, (ii)의 확률은 약 0.6131이다. 다이아몬드 개수를 푸아송 분포로 생성하여 구매자 2×10^5명을 시뮬레이션한 결과 0.7151과 0.6136을 얻었고, 다이아몬드 개수의 평균은 0.350과 0.599였다.
- (ii)의 확률이 (i)의 확률보다 작다. 이는 타당하다. 평균적으로 350 g이 아니라 600 g을 사기 때문이다.
- 두 기댓값 모두 확률적인 배수의 평균에 0.1을 곱한 값이다. 질량이 주어졌을 때 다이아몬드 개수의 기댓값이 질량에 대해 선형(linear)이므로, 다이아몬드 개수의 기댓값은 킬로그램 수의 기댓값과 같다.

흔한 실수

- 다이아몬드가 하나도 없을 확률로 $\mathrm{e}^{-0.35}$(또는 $\mathrm{e}^{-0.6}$)를 계산하는 것. 이는 평균 질량만 사용한 것인데, e^{-m}은 m에 대해 선형이 아니므로 평균은 e^{-m}을 취하기 전이 아니라 취한 뒤에 내야 한다. 기댓값은 선형이므로 $\mathrm{E}(D)$에 대해서는 이 지름길이 실제로 통한다.
- 그램을 킬로그램으로 바꾸는 것을 잊어서 평균을 $0.1n$이 아니라 $100n$으로 얻는 것.
- $\mathrm{P}(T=t) = (\frac{5}{6})^t \frac{1}{6}$이라고 쓰는 것. 이는 6이 아닌 눈을 하나 더 센 것이다.

STEP 너머

여기서는 푸아송 분포의 모수 자체가 확률적이므로 D는 **혼합**(mixture) 분포가 된다. 사용한 도구인 전확률 법칙과 탑 법칙(tower law) $\mathrm{E}(D) = \mathrm{E}(\mathrm{E}(D \mid T))$은 대학 1학

년 확률론의 핵심이다. 확률생성함수(probability generating function)를 쓰면 이런 문제를 깔끔하게 다룰 수 있다. $\mathrm{P}(D = 0)$은 T의 생성함수에 $\mathrm{e}^{-0.1}$을 대입한 값이다. 시행 횟수를 확률적으로 정하고 각 시행에서 다시 확률적인 개수를 세는 구조는 복합 분포(compound distribution)와 분기 과정(branching process)으로 이어진다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)에서 이 모든 내용을 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.