풀이

STEP II 1994, 문제 13

확률과 통계 • 쿠폰 수집가 문제

> **STEP II 1994, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/94-s2-q13

이 문제는 *Beyond STEP*의 9.11 절 **한 세트 모두 모으기**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch09/coupon-collector.html

핵심 아이디어

이 문제는 고전적인 "쿠폰 수집가 문제(coupon collector problem)"이며, 데이터베이스라는 분류와 달리 순수한 확률 문제이다. 카드 한 세트를 모두 모으기까지의 긴 기다림을 n개의 짧은 기다림으로 나눈다. 첫 번째 새 카드를 기다리는 시간, 두 번째 새 카드를 기다리는 시간, 이런 식이다. 각각의 짧은 기다림은 기하분포(geometric distribution)를 따르므로 평균을 쉽게 구할 수 있고, 기댓값의 선형성(linearity of expectation)으로 이들을 더하면 된다. 그 합은 n에 조화합(harmonic sum)을 곱한 것이고, 적분 부등식은 이 합이 얼마나 큰지 알려 주기 위해 주어진 것이다.

풀이

새 카드 기다리기

각 봉지에는 n장의 카드 중 한 장이 들어 있고, 각 카드가 들어 있을 확률은 똑같이 $\frac{1}{n}$이며, 다른 모든 봉지와 독립(independent)이라고 모형화한다.

내가 이미 서로 다른 카드 r장을 가지고 있다고 하자. 여기서 $0 \le r \le n-1$이다. 봉지 하나에서 새 카드가 나올 확률은

$$p = \frac{n-r}{n} > 0,$$

이고, 이미 가진 카드가 나올 확률은 $q = 1 - p = \frac{r}{n}$이다. 첫 번째 새 카드가 나올 때까지(그

봉지를 포함하여) 뜯는 봉지의 개수를 X라 하자. 봉지들은 서로 독립이므로 $X = k$는 이미 가진 카드가 $k-1$번 나온 뒤 새 카드가 나온다는 뜻이고, 따라서

$$\mathrm{P}(X = k) = q^{k-1}p, \qquad k = 1, 2, 3, \ldots$$

이다. 그러므로

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} kpq^{k-1} = p\,(1-q)^{-2} = \frac{p}{p^2} = \frac{1}{p},$$

이다. 여기서 $|q| < 1$일 때 성립하는 $\sum_{k\geq 1} kq^{k-1} = (1-q)^{-2}$를 이용하였다(기하급수(geometric series)를 미분하거나, $(1-q)^{-2}$를 이항전개(binomial expansion)하면 된다). $r \leq n-1$이므로 여기서 $0 \leq q < 1$이다. 따라서

$$\mathrm{E}(X) = \frac{n}{n-r},$$

이고, 이것이 구하려던 결과이다.

(더 간결한 방법도 있다. 첫 번째 봉지에서 새 카드가 나오면 멈추고, 그렇지 않으면 확률 q로 처음 상태로 되돌아가므로 $\mathrm{E}(X) = 1 + q\,\mathrm{E}(X)$이고, 따라서 $\mathrm{E}(X) = 1/(1-q) = 1/p$이다. 이 방법은 $\mathrm{E}(X)$가 유한하다고 가정하는데, 이는 위의 급수로 확인된다.)

적분 부등식

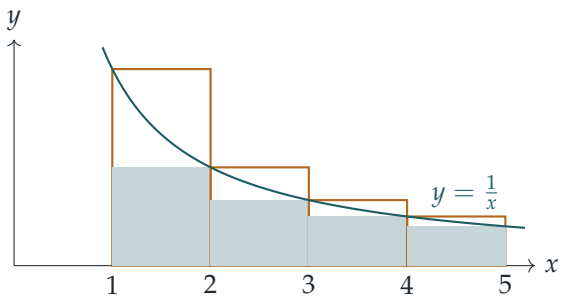
$r \geq 1$이라 하자. $r \leq x \leq r+1$일 때

$$\frac{1}{r+1} \leq \frac{1}{x} \leq \frac{1}{r},$$

이 성립한다. $x > 0$에서 $1/x$가 감소하기 때문이다. 길이가 1인 구간에서 적분해도 이 부등식은 유지되므로

$$\frac{1}{r+1} \leq \int_r^{r+1} \frac{1}{x}\,\mathrm{d}x \leq \frac{1}{r}.$$

이다. 그림으로 보면, $x = r$와 $x = r+1$ 사이에서 $y = 1/x$ 아래의 넓이는 높이가 $\frac{1}{r+1}$인 낮은 직사각형과 높이가 $\frac{1}{r}$인 높은 직사각형의 넓이 사이에 있다.



(알아보기 쉽도록 세로 방향을 늘려 그렸다.) 색칠한 직사각형은 곡선 아래에 있고, 테두리만 그린 직사각형은 곡선 위로 솟아 있다.

이제 $n \geq 2$에 대하여 $r = 1, 2, \ldots, n-1$일 때의 부등식을 모두 더한다. 적분들은 하나로 이어진다.

$$\sum_{r=1}^{n-1} \frac{1}{r+1} \leq \int_1^n \frac{1}{x}\,\mathrm{d}x \leq \sum_{r=1}^{n-1} \frac{1}{r}.$$

가운데 항은 $\ln n - \ln 1 = \ln n$이고, 왼쪽 합에서 $r+1$을 새로 r로 바꾸어 쓰면 범위는 2부터 n까지가 된다. 따라서

$$\sum_{r=2}^{n} \frac{1}{r} \leq \ln n \leq \sum_{r=1}^{n-1} \frac{1}{r},$$

이고, 이것이 구하려던 결과이다.

한 세트 모두 모으기

서로 다른 카드를 정확히 r장 가지고 있는 동안 뜯는 봉지의 개수를 X_r라 하자. 이때 마지막 봉지는 $r+1$번째 카드가 나오는 봉지이다. 한 세트를 모두 모으는 데 필요한 봉지의 총개수는

$$T = X_0 + X_1 + \cdots + X_{n-1}.$$

이다.

> **생각해 보기** X_r들은 앞에서 일어난 일에 따라 달라진다. 그래도 그냥 기댓값을 더해도 될까?

그렇다. 기댓값의 선형성 $\mathrm{E}(A+B) = \mathrm{E}(A) + \mathrm{E}(B)$에는 독립성이 전혀 필요 없다. (사실 지금까지의 과정이 주어지더라도 X_r는 언제나 $p = \frac{n-r}{n}$인 기하분포를 따르므로, 첫 부분의 결과를 각 X_r에 적용할 수 있다.) 그러므로 $k = n - r$로 놓으면

$$\mathrm{E}(T) = \sum_{r=0}^{n-1} \frac{n}{n-r} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{1}\right) = n\sum_{k=1}^{n} \frac{1}{k},$$

이다.

$H_n = \sum_{k=1}^{n} \frac{1}{k}$로 쓰자. 부등식의 두 부분으로부터

$$H_n = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \ln n, \qquad H_n = \sum_{k=1}^{n-1} \frac{1}{k} + \frac{1}{n} \geq \ln n + \frac{1}{n}.$$

을 얻는다. 따라서

$$n\ln n + 1 \leq \mathrm{E}(T) \leq n\ln n + n.$$

이다. 두 경계의 차는 n보다 작은데, n이 크면 이는 $n\ln n$에 비해 아주 작다(두 값의 비 $1/\ln n \to 0$). 그러므로 n이 크면 대략

$$n\ln n \text{ 봉지를 먹을 것으로 기대해야 한다.}$$

예를 들어 수학자가 100명이면 Cruncho를 약 460봉지 먹어야 한다.

검산

- $r = 0$: 첫 번째 봉지는 언제나 새 카드이고, $n/(n-0) = 1$이다. $r = n-1$: 특정한 카드 한 장만 빠져 있고 이 카드는 확률 $\frac{1}{n}$으로 나오므로 n봉지를 기대하게 되며, 이는 $n/1$과 일치한다.
- $n = 2$: $\mathrm{E}(T) = 2(1+\frac{1}{2}) = 3$이고, 경계는 $2\ln 2 + 1 \approx 2.39$와 $2\ln 2 + 2 \approx 3.39$이다.
- $n = 20$으로 전체 수집을 10^5번 시뮬레이션한 결과 평균 71.9봉지가 나왔는데, 이론값은 $20H_{20} \approx 71.95$이다. 경계는 60.9와 79.9이다.

흔한 실수

- 적분 부등식을 $r = 0$부터 더하는 것: 그곳에서는 경계 $\frac{1}{r}$이 성립하지 않고, $\int_0^1 \frac{1}{x}\,\mathrm{d}x$는 발산한다. 합은 $r = 1$부터 시작해야 한다.
- 기댓값을 더하려면 X_r들이 독립이어야 한다고 생각하는 것. 선형성에는 독립성이 필요 없다.
- 답을 $\ln n$이나 n으로 내는 것: 기대되는 봉지 수는 $n\ln n$처럼 증가하며, n보다 빠르게 커진다.

STEP 너머

차 $H_n - \ln n$은 감소하면서 어떤 극한, 즉 오일러-마스케로니 상수(Euler–Mascheroni constant) $\gamma \approx 0.5772$에 수렴(converge)한다. 따라서 더 정밀한 답은 $\mathrm{E}(T) \approx n(\ln n + \gamma)$이고, $n = 100$이면 460봉지가 아니라 약 519봉지가 된다. 이처럼 합과 적분을 비교하는 것이 급수의 적분판정법(integral test)이며, 대학의 첫 해석학(analysis) 강의에서 일찍 배운다. 쿠폰 수집가 문제 자체도 알고리즘 분석(analysis of algorithms)과, 카드 한 벌을 무작위로 몇 번 섞어야 충분히 섞이는지에 관한 연구에서 다시 등장한다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 기하분포를 따르는 대기 시간과 기댓값의 선형성이 함께 쓰이는 모습을 보기에 좋은 책이다.

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.