풀이

# STEP III 1997, 문제 12

확률과 통계 • 쿠폰 수집가 문제

**STEP III 1997, 12번 문제**

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

step.maths.org/questions/97-s3-q12

## 핵심 아이디어

소문항 (i)은 기하분포(geometric distribution)이고, 그 확률생성함수(probability generating function)를 미분하면 평균과 분산(variance)을 얻는다: $\mathrm{E}(K) = G'(1)$이고 $\mathrm{E}(K(K-1)) = G''(1)$이다. 소문항 (ii)에서 핵심 요령은 카드 한 벌을 모두 모을 때까지의 긴 기다림을 52개의 짧은 기다림으로 나누는 것이다. 각 기다림은 **새로운** 카드 한 장을 얻을 때까지의 기다림이다. 짧은 기다림 하나하나가 "첫 앞면이 나올 때까지 던지기" 실험이므로, 소문항 (i)이 모든 일을 해 준다.

## 풀이

### 소문항 (i)

첫 앞면이 $k$번째 던지기에서 나오는 것은 처음 $k-1$번이 모두 뒷면이고 $k$번째가 앞면일 때, 그리고 그때에만이다. 각 던지기는 독립(independent)이므로

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

이다. 따라서

$$G(s) = \sum_{k=1}^{\infty} pq^{k-1}s^k = ps\sum_{j=0}^{\infty}(qs)^j = \frac{ps}{1-qs},$$

이다. 여기서 등비급수(geometric series)를 합했는데, 이 급수는 $|qs| < 1$일 때 수렴한다(특히 $p > 0$이면 1 근처의 $s$에 대해 수렴한다). 이것이 구하려던 결과이다.

미분하면

$$\mathrm{G}'(s) = \frac{p(1-qs)+psq}{(1-qs)^2} = \frac{p}{(1-qs)^2},$$
$$\mathrm{G}''(s) = \frac{2pq}{(1-qs)^3}.$$

이다. $\mathrm{G}'(s) = \sum k\,\mathrm{P}(K=k)s^{k-1}$이고 $\mathrm{G}''(s) = \sum k(k-1)\mathrm{P}(K=k)s^{k-2}$이므로, $s=1$을 대입하고 $1-q=p$를 이용하면

$$\mathrm{E}(K) = \mathrm{G}'(1) = \frac{p}{p^2} = \frac{1}{p}, \qquad \mathrm{E}(K(K-1)) = \mathrm{G}''(1) = \frac{2pq}{p^3} = \frac{2q}{p^2}.$$

을 얻는다. 그러므로

$$\begin{aligned}\mathrm{Var}(K) &= \mathrm{E}(K(K-1)) + \mathrm{E}(K) - (\mathrm{E}(K))^2 \\ &= \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q+p-1}{p^2} = \frac{q}{p^2}.\end{aligned}$$

이다. 즉 $\mathrm{E}(K) = \frac{1}{p}$이고 $\mathrm{Var}(K) = \frac{q}{p^2}$이다.

**소문항 (ii)**

서로 다른 카드를 정확히 $i-1$장 본 뒤부터, $i$번째 새로운 카드가 나오는 뽑기까지(그 뽑기를 포함하여) 뽑은 횟수를 $N_i$라 하자. 그러면 $N_1 = 1$이고(첫 카드는 언제나 새로운 카드이다),

$$N = N_1 + N_2 + \cdots + N_{52}.$$

이다.

? **생각해 보기** 왜 각 $N_i$는 $K$와 같은 "첫 앞면" 확률변수인가?

서로 다른 카드를 $i-1$장 보았다면, 52장 가운데 $53-i$장은 아직 보지 못한 카드이다. 복원추출(drawing with replacement)을 하므로 각 뽑기는 과거와 독립이고, 언제나 같은 확률

$$p_i = \frac{53-i}{52}.$$

로 아직 보지 못한 카드가 나온다. 이런 뽑기를 "앞면"이라 부르자. 그러면 $N_i$는 첫 앞면이 나올 때까지(그 뽑기를 포함하여) 뽑은 횟수이며, 이는 정확히 $p = p_i$인 소문항 (i)의 확률변수 $K$이다. 따라서

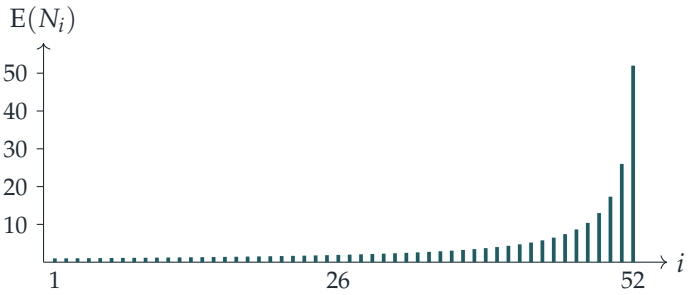
$$\mathrm{E}(N_i) = \frac{1}{p_i} = \frac{52}{53-i}.$$

이다.

기댓값(expectation)은 선형적이므로(이를 위해 $N_i$들이 독립일 필요는 없다)

$$\begin{aligned}\mathrm{E}(N) &= \sum_{i=1}^{52} \frac{52}{53-i} = 52\left(\frac{1}{52}+\frac{1}{51}+\cdots+\frac{1}{2}+\frac{1}{1}\right)\\ &= 52\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{52}\right).\end{aligned}$$

이다. 초반의 기다림은 짧고 마지막 몇 번의 기다림이 대부분을 차지한다. 마지막 카드 한 장만 해도 평균 52번을 뽑아야 한다.




어림하기 위해 $n = 52$로 두고 $1+\frac{1}{2}+\cdots+\frac{1}{n} \approx 0.5+\ln n$을 이용한다. $\ln 52$에 대해서는 $2.7^2 = 7.29$이고 $7.29^2 \approx 53.1$이므로 $\mathrm{e}^4 \approx 53 \approx 52$이고, 따라서 $\ln 52 \approx 4$이다. 그러므로

$$\mathrm{E}(N) \approx 52\,(0.5+4) = 52 \times 4.5 = 234.$$

이다. 즉 모든 카드를 보려면 평균적으로 대략 234번, 카드 한 벌 크기의 약 네 배 반을 뽑아야 한다.

## 검산

- (i)에서 $\mathrm{G}(1) = p/(1-q) = 1$인데, 모든 확률생성함수는 마땅히 그래야 한다. 공정한 동전($p = \frac{1}{2}$)이면 $\mathrm{E}(K) = 2$이고 $\mathrm{Var}(K) = 2$이다. 또 $p \to 1$이면 분산 $q/p^2 \to 0$인데, 첫 던지기가 거의 확실히 앞면일 때 당연히 그래야 한다.
- 정확한 값은 $52H_{52} \approx 235.98$이고, 카드 한 벌 모으기를 $10^5$번 모의실험한 결과 평균은 235.7이었다. 우리의 어림값 234는 이에 가깝다. 약간 모자란 것은 근사 $0.5+\ln n$(더 정밀한 근사는 $\ln n + 0.577$이다) 때문이고, 또 $\ln 52 \approx 3.95$를 4로 올림했기 때문이다.

## 흔한 실수

- $N_i$의 성공 확률을 거꾸로 잡는 실수: **새로운** 카드가 나올 확률은 $(53-i)/52$이며, 모은 카드가 늘어날수록 줄어든다.

- $+G'(1)$ 항을 잊고 $\mathrm{Var}(K) = G''(1) - G'(1)^2$로 쓰는 실수: $G''(1)$은 $\mathrm{E}(K^2)$가 아니라 $\mathrm{E}(K(K-1))$이다.
- 기댓값을 더하기 전에 $N_i$들이 독립이어야 한다고 생각하는 실수: 기댓값의 선형성(linearity of expectation)은 독립 여부와 관계없이 성립한다(다만 여기서는 실제로 독립이기도 하다).

## STEP 너머

이것은 **쿠폰 수집가 문제**(coupon collector's problem)이다. 똑같이 나올 법한 종류가 $n$가지 있을 때, 모든 종류를 모으는 데 걸리는 시간의 기댓값은 $nH_n \approx n\ln n$이다. 대학 학부 과정에서는 더 나아가 이 시간이 $n\ln n$ 근처에 날카롭게 집중되어 있으며 그 요동은 $n$ 정도의 크기임을 보이고, 확률생성함수는 이와 같은 독립 확률변수들의 합을 다루는 표준 도구가 된다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 생성함수와 기하분포를 꼼꼼히 다루고, Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 쿠폰 수집가 문제를 여러 변형과 함께 논의한다.

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.