풀이

STEP III 2000, 문제 13

확률과 통계 • 확률생성함수

> **STEP III 2000, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/00-s3-q13

핵심 아이디어

두 확률변수(random variable) 모두 "꼬리(tail)" 사건 또는 "누적(cumulative)" 사건을 통해 다루는 것이 가장 쉽다. 첫 번째 6이 r번째 던지기 이후에 나온다는 것은 처음 r번의 던지기에서 **모든** 주사위가 매번 6이 아닌 눈을 냈다는 것과 정확히 같다. 마지막 6이 r번째 던지기까지 이미 나왔다는 것은 **각** 주사위가 이미 6을 한 번 이상 냈다는 것과 정확히 같다. 이 사건들은 각각 독립(independent)인 사건들의 교집합이므로 그 확률은 곱으로 주어지고, 분포는 연속한 값끼리 빼서 얻는다. 첫 번째 6에 대한 확률변수가 기하분포(geometric distribution)를 따른다는 것을 알아차리면, 그에 관한 모든 것은 금방 구할 수 있다.

풀이

전체에 걸쳐 주사위들은 서로 독립이고, 던질 때마다도 독립이라고 가정한다. 주사위 하나에 대하여, 주사위 i가 처음으로 6을 내는 던지기의 번호를 X_i라 하면

$$\mathrm{P}(X_i > r) = q^r \qquad (r = 0, 1, 2, \dots),$$

이다. $X_i > r$는 연속으로 r번 6이 아닌 눈이 나왔다는 뜻이기 때문이다. 그러면 $R = \min_i X_i$이고 $S = \max_i X_i$이다.

첫 번째 6

사건 $R > r$는 처음 r번의 던지기에서 n개의 주사위 중 어느 것도 6을 내지 않았다는 것, 즉 nr번의 개별 던지기가 모두 6이 아니었다는 것을 뜻한다. 독립성에 의하여

$$\mathrm{P}(R > r) = q^{nr}.$$

따라서 $r \geq 1$일 때

$$\mathrm{P}(R = r) = \mathrm{P}(R > r-1) - \mathrm{P}(R > r) = q^{n(r-1)} - q^{nr} = q^{nr}(q^{-n} - 1),$$

이고, 이것이 구하려던 결과이다.

$Q = q^n$이라 쓰면 이것은 $\mathrm{P}(R = r) = Q^{r-1}(1-Q)$이다. 즉 R는 기하분포를 따르며, 주사위 전체를 한 번 던지는 것이 확률 $1-Q$로 "성공"(적어도 하나의 6)이 된다. 따라서 확률생성함수(probability generating function)는 기하급수(geometric series)이다.

$$\begin{aligned} G_R(t) = \sum_{r=1}^{\infty} t^r Q^{r-1}(1-Q) &= (1-Q)\, t \sum_{r=1}^{\infty} (Qt)^{r-1} \\ &= \frac{(1-q^n)\, t}{1-q^n t}, \end{aligned}$$

이 식은 $|t| < q^{-n}$에서 (특히 $t = 1$ 근처에서) 성립한다.

평균(mean)을 구하기 위해 몫의 미분법(quotient rule)으로 미분하면

$$G_R'(t) = (1-q^n)\, \frac{(1-q^n t) + q^n t}{(1-q^n t)^2} = \frac{1-q^n}{(1-q^n t)^2},$$

이므로

$$\mathrm{E}(R) = G_R'(1) = \frac{1}{1-q^n}.$$

$n = 2$, $p = \frac{1}{6}$이면 $q^2 = \frac{25}{36}$이므로 구하려던 대로 $\mathrm{E}(R) = 1/\frac{11}{36} = \frac{36}{11}$이다.

마지막 6

사건 $S \leq r$는 처음 r번의 던지기 중 어딘가에서 모든 주사위가 6을 냈다는 것을 뜻한다. 주사위 하나에 대해 이 확률은 $1-q^r$이고, 주사위들은 독립이므로

$$\mathrm{P}(S \leq r) = (1-q^r)^n.$$

연속한 값끼리 빼면 $r \geq 1$일 때

$$\mathrm{P}(S = r) = \mathrm{P}(S \leq r) - \mathrm{P}(S \leq r-1) = (1-q^r)^n - (1-q^{r-1})^n,$$

이고, 이것이 구하려던 결과이다. ($r = 1$일 때 둘째 항은 $0^n = 0$으로, 올바르게 된다.)

> **생각해 보기** $t^r[(1-q^r)^2 - (1-q^{r-1})^2]$를 더하는 것은 복잡해 보인다. 어떻게 하면 이것을 기하급수로 바꿀 수 있을까?

$n = 2$일 때 제곱을 전개하면

$$\begin{aligned} \mathrm{P}(S = r) &= (1 - 2q^r + q^{2r}) - (1 - 2q^{r-1} + q^{2r-2}) \\ &= 2q^{r-1}(1-q) - q^{2(r-1)}(1-q^2) \\ &= 2p\, q^{r-1} - (1-q^2)\, (q^2)^{r-1}. \end{aligned}$$

각 항은 기하분포(의 상수배)이므로

$$G_S(t) = \sum_{r=1}^{\infty} t^r \,\mathrm{P}(S=r) = \frac{2pt}{1-qt} - \frac{(1-q^2)\,t}{1-q^2t}.$$

이것은 이치에 맞는다. $S = X_1 + X_2 - R$이므로 그 분포는 "모수 p인 기하분포의 두 배에서 R의 분포를 뺀 것"이기 때문이다. 간단히 하기 위해 모든 것을 $(1-qt)(1-q^2t)$로 통분하고 $1-q^2 = p(1+q)$를 이용하면

$$\begin{aligned} 2p(1-q^2t) - p(1+q)(1-qt) &= p[\,2 - 2q^2t - 1 - q + qt + q^2t\,] \\ &= p[(1-q) + qt(1-q)] = p^2(1+qt). \end{aligned}$$

따라서

$$G_S(t) = \frac{p^2\,t\,(1+qt)}{(1-qt)(1-q^2t)}.$$

간단한 확인으로, $G_S(1) = \dfrac{p^2(1+q)}{p\cdot p(1+q)} = 1$이다.

평균을 구할 때는 나누어 쓴 꼴을 미분하는 것이 가장 쉽다. 각 조각은 기하분포의 확률생성함수(PGF)이고, 그 1에서의 도함수 값은 평균에 계수를 곱한 것이기 때문이다. $\dfrac{(1-a)t}{1-at}$의 $t=1$에서의 도함수 값은 (위의 G_R와 꼭 마찬가지로) $\dfrac{1}{1-a}$이므로

$$\mathrm{E}(S) = G_S'(1) = 2\cdot\frac{1}{p} - \frac{1}{1-q^2} = \frac{2(1+q)-1}{p(1+q)} = \frac{1+2q}{p(1+q)}.$$

$p = \frac{1}{6}, q = \frac{5}{6}$이면

$$\mathrm{E}(S) = 12 - \frac{36}{11} = \frac{96}{11} \approx 8.73.$$

검산

- $\mathrm{E}(R) + \mathrm{E}(S) = \frac{36}{11} + \frac{96}{11} = 12 = \mathrm{E}(X_1) + \mathrm{E}(X_2)$이다. $\min + \max = X_1 + X_2$이므로 당연히 그래야 한다.
- $n = 1$이면 첫 번째 6과 마지막 6이 일치하고, G_R는 주사위 하나의 기하분포 PGF인 $pt/(1-qt)$가 된다.
- 주사위 쌍 2×10^5개에 대한 몬테카를로 시뮬레이션(Monte Carlo simulation)에서 표본평균(sample mean)은 3.274와 8.709로, $\frac{36}{11} \approx 3.273$ 및 $\frac{96}{11} \approx 8.727$과 비교된다. 또 G_S의 급수를 수치적으로 더한 값은 여러 t 값에서 닫힌 꼴과 일치한다.

흔한 실수

- $\mathrm{P}(S = r)$를 "어떤 주사위가 r번째 던지기에서 처음 6을 내고 나머지는 그 전에 냈다"로 쓰면서, 두 주사위가 같은 던지기에서 처음 6을 내는 경우를 빠뜨리는 것. $\mathrm{P}(S \leq r)$로 계산하면 이 문제를 완전히 피할 수 있다.
- 나누어 쓴 꼴 대신 간단히 정리한 G_S를 미분하는 것. 틀리지는 않지만 계산이 훨씬 많아지고 실수할 여지도 많다.

STEP 너머

R와 S는 독립인 확률변수들의 최솟값과 최댓값으로, **순서통계량**(*order statistics*)의 가장 간단한 예이다. $\mathrm{P}(\min > r)$와 $\mathrm{P}(\max \leq r)$를 이용하는 기법은 어떤 분포에도 통한다. 무기억성(memoryless)을 가진 기하분포는 지수분포(exponential distribution)의 이산 사촌이며, 연속 극한에서 n개의 지수분포 확률변수의 최댓값의 기댓값은 조화합(harmonic sum) $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ (에 평균을 곱한 것)이 된다. 이는 쿠폰 수집가 문제(coupon collector's problem)의 핵심 단계이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 생성함수와 이러한 아이디어들을 깊이 있게 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.