풀이

STEP III 2009, 문제 12

확률과 통계 • 확률생성함수

> **STEP III 2009, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/09-s3-q12

핵심 아이디어

소문항 (i)에서는 한 번 던질 때마다 동전 개수의 기댓값(expected value)이 절반으로 줄어든다. 그러나 던지는 횟수 자체가 확률변수이므로 그 값에 대해 조건부(conditioning)로 생각한다. 이전 개수가 주어지면 다음 개수는 이항분포(binomial distribution)를 따르고, 그 평균은 쉽게 구할 수 있다. 이렇게 하면 기댓값들이 등비수열(geometric sequence)을 이룬다. 소문항 (ii)에서는 합계가 서로 독립(independent)이고 같은 분포를 따르는 조각들의 합이다. 확률생성함수(probability generating function)를 쓰면 "독립인 확률변수들의 합"이 "생성함수들의 곱"으로 바뀐다.

풀이

소문항 (i)

각 시행에서 앞면이 나올 확률은 $\frac{1}{2}$이고 시행들은 서로 독립이므로 $X_1 \sim \mathrm{B}(k, \frac{1}{2})$이고

$$\mathrm{E}(X_1) = \tfrac{1}{2}k.$$

이다.

$X_1 = x_1$이 주어지면 Albert는 새로 서로 독립인 x_1번의 시행을 하므로 $X_2 \mid X_1 = x_1 \sim \mathrm{B}(x_1, \frac{1}{2})$이고

$$\mathrm{E}(X_2 \mid X_1 = x_1) = \tfrac{1}{2}x_1.$$

이다. ($x_1 = 0$이면 Albert는 동전을 던지지 않고 $X_2 = 0$이므로 이 식과 일치한다.) 전체 기

댓값의 법칙(law of total expectation)에 따라 X_1이 가질 수 있는 값들에 대해 평균을 내면

$$\mathrm{E}(X_2) = \sum_{x_1=0}^{k} \mathrm{E}(X_2 \mid X_1 = x_1)\,\mathrm{P}(X_1 = x_1) = \tfrac{1}{2}\sum_{x_1=0}^{k} x_1\,\mathrm{P}(X_1 = x_1) = \tfrac{1}{2}\,\mathrm{E}(X_1) = \tfrac{1}{4}k,$$

이므로 요구한 결과를 얻는다.

? **생각해 보기** 왜 그냥 "X_2는 $\mathrm{B}(\frac{1}{2}k, \frac{1}{2})$를 따른다"고 말할 수 없을까?

두 번째 단계에서 던지는 횟수는 수 $\frac{1}{2}k$가 아니라 확률변수 X_1이기 때문이다. 실제로 $\frac{1}{2}k$는 정수가 아닐 수도 있다. X_1에 대해 조건을 걸어야 던지는 횟수가 고정되고, 그때에야 비로소 이항분포의 평균을 쓸 수 있다.

모든 단계에서 똑같은 논증이 성립한다. $X_i = x_i$가 주어지면 $X_{i+1} \sim \mathrm{B}(x_i, \frac{1}{2})$이므로 $\mathrm{E}(X_{i+1}) = \frac{1}{2}\mathrm{E}(X_i)$이다. 수학적 귀납법(induction)에 의해

$$\mathrm{E}(X_i) = \frac{k}{2^i} \qquad (i \geq 1),$$

이고, 공비가 $\frac{1}{2}$인 등비급수(geometric series)를 더하면

$$\sum_{i=1}^{\infty} \mathrm{E}(X_i) = k\left(\tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \cdots\right) = k.$$

이다.

소문항 (ii)

동전 하나에 대해 $Y_i = n$은 앞면이 n번 나온 뒤 뒷면이 나온다는 뜻이므로

$$\mathrm{P}(Y_i = n) = \left(\tfrac{1}{2}\right)^n \cdot \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^{n+1}, \qquad n = 0, 1, 2, \ldots$$

이다. 이 분포의 확률생성함수는 $|t| < 2$에서 수렴(convergent)하는 등비급수이다.

$$G_{Y_i}(t) = \sum_{n=0}^{\infty} \left(\tfrac{1}{2}\right)^{n+1} t^n = \frac{\frac{1}{2}}{1 - \frac{1}{2}t} = \frac{1}{2 - t}.$$

k개의 동전은 따로따로 던지므로 $Y_1, \ldots, Y_k$는 서로 독립이고, 독립인 확률변수들의 합의 생성함수는 각 생성함수의 곱이다. 따라서

$$G_Y(t) = \mathrm{E}(t^{Y_1 + \cdots + Y_k}) = \prod_{i=1}^{k} \mathrm{E}(t^{Y_i}) = (2 - t)^{-k}.$$

이다.

평균과 분산. 미분하면

$$G_Y'(t) = k(2-t)^{-k-1}, \qquad G_Y''(t) = k(k+1)(2-t)^{-k-2},$$

이므로 $G_Y'(1) = k$, $G_Y''(1) = k(k+1)$이다. 따라서

$$\mathrm{E}(Y) = G_Y'(1) = k,$$

이고, $G_Y''(1) = \mathrm{E}(Y(Y-1))$이므로

$$\mathrm{Var}(Y) = G_Y''(1) + G_Y'(1) - (G_Y'(1))^2 = k(k+1) + k - k^2 = 2k.$$

이다.

분포. $\mathrm{P}(Y=r)$는 $G_Y(t)$에서 t^r의 계수이다.

$$G_Y(t) = 2^{-k}\left(1 - \tfrac{1}{2}t\right)^{-k},$$

로 쓰고, 음의 지수에 대한 이항급수(binomial series)를 이용한다($|t| < 2$에서 성립).

$$(1-x)^{-k} = \sum_{r=0}^{\infty} \binom{k+r-1}{r} x^r.$$

(계수는 $\dfrac{k(k+1)\cdots(k+r-1)}{r!}$이고, 이는 $\binom{k+r-1}{r}$과 같다.) $x = \frac{1}{2}t$를 대입하면

$$\mathrm{P}(Y=r) = \binom{k+r-1}{r}\frac{1}{2^{k+r}}, \qquad r = 0, 1, 2, \ldots$$

이다.

? 생각해 보기 이 공식을 직접 알아볼 방법이 있을까?

있다. 모든 시행을 한 줄로 길게 늘어놓았다고 생각하자. $Y = r$은 k번째 뒷면이 $k+r$번째 시행에서 나온다는 뜻이다. $k+r$번의 시행으로 이루어진 특정한 결과열 하나의 확률은 $2^{-(k+r)}$이고, 마지막 시행은 반드시 뒷면이어야 하므로 처음 $k+r-1$번의 시행 중 어느 r번이 앞면인지만 고르면 된다. 그 방법의 수는 $\binom{k+r-1}{r}$가지이다. 이것이 음이항분포(negative binomial distribution)이다.

검산

- $k = 1$이면 $\mathrm{P}(Y=r) = 2^{-(r+1)}$로 동전 하나의 기하분포(geometric distribution)이고, $\mathrm{E}(Y) = 1$, $\mathrm{Var}(Y) = 2$는 0부터 세는 $p = \frac{1}{2}$인 기하분포 확률변수의 익숙한 값이다.
- 모든 생성함수가 만족해야 하듯이 $G_Y(1) = 1$이다.

- $k = 2$이면 $\mathrm{P}(Y = 0) = \frac{1}{4}$(두 동전 모두 첫 시행에서 뒷면)이고 $\mathrm{P}(Y = 1) = \binom{2}{1}\frac{1}{8} = \frac{1}{4}$(한 동전은 앞면 다음 뒷면, 다른 동전은 뒷면: 두 가지 경우가 있고 각각 $\frac{1}{8}$)이다.
- 분수를 이용한 정확한 계산으로 $k \leq 7, i < 40$에 대해 $\mathrm{E}(X_i) = k/2^i$임을 확인했고(따라서 부분합이 k에 가까워진다), $k \leq 6$에 대해 $\mathrm{P}(Y = r)$의 공식, 평균 k, 분산 $2k$를 확인했다.

흔한 실수

- 마지막 뒷면을 잊고 $\mathrm{P}(Y_i = n) = (\frac{1}{2})^n$으로 놓는 것. 이렇게 하면 확률의 합이 2가 된다.
- $\mathrm{Var}(Y) = G''(1) - \big(G'(1)\big)^2$을 쓰는 것. 이는 $G'(1)$ 항을 빠뜨려 $2k$ 대신 k를 준다.
- 계수를 $\binom{k+r}{r}$로 쓰는 것. 마지막 시행은 뒷면으로 고정되어 있으므로 자유로운 시행은 $k + r - 1$번뿐이다.

STEP 너머

소문항 (i)은 분기 과정(branching process)이다. 각 "개체"(앞면)는 다음 세대에 무작위 개수의 "자손"을 남기는데, 여기서는 그 개수가 $\mathrm{B}(1, \frac{1}{2})$를 따른다. 자손 수의 평균이 1보다 작으므로 개체군은 결국 소멸한다. 조건부 단계 $\mathrm{E}(X_{i+1}) = \mathrm{E}\big(\mathrm{E}(X_{i+1} \mid X_i)\big)$는 탑 성질(tower property)이며, 대학 1학년 과정에서는 X_i의 생성함수가 자손 생성함수를 i번 합성한 것임을 보인다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 생성함수와 분기 과정을 모두 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.