풀이

STEP III 2011, 문제 12

확률과 통계 • 확률생성함수를 이용한 복합분포

> **STEP III 2011, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/11-s3-q12

핵심 아이디어

분포의 평균과 분산에 관해 필요한 모든 정보는 확률생성함수(probability generating function, pgf)의 $t = 1$에서의 처음 두 도함수(derivative)에 담겨 있다. 따라서 합성함수 $G(H(t))$를 연쇄법칙(chain rule)으로 (두 번) 미분하고, $t = 1$을 대입한 뒤 $H(1) = 1$을 이용한다. 동전 문제에서는 N이 기하분포를 따르고 각 X_i는 공정한 동전 한 번에 대한 지시변수(indicator)이므로, 두 확률생성함수가 모두 간단하며 $G(H(t))$는 등비급수(geometric series)로 전개할 수 있는 유리함수가 된다.

풀이

일반적인 결과

음이 아닌 정수 값을 가지고 확률생성함수가 $P(t) = \mathrm{E}(t^Z)$인 임의의 확률변수 Z에 대하여 $P(1) = 1$,

$$P'(1) = \mathrm{E}(Z), \qquad P''(1) = \mathrm{E}(Z(Z-1)),$$

이고, 따라서

$$\mathrm{Var}(Z) = P''(1) + P'(1) - (P'(1))^2.$$

이다. $\mu = \mathrm{E}(X_i) = H'(1)$, $\sigma^2 = \mathrm{Var}(X_i)$라 하고, Y의 확률생성함수를 $K(t) = G(H(t))$라 하자.

연쇄법칙에 의해 $K'(t) = G'(H(t))\,H'(t)$이다. $H(1) = 1$이므로

$$\mathrm{E}(Y) = K'(1) = G'(1)H'(1) = \mathrm{E}(N)\,\mathrm{E}(X_i),$$

이고, 이것이 첫 번째 결과이다.

한 번 더 미분하면 (곱의 미분법을 쓴 뒤 연쇄법칙을 쓴다)

$$K''(t) = G''(H(t))(H'(t))^2 + G'(H(t))\,H''(t),$$

이므로, 다시 $H(1) = 1$을 이용하면

$$K''(1) = G''(1)\mu^2 + \mathrm{E}(N)\,H''(1).$$

이다. 위의 분산 공식을 N과 X_i에 적용하면

$$G''(1) = \mathrm{Var}(N) - \mathrm{E}(N) + (\mathrm{E}(N))^2, \qquad H''(1) = \sigma^2 - \mu + \mu^2.$$

이다.

생각해 보기 $G''(1)$과 $H''(1)$을 다시 쓰지 않고 $K''(1)$에서 곧바로 $\mathrm{Var}(Y)$를 계산하면 안 되는 이유는 무엇일까?

우리가 원하는 답은 계승 모멘트(factorial moment)가 아니라 분산으로 표현된 것이기 때문이다. $G''(1)$과 $H''(1)$을 $\mathrm{Var}(N)$과 σ^2로 나타내면 불필요한 항들이 소거된다. $m = \mathrm{E}(N)$이라 쓰면

$$\begin{aligned}\mathrm{Var}(Y) &= K''(1) + K'(1) - (K'(1))^2 \\ &= (\mathrm{Var}(N) - m + m^2)\mu^2 + m(\sigma^2 - \mu + \mu^2) + m\mu - m^2\mu^2 \\ &= \mathrm{Var}(N)\mu^2 + m\sigma^2 + (-m\mu^2 + m\mu^2) + (m^2\mu^2 - m^2\mu^2) + (-m\mu + m\mu) \\ &= \mathrm{Var}(N)(\mathrm{E}(X_i))^2 + \mathrm{E}(N)\,\mathrm{Var}(X_i),\end{aligned}$$

이 되어 원하는 결과를 얻는다.

동전

N의 확률생성함수. 첫 번째 앞면이 n번째 던지기에서 나오려면 처음 $n-1$번은 뒷면이 나오고 n번째에 앞면이 나와야 하므로, $n = 1, 2, 3, \ldots$에 대하여 $\mathrm{P}(N = n) = (\frac{1}{2})^n$이다(기하분포(geometric distribution)). 따라서 $|t| < 2$일 때

$$G(t) = \sum_{n=1}^{\infty} \left(\frac{t}{2}\right)^n = \frac{t/2}{1 - t/2} = \frac{t}{2 - t}.$$

이다.

각 X_i의 확률생성함수. 추가로 던지는 것 중 i번째에서 앞면이 나오면 $X_i = 1$, 그렇지 않으면 $X_i = 0$이라 하자. 그러면 $Y = X_1 + \cdots + X_N$이고 $H(t) = \frac{1}{2} + \frac{1}{2}t = \frac{1}{2}(1 + t)$이다. 추가로 던지는 것들은 서로 독립이고 N을 결정한 던지기들과도 독립이므로, 일반적인 결과의 조건이 성립한다.

Y의 확률생성함수.

$$G(H(t)) = \frac{\frac{1}{2}(1+t)}{2-\frac{1}{2}(1+t)} = \frac{1+t}{3-t}.$$

평균과 분산. $p = \frac{1}{2}$인 기하분포에서 $\mathrm{E}(N) = 1/p = 2$이고 $\mathrm{Var}(N) = (1-p)/p^2 = 2$이다(또는 G를 미분하면 $G'(t) = 2/(2-t)^2$, $G''(t) = 4/(2-t)^3$이므로 $G'(1) = 2$, $G''(1) = 4$이고 $\mathrm{Var}(N) = 4 + 2 - 4 = 2$이다). 각 X_i는 베르누이$(\frac{1}{2})$ 분포를 따르므로 $\mathrm{E}(X_i) = \frac{1}{2}$이고 $\mathrm{Var}(X_i) = \frac{1}{4}$이다. 따라서

$$\mathrm{E}(Y) = 2 \cdot \tfrac{1}{2} = 1, \qquad \mathrm{Var}(Y) = 2 \cdot \tfrac{1}{4} + 2 \cdot \tfrac{1}{4} = 1.$$

이다.

Y의 분포. $|t| < 3$일 때 다음과 같이 전개한다.

$$\frac{1+t}{3-t} = \frac{1+t}{3} \cdot \frac{1}{1-t/3} = \frac{1+t}{3} \sum_{k=0}^{\infty} \frac{t^k}{3^k} = \sum_{k=0}^{\infty} \frac{t^k}{3^{k+1}} + \sum_{k=0}^{\infty} \frac{t^{k+1}}{3^{k+1}}.$$

t^0의 계수는 첫 번째 합에서만 나온다. $r \geq 1$일 때 t^r의 계수는 $\dfrac{1}{3^{r+1}} + \dfrac{1}{3^r} = \dfrac{4}{3^{r+1}}$이다. 따라서

$$\mathrm{P}(Y = 0) = \frac{1}{3}, \qquad \mathrm{P}(Y = r) = \frac{4}{3^{r+1}} \quad (r = 1, 2, 3, \ldots).$$

이다.

검산

- 확률의 합이 1이다: $\frac{1}{3} + 4\sum_{r\geq 1} 3^{-(r+1)} = \frac{1}{3} + 4 \cdot \frac{1}{6} = 1$.
- $\mathrm{P}(Y = 0)$를 직접 구하면 $\sum_{n\geq 1} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \sum_{n\geq 1} 4^{-n} = \frac{1}{3}$이다. 마찬가지로 $\mathrm{P}(Y = 1) = \sum_{n\geq 1} n\, 4^{-n} = \frac{4}{9}$이고, 이는 $4/3^2$와 일치한다.
- $n \leq 200$에 대하여 $\mathrm{P}(N = n)\binom{n}{r}2^{-n}$을 정확히 더하면 $r \leq 7$에 대한 $\mathrm{P}(Y = r)$가 재현되었고, 평균과 분산이 모두 1.000으로 나왔다. 200 000회 시뮬레이션에서는 0.996과 0.994가 나왔다.

흔한 실수

- N이 0부터 시작한다고 보는 것(확률생성함수 $1/(2-t)$): 여기서 N은 앞면이 나온 던지기까지 세므로 $N \geq 1$이다.
- $P''(1)$이 $\mathrm{E}(Z^2)$가 아니라 $\mathrm{E}(Z^2) - \mathrm{E}(Z)$라는 것을 잊어서, 분산 공식의 $+P'(1)$을 빠뜨리는 것.
- 모든 $r \geq 0$에 대하여 $4/3^{r+1}$이라고 답하는 것: $r = 0$인 항은 다르다.

STEP 너머

Y는 **무작위 합**(random sum, 복합 분포(compound distribution))이며, 분산 공식은 전분산 법칙(law of total variance) $\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid N)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid N)\big)$의 특수한 경우이다. 확률생성함수의 합성은 분기 과정(branching process)의 원동력이기도 하다. 분기 과정에서는 $G(G(\cdots G(t)))$가 n번째 세대의 크기를 나타내고, 소멸 확률은 G의 고정점(fixed point)이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 생성함수, 무작위 합, 분기 과정을 정확히 이 수준에서 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.