풀이

STEP III 2008, 문제 12

확률과 통계 • 적률생성함수

STEP III 2008, 12번 문제

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

step.maths.org/questions/08-s3-q12

이 문제는 *Beyond STEP*의 12.24 절 **중심극한정리**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch12/clt.html

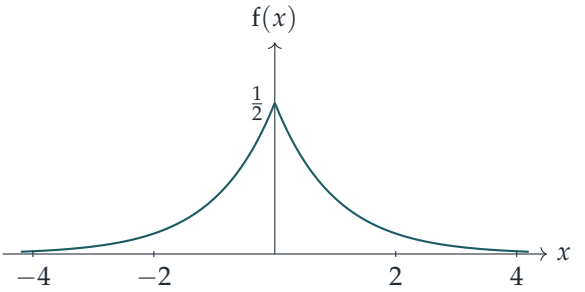
핵심 아이디어

확률밀도함수(density)에 있는 절댓값(absolute value) 때문에 모든 적분은 $x < 0$과 $x > 0$의 두 부분으로 나뉘고, 각 부분은 단순한 지수함수의 적분이 된다. 그다음부터 문제 전체는 적률생성함수(moment generating function, m.g.f.)에 관한 두 가지 사실로 풀린다. 하나는 독립(independent)인 확률변수들의 합의 적률생성함수가 각 적률생성함수의 곱이라는 것이고, 다른 하나는 $\mathrm{M}_{aY}(\theta) = \mathrm{M}_{Y}(a\theta)$라는 것이다. 로그를 취하면 n제곱이 작은 값의 n배로 바뀌는데, 이것이 바로 매클로린 급수(Maclaurin series)가 다루기 좋은 형태이다. 이것은 중심극한정리(central limit theorem)를, 모든 단계를 손으로 직접 계산할 수 있는 경우에서 보는 것이다.

풀이

그래프 개형

이 확률밀도함수는 우함수이고, $x = 0$에서 최댓값 $\frac{1}{2}$를 가지며 그곳에서 뾰족한 꼭짓점을 이룬다(그 점에서의 좌미분계수와 우미분계수는 각각 $+\frac{1}{2}$와 $-\frac{1}{2}$이다). 그리고 양쪽으로 지수적으로 감소하여 0에 가까워진다.



적률생성함수

정의에 따라 $\mathrm{M}_X(\theta) = \mathrm{E}(\mathrm{e}^{\theta X})$이다. 0을 기준으로 나누면

$$\begin{aligned}\mathrm{M}_X(\theta) &= \tfrac{1}{2}\int_{-\infty}^{0} \mathrm{e}^{(1+\theta)x}\,\mathrm{d}x + \tfrac{1}{2}\int_{0}^{\infty} \mathrm{e}^{-(1-\theta)x}\,\mathrm{d}x \\ &= \frac{1}{2(1+\theta)} + \frac{1}{2(1-\theta)} = \frac{(1-\theta)+(1+\theta)}{2(1-\theta^2)} = \frac{1}{1-\theta^2}.\end{aligned}$$

첫 번째 적분은 $1+\theta > 0$일 때만, 두 번째 적분은 $1-\theta > 0$일 때만 수렴하므로, 이 식은 $-1 < \theta < 1$에서 성립한다. $\theta = 0$ 근방에서는 이것으로 충분하다. 이로써 요구된 결과를 얻는다.

분산(variance)을 구하기 위해 $|\theta| < 1$에서 등비급수(geometric series)로 전개하면

$$\mathrm{M}_X(\theta) = 1 + \theta^2 + \theta^4 + \cdots.$$

이것을 $\mathrm{M}_X(\theta) = 1 + \mathrm{E}(X)\,\theta + \frac{1}{2}\mathrm{E}(X^2)\,\theta^2 + \cdots$와 비교하면 $\mathrm{E}(X) = 0$이고 $\frac{1}{2}\mathrm{E}(X^2) = 1$이다. 따라서

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = 2.$$

(두 번 미분한 뒤 $\theta = 0$을 대입해도 같은 결과를 얻지만, 급수를 쓰는 편이 더 빠르다.)

개구리: $Y/\sqrt{2n}$의 적률생성함수

점프 $X_1, \ldots, X_n$이 독립이므로 $\mathrm{e}^{\theta X_1}, \ldots, \mathrm{e}^{\theta X_n}$도 독립이고, 따라서 이들의 곱의 기댓값은 기댓값들의 곱이다.

$$\mathrm{M}_Y(\theta) = \mathrm{E}\big(\mathrm{e}^{\theta X_1}\cdots\mathrm{e}^{\theta X_n}\big) = \prod_{i=1}^{n}\mathrm{M}_{X_i}(\theta) = (1-\theta^2)^{-n}.$$

$Z = Y/\sqrt{2n}$으로 두면 $\mathrm{E}(\mathrm{e}^{\theta Z}) = \mathrm{E}\big(\mathrm{e}^{(\theta/\sqrt{2n})Y}\big) = \mathrm{M}_Y\big(\theta/\sqrt{2n}\big)$이므로

$$\mathrm{M}_Z(\theta) = \left(1 - \frac{\theta^2}{2n}\right)^{-n}, \qquad |\theta| < \sqrt{2n}.$$

$\sqrt{2n}$으로 나누는 것은 자연스럽다. $\mathrm{Var}(Y) = 2n$이므로 Z의 평균은 0, 분산은 1이 되기 때문이다.

극한

θ를 고정한다. $n > \frac{1}{2}\theta^2$이면 $0 \le \theta^2/(2n) < 1$이므로 로그를 취할 수 있고, $|u| < 1$에서 $\ln(1-u) = -u - \frac{1}{2}u^2 - \frac{1}{3}u^3 - \cdots$를 이용하면

$$\begin{aligned}\ln \mathrm{M}_Z(\theta) &= -n \ln\left(1 - \frac{\theta^2}{2n}\right) = n\left(\frac{\theta^2}{2n} + \frac{\theta^4}{8n^2} + \frac{\theta^6}{24n^3} + \cdots\right) \\ &= \frac{\theta^2}{2} + \frac{\theta^4}{8n} + \frac{\theta^6}{24n^2} + \cdots .\end{aligned}$$

> ? **생각해 보기** $n \to \infty$일 때 이 급수의 나머지 부분이 정말로 사라진다는 것을 어떻게 알 수 있는가?

$u = \theta^2/(2n)$으로 두면, 첫 번째 항 이후의 모든 항의 합은 $n\left(\frac{1}{2}u^2 + \frac{1}{3}u^3 + \cdots\right) \le n \cdot \frac{1}{2}u^2(1 + u + u^2 + \cdots) = \dfrac{nu^2}{2(1-u)} = \dfrac{\theta^4}{8n(1-u)}$이고, $u \to 0$이므로 이 값은 0으로 수렴한다. 따라서 $\ln \mathrm{M}_Z(\theta) \to \frac{1}{2}\theta^2$이고, exp가 연속이므로

$$\mathrm{M}_{Y/\sqrt{2n}}(\theta) \to \exp\left(\tfrac{1}{2}\theta^2\right) \quad \text{단, } n \to \infty,$$

이므로 요구된 결과를 얻는다.

최소 점프 횟수

이 극한은 표준정규분포(standard Normal)의 적률생성함수이므로, n이 클 때 $Z = Y/\sqrt{2n}$이 근사적으로 $\mathrm{N}(0,1)$을 따른다고 본다. "출발점에서 25 cm 이상 떨어진 곳에 착지한다"는 것은 어느 방향으로든 $|Y| \ge 25$라는 뜻이므로

$$\mathrm{P}(|Y| \ge 25) = \mathrm{P}\left(|Z| \ge \frac{25}{\sqrt{2n}}\right) \approx 2\left(1 - \Phi\left(\frac{25}{\sqrt{2n}}\right)\right).$$

이 확률은 n에 따라 증가하므로(개구리가 점점 더 멀리 퍼져 나가므로), 이 확률이 5%에 이르는 최소의 n을 구하면 된다. $\Phi(1.96) = 0.975$이므로 다음이 필요하다.

$$\frac{25}{\sqrt{2n}} \le 1.96 \iff 2n \ge \left(\frac{25}{1.96}\right)^2 \approx 162.7 \iff n \ge 81.3.$$

따라서 최소 점프 횟수는 $n = 82$이다. ($n = 81$일 때 근사 확률은 약 4.95%이고, $n = 82$일 때는 약 5.09%이다.)

검산

- 모든 적률생성함수가 그래야 하듯이 $\mathrm{M}_X(0) = 1$이고, M_X는 우함수로서 그래프의 대칭성과 일치한다.

- $\int x^2 \mathrm{f}(x)\,\mathrm{d}x$를 수치 구적(numerical quadrature)으로 계산하면 2.0000이고, $\int \mathrm{e}^{0.3x}\mathrm{f}(x)\,\mathrm{d}x$는 $1.0989 = 1/0.91$이다.
- $\theta = 1$일 때 $(1 - \frac{1}{2n})^{-n}$의 값은 $n = 10$이면 1.6702, $n = 100$이면 1.6508로, $\mathrm{e}^{1/2} = 1.6487$에 가까워진다.
- 각각 82번의 라플라스(Laplace) 점프를 하는 개구리 10^5마리를 시뮬레이션하면, 그중 약 5.1%가 출발점에서 25 cm 이상 떨어진 곳에서 끝났다.

흔한 실수

- 단측(one-tailed) 값 1.645를 쓰는 것: 개구리는 어느 쪽으로든 25 cm 떨어진 곳에 도달할 수 있으므로 양쪽 꼬리를 모두 세어야 한다.
- $\mathrm{M}_{Y/\sqrt{2n}}(\theta) = \mathrm{M}_Y(\theta)/\sqrt{2n}$으로 쓰는 것. 확률변수에 상수를 곱하면 적률생성함수의 **변수**가 그만큼 곱해진다.
- 81.3을 가장 가까운 정수로 반올림하여 $n = 81$이라고 답하는 것: $n = 81$에서는 확률이 5%보다 약간 작다.

STEP 너머

여러분이 증명한 것은 라플라스 걸음에 대한 중심극한정리이며, 일반적인 증명도 같은 길을 따른다. 즉 $\ln \mathrm{M}(\theta/\sqrt{n})$을 2차항까지 전개하고 나머지가 사라짐을 보이는 것이다. 대학 1학년 과정에서는 "연속성 정리(continuity theorem)"를 배우게 되는데, 이 정리는 적률생성함수(또는 항상 존재하는 특성함수(characteristic function))의 수렴으로부터 확률의 수렴으로 넘어가는 것을 정당화한다. 이것이 여기서 믿고 넘어간 단계이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 두 가지를 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.