풀이

STEP II 2000, 문제 14

확률과 통계 • 중앙값의 분포

> **STEP II 2000, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/00-s2-q14

이 문제는 *Beyond STEP*의 13.4 절 **평균인가, 중앙값인가?**에서 단계별 안내와 함께 자세히 다룬다: https://beyondstep.keithmcnulty.org/ko/ch13/mean-or-median.html

핵심 아이디어

중앙값(median)의 밀도함수(density)는 $y^r(1-y)^s$ 꼴의 다항식이므로, 주어진 적분 하나로 모든 계산을 할 수 있다. 먼저 정규화 상수(normalising constant)를 구하는 데 한 번 쓰고, 이어서 y의 지수를 하나, 그리고 둘 올려서 처음 두 적률(moment)을 구하는 데 다시 쓴다. 마지막 부분에서는 Y와 $\bar{X}$ 모두 $\frac{1}{2}$을 중심으로 하며, 정규분포(normal distribution)에서는 중심에서 일정한 거리 안에 떨어질 확률이 표준편차(standard deviation)에만 의존한다. 퍼짐이 더 작은 확률변수가 이긴다.

풀이

상수 k

밀도함수를 적분하면 1이 되어야 한다. 주어진 결과에 $r = s = n$을 대입하면

$$1 = \int_0^1 k y^n (1-y)^n \, \mathrm{d}y = k \, \frac{n!\, n!}{(2n+1)!}, \qquad \text{따라서} \qquad k = \frac{(2n+1)!}{(n!)^2},$$

이므로 요구한 결과를 얻는다.

평균

밀도함수는 $y = \frac{1}{2}$에 대하여 대칭(symmetric)이다(y를 $1-y$로 바꾸어도 변하지 않는다). 따라서 $\mathrm{E}(Y) = \frac{1}{2}$이다. 이를 적분으로 확인하기 위해 $r = n+1, s = n$으로 두면

$$\mathrm{E}(Y) = k\int_0^1 y^{n+1}(1-y)^n \,\mathrm{d}y = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+1)!\,n!}{(2n+2)!} = \frac{n+1}{2n+2} = \frac{1}{2}.$$

분산

이제 $r = n+2, s = n$으로 두면

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

따라서

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}.$$

중앙값과 평균의 비교

각 X_i는 $[0,1]$에서 균등분포(uniform distribution)를 따르므로 $\mathrm{E}(X_i) = \frac{1}{2}$이고 $\mathrm{Var}(X_i) = \frac{1}{12}$이다. X_i들은 서로 독립(independent)이므로

$$\mathrm{E}(\bar{X}) = \tfrac{1}{2}, \qquad \mathrm{Var}(\bar{X}) = \frac{1}{(2n+1)^2}\cdot(2n+1)\cdot\frac{1}{12} = \frac{1}{12(2n+1)}.$$

생각해 보기 두 확률변수 모두 $\frac{1}{2}$을 중심으로 한다. 어느 쪽이 중심에서 $d/\sqrt{n}$ 이내에 떨어질 가능성이 더 큰지를 결정하는 수 하나는 무엇인가?

바로 표준편차이다. n이 클 때 $Y \sim \mathrm{N}(\frac{1}{2}, \sigma_Y^2)$, $\bar{X} \sim \mathrm{N}(\frac{1}{2}, \sigma_{\bar{X}}^2)$로 볼 수 있다. $W \sim \mathrm{N}(\frac{1}{2}, \sigma^2)$이고 $c > 0$일 때, 표준정규분포의 분포함수(distribution function)를 Φ로 쓰면

$$\mathrm{P}\left(|W - \tfrac{1}{2}| < c\right) = \mathrm{P}\left(|Z| < \frac{c}{\sigma}\right) = 2\Phi\left(\frac{c}{\sigma}\right) - 1,$$

이다. 여기서 $Z \sim \mathrm{N}(0,1)$이다. Φ는 순증가하므로 이것은 σ에 대한 순감소함수(decreasing function)이다. 따라서 $c = d/\sqrt{n}$일 때 요구한 부등식은 정확히 $\sigma_Y > \sigma_{\bar{X}}$일 때, 즉 $\mathrm{Var}(Y) > \mathrm{Var}(\bar{X})$일 때 성립한다.

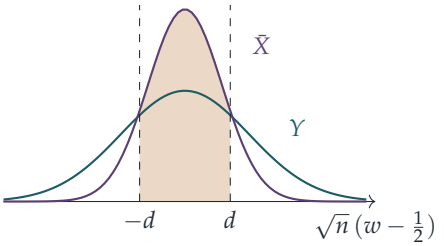
$$\begin{aligned}\frac{1}{4(2n+3)} > \frac{1}{12(2n+1)} &\iff 12(2n+1) > 4(2n+3)\\ &\iff 24n+12 > 8n+12\\ &\iff n > 0.\end{aligned}$$

이는 모든 $n \geq 1$에 대하여 참이다. 따라서 n이 정규근사(normal approximation)가 타당할 만큼 크기만 하면, 양수 d가 무엇이든 Y에 대한 확률은 $\bar{X}$에 대한 확률보다 순전히 작다. 이로써 증명이 끝난다.

차이가 얼마나 큰지 보기 위해 표준화(standardise)하면

$$\frac{d/\sqrt{n}}{\sigma_Y} = 2d\sqrt{\frac{2n+3}{n}} \to 2\sqrt{2}\,d, \qquad \frac{d/\sqrt{n}}{\sigma_X} = d\sqrt{\frac{12(2n+1)}{n}} \to 2\sqrt{6}\,d.$$

점근적으로(asymptotically) 평균의 표준편차가 $\sqrt{3}$배만큼 더 작다.



그림은 두 극한 밀도함수를 $\sqrt{n}\,(w-\frac{1}{2})$ 척도로 나타낸 것이다. 더 좁은 곡선(평균)이 어떤 중심 구간에서든 더 많은 확률을 갖는다.

검산

- $n = 0$: 균등분포 확률변수 하나의 "중앙값"은 그냥 X_1이며, 우리가 구한 공식은 $k = 1$과 $\mathrm{Var}(Y) = \frac{1}{12}$, 즉 균등분포의 분산을 주므로 마땅히 그래야 하는 결과와 맞다.
- 두 분산은 $n = 0$일 때 모두 $\frac{1}{12}$이고 $n \geq 1$이면 서로 달라지는데, 이는 위의 부등식 $n > 0$과 일치한다.
- $n = 50$(101개의 균등분포 확률변수)으로 모의실험(simulation)을 한 결과 $\mathrm{Var}(Y) \approx 0.00243$으로, $1/412 \approx 0.00243$과 같았다. $d = 0.3$일 때 $\mathrm{P}(|Y - \frac{1}{2}| < d/\sqrt{n}) \approx 0.607$, $\mathrm{P}(|\bar{X} - \frac{1}{2}| < d/\sqrt{n}) \approx 0.859$를 얻었는데, 이는 정규근사 값 0.611, 0.860에 가깝다.

흔한 실수

- $\mathrm{Var}(\bar{X}) = \frac{1}{12}$로 쓰거나, 분모에 $(2n+1)^2$을 얻은 뒤 $2n+1$을 곱하는 대신 $2n+1$로 한 번만 나누는 것.
- 방향을 반대로 잡는 것: 분산이 **클수록** 중심 근처에 있을 확률은 **작아진다**.
- $\mathrm{E}(Y^2)$에는 $r = 2n$이 아니라 $r = n+2$를 써야 한다는 것을 잊는 것.

STEP 너머

중앙값의 밀도함수는 베타분포(Beta distribution) Beta$(n+1, n+1)$이고, 주어진 적분은 베타함수(Beta function)이다. 방금 한 비교는 **점근 상대 효율**(asymptotic relative efficiency)이라는 개념이다. 균등분포 자료에서는 표본평균(sample mean)이 중앙값 분산의 3분의 1의 분산으로 중심을 추정하지만, 꼬리가 두꺼운(heavy-tailed) 자료에서는 중앙값이 이길 수 있다. 순서통계량(order statistics)과 베타함수는 Blitzstein and Hwang, *Introduction to Probability*(확률론 입문)에 잘 설명되어 있다.

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.