풀이

STEP III 1989, 문제 15

확률과 통계 • 중앙값의 분포

> **STEP III 1989, 15번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/89-s3-q15

핵심 아이디어

$2n+1$개 값의 중앙값(median)이 z 근처에 놓이는 것은 정확히 한 값이 z에 있는 아주 작은 구간에 들어가고, n개의 값이 그보다 작고, n개의 값이 그보다 클 때이다. 소문항 (i)과 (ii)는 "모두 아래"와 "모두 위"에 해당하는 부분의 확률을 주고, 어떤 값이 어떤 역할을 맡을지 고르는 경우의 수를 세면 밀도(density)를 얻는다. 그다음에는 밀도를 적분하면 1이 된다는 사실로 적분값을 거저 구할 수 있고, 같은 방법을 n 대신 $n+1$로 쓰면 분산(variance)을 얻는다.

풀이

이 풀이 전체에서 X는 $[-c, c]$에서 밀도 $\frac{1}{2c}$를 가지므로, $-c \le k \le c$에 대하여

$$\mathrm{P}(X > k) = \frac{c-k}{2c}, \qquad \mathrm{P}(X < k) = \frac{c+k}{2c}.$$

소문항 (i)

n개의 값은 서로 독립(independent)이므로 확률을 곱한다.

$$\mathrm{P}(n\text{개의 값이 모두} > k) = \left(\frac{c-k}{2c}\right)^n.$$

소문항 (ii)

같은 방법으로,

$$\mathrm{P}(n\text{개의 값이 모두} < k) = \left(\frac{c+k}{2c}\right)^n.$$

중앙값의 밀도

$X_1, \ldots, X_{2n+1}$을 표본(sample)이라 하고, Z를 그 중앙값, 즉 $(n+1)$번째로 작은 값이라 하자. X는 연속(continuous)이므로 같은 값이 나올 확률은 0이고, 따라서 이 경우는 무시한다. 모든 X_i가 $[-c, c]$에 있으므로 Z도 그렇다.

$(-c, c)$에서 z를 잡고 작은 $\delta z > 0$을 잡자. $(\delta z)^2$ 차수의 항을 무시하면, 사건 $z < Z \leq z + \delta z$가 일어나는 것은 정확히 다음과 같을 때이다.

- 값 하나가 $(z, z+\delta z]$에 놓이고,
- 나머지 값 중 n개가 z보다 작고,
- 남은 n개의 값이 $z+\delta z$보다 크다.

(작은 구간에 두 개 이상의 값이 들어가는 배치의 확률은 $(\delta z)^2$ 차수이므로 극한에서는 영향을 주지 않는다.)

? 생각해 보기 $2n+1$개의 값을 이 세 가지 역할에 배정하는 방법은 몇 가지인가?

작은 구간에 들어갈 값을 고르는 방법이 $2n+1$가지이고, 그다음 나머지 중 어느 n개가 z 아래로 갈지 고르는 방법이 $\binom{2n}{n}$가지이다. 따라서 배정의 수는

$$(2n+1)\binom{2n}{n} = \frac{(2n+1)!}{(n!)^2}$$

이다. 이 배정들은 서로 배반이고, 각각의 확률은 같다. 하나의 배정을 고정하면, 독립성과 $k = z$로 둔 소문항 (i), (ii)에 의해 그 확률은

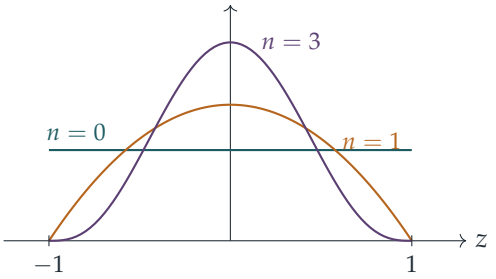
$$\frac{\delta z}{2c}\left(\frac{c+z}{2c}\right)^n\left(\frac{c-z-\delta z}{2c}\right)^n.$$

이다. 모든 배정에 대하여 더하고, δz로 나눈 뒤 $\delta z \to 0$으로 보내면 Z의 밀도는

$$\begin{aligned} g(z) &= \frac{(2n+1)!}{(n!)^2}\cdot\frac{1}{2c}\cdot\frac{(c+z)^n(c-z)^n}{(2c)^{2n}} \\ &= \frac{(2n+1)!}{(n!)^2(2c)^{2n+1}}(c^2-z^2)^n, \qquad -c \leq z \leq c, \end{aligned}$$

이다. 여기서 $(c+z)(c-z) = c^2 - z^2$을 이용하였다. 이것이 구하려던 결과이다.

아래 그림은 $c = 1$일 때의 밀도를 보여 준다. $n = 0$이면 "중앙값"은 값 하나이고, 이는 균등분포(uniform distribution)를 따른다. n이 커질수록 중앙값은 0 근처에 모인다.



적분

확률밀도는 그 범위 전체에서 적분하면 1이므로

$$\frac{(2n+1)!}{(n!)^2(2c)^{2n+1}}\int_{-c}^{c}(c^2-z^2)^n\,\mathrm{d}z = 1,$$

이고, 따라서

$$\int_{-c}^{c}(c^2-z^2)^n\,\mathrm{d}z = \frac{(n!)^2(2c)^{2n+1}}{(2n+1)!}.$$

이다. 이 식은 모든 정수 $n \geq 0$에 대하여 성립하며, 곧 이 사실을 이용한다.

평균과 분산

밀도 g는 z의 우함수(even function)이므로 $zg(z)$는 기함수(odd function)이고, 대칭 구간 $[-c, c]$에서의 적분은 0이다.

$$\mathrm{E}(Z) = 0.$$

따라서 $\mathrm{var}(Z) = \mathrm{E}(Z^2)$이다. $K_n = \dfrac{(2n+1)!}{(n!)^2(2c)^{2n+1}}$, $I_n = \displaystyle\int_{-c}^{c}(c^2-z^2)^n\,\mathrm{d}z = \frac{1}{K_n}$이라 두자.

? **생각해 보기** $z^2(c^2-z^2)^n$을 전개하지 않고 어떻게 적분할 수 있을까?

$z^2 = c^2 - (c^2 - z^2)$으로 쓴다. 그러면

$$\begin{aligned}\mathrm{E}(Z^2) &= K_n\int_{-c}^{c}(c^2-(c^2-z^2))(c^2-z^2)^n\,\mathrm{d}z \\ &= K_n\left(c^2 I_n - I_{n+1}\right) = c^2 - \frac{K_n}{K_{n+1}},\end{aligned}$$

이다. 여기서 $K_n I_n = 1$과, 적분 결과에서 n 대신 $n+1$을 넣은 $I_{n+1} = 1/K_{n+1}$을 이용하였다. 이제

$$\frac{K_n}{K_{n+1}} = \frac{(2n+1)!}{(2n+3)!}\cdot\frac{((n+1)!)^2}{(n!)^2}\cdot\frac{(2c)^{2n+3}}{(2c)^{2n+1}} = \frac{(n+1)^2\cdot 4c^2}{(2n+2)(2n+3)} = \frac{2(n+1)c^2}{2n+3}.$$

따라서

$$\mathrm{var}(Z) = c^2\left(1 - \frac{2n+2}{2n+3}\right) = \frac{c^2}{2n+3}.$$

검산

- $n = 0$이면 중앙값은 X의 값 하나일 뿐이다. 밀도는 $\frac{1}{2c}$가 되고, 적분값은 $2c$, 분산은 $\frac{c^2}{3}$이며, 이는 $[-c, c]$에서의 균등분포의 분산이다.
- $n = 1, c = 1$일 때 적분 공식은 $\int_{-1}^{1}(1 - z^2)\,\mathrm{d}z = \frac{1 \cdot 8}{6} = \frac{4}{3}$을 주며, 이는 옳다.
- $c = 1.7$로 크기 $5(n = 2)$인 표본 10^5개를 시뮬레이션(simulation)한 결과 중앙값의 분산은 0.415였고, 이론값은 $c^2/7 \approx 0.413$이다.

흔한 실수

- 가운데 값을 고르는 인수 $2n + 1$을 빠뜨리고 $\binom{2n}{n}$만 쓰는 실수.
- 가운데 값 자체의 밀도에 해당하는 인수 $\frac{1}{2c}$를 빠뜨려, 거듭제곱이 $(2c)^{2n+1}$ 대신 $(2c)^{2n}$으로 남는 실수.
- $\mathrm{E}(Z^2)$을 구하려고 $(c^2 - z^2)^n$을 이항전개(binomial expansion)하는 것. 가능은 하지만 그 합을 간단히 하기가 어렵다.

STEP 너머

중앙값은 표본의 **순서통계량**(*order statistics*) 가운데 하나이다. 같은 경우의 수 논법으로, 임의의 연속분포에서 뽑은 m개의 값 중 r번째로 작은 값의 밀도가 $\frac{m!}{(r-1)!\,(m-r)!}F^{r-1}(1 - F)^{m-r}f$임을 얻는다. 균등분포의 경우 이들은 **베타분포**(*beta distribution*)이고, 여러분이 유도한 적분은 베타함수(beta function) $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$의 한 경우이다. 분산 $c^2/(2n + 3)$은 $1/n$처럼 줄어드는데, 이는 표본 중앙값의 점근적 정규성(asymptotic normality)으로 가는 첫걸음이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 순서통계량과 베타분포를 잘 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.