풀이

STEP II 2023, 문제 12

확률과 통계 • 최댓값의 분포

> **STEP II 2023, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2023년 STEP 시험지 묶음(zip 파일) 가운데 STEP II 시험지에 있다:
>
> www.ocr.org.uk/Images/707102-2023-past-papers.zip

핵심 아이디어

여러 수의 최댓값(maximum)이 t 이하일 필요충분조건은 그 수들 **하나하나가 모두** t 이하인 것이다. 따라서 Y의 누적분포함수(cumulative distribution function)는 X_i 하나의 누적분포함수의 거듭제곱이다. 나머지는 모두 밀도함수 대신 누적분포함수를 다루는 데서 나온다. 중앙값(median)은 $\mathrm{P}(Y \le m) = \frac{1}{2}$를 직접 풀어서 얻고, 평균(mean)은 부분적분(integration by parts)을 하여 적분 안에 밀도함수가 아닌 분포함수가 나타나도록 하면 깔끔하게 구해진다.

풀이

소문항 (i)

$F(t) = \mathrm{P}(X_1 \le t)$라 하자. $0 \le t \le \pi$일 때

$$F(t) = \int_0^t \tfrac{1}{2} \sin x \, \mathrm{d}x = \tfrac{1}{2}(1 - \cos t).$$

사건 $\{Y \le t\}$는 $X_1 \le t$, $X_2 \le t$, ..., $X_n \le t$가 모두 일어날 때, 그리고 그때에만 일어난다. 가장 큰 값이 t 이하이면 나머지 값도 모두 t 이하이고, 그 역도 성립하기 때문이다. X_i들은 독립(independent)이고 동일한 분포를 따르므로(identically distributed), 이 교사건의 확률은 각 확률의 곱이다.

$$\mathrm{P}(Y \le t) = \mathrm{P}(X_1 \le t)\,\mathrm{P}(X_2 \le t) \cdots \mathrm{P}(X_n \le t) = [\mathrm{P}(X_1 \le t)]^n.$$

따라서 Y의 누적분포함수를 G라 하면

$$G(t) = \frac{(1 - \cos t)^n}{2^n} \qquad (0 \le t \le \pi),$$

이고, $t < 0$일 때 $G(t) = 0$, $t > \pi$일 때 $G(t) = 1$이다. 연쇄법칙(chain rule)으로 미분하면 Y의 확률밀도함수(probability density function)는

$$g(t) = \frac{n\,(1-\cos t)^{n-1}\sin t}{2^n} \quad \text{단, } 0 \le t \le \pi,$$

이고, 그 밖에서는 0이다.

소문항 (ii)

중앙값 $m = m(n)$은 $G(m) = \frac{1}{2}$를 만족한다.

$$\left(\frac{1-\cos m}{2}\right)^n = \frac{1}{2} \iff \frac{1-\cos m}{2} = 2^{-1/n}$$
$$\iff \cos m = 1 - 2^{1-1/n},$$

여기서 $(1-\cos m)/2 \ge 0$이므로 양의 실수 n제곱근을 취하였다. m은 코사인이 일대일(one-to-one)인 구간 $[0, \pi]$에 있어야 하므로

$$m(n) = \arccos(1 - 2^{1-1/n}).$$

n이 커지면 $1/n$은 작아지므로 $2^{1-1/n}$은 (2를 향해) 커지고 $1 - 2^{1-1/n}$은 (-1을 향해) 작아진다. arccos는 감소함수(decreasing function)이므로 $m(n)$은 n에 따라 **증가하고**, $n \to \infty$일 때 $m(n) \to \arccos(-1) = \pi$이다. ($n = 1$이면 $m(1) = \arccos 0 = \pi/2$인데, 이는 $\frac{1}{2}\sin x$가 $\pi/2$에 대하여 대칭(symmetric)이라는 사실과 맞아떨어진다.) 이는 자연스러운 결과이다. 점점 더 많은 값의 최댓값을 취하면 그 값은 범위의 위쪽 끝으로 밀려 올라간다.

소문항 (iii)

정의에 따라 $\mu(n) = \int_0^\pi t\,g(t)\,\mathrm{d}t$이다. $g = G'$이므로 $u = t$, $\mathrm{d}v = g(t)\,\mathrm{d}t$, $v = G(t)$로 두고 부분적분하면

$$\mu(n) = \Big[t\,G(t)\Big]_0^\pi - \int_0^\pi G(t)\,\mathrm{d}t$$
$$= \pi \cdot 1 - 0 - \frac{1}{2^n}\int_0^\pi (1-\cos x)^n\,\mathrm{d}x,$$

이다. 여기서 $G(\pi) = (\frac{1+1}{2})^n = 1$을 이용하였다. 이것이 구하는 공식이다.

> **생각해 보기** 왜 $\int t\,(1-\cos t)^{n-1}\sin t\,\mathrm{d}t$를 직접 계산하려 하지 않고 부분적분을 할까?

목표 공식에는 $(1-\cos x)^n$이 들어 있는데, 이는 밀도함수가 아니라 **분포**함수인 $2^n G(x)$이다. 부분적분은 바로 $t\,g(t)$를 $G(t)$로 바꾸어 주는 방법이며, 동시에 다루기 번거로운 인수 t도 없애 준다.

(a) $h(x) = \frac{1}{2}(1-\cos x)$라 두면 $[0,\pi]$에서 $0 \le h(x) \le 1$이고, $0 < x < \pi$일 때 $0 < h(x) < 1$이다. 공식은 $\mu(n) = \pi - \int_0^\pi h(x)^n \,\mathrm{d}x$가 되므로

$$\mu(n+1) - \mu(n) = \int_0^\pi (h(x)^n - h(x)^{n+1})\,\mathrm{d}x = \int_0^\pi h(x)^n(1-h(x))\,\mathrm{d}x.$$

$0 < x < \pi$에서 $h(x) > 0$이고 $1 - h(x) > 0$이므로 피적분함수(integrand)는 연속이고 순양수이다. 따라서 적분값은 양수이고, 모든 $n \ge 1$에 대하여 $\mu(n+1) > \mu(n)$이다. 즉 $\mu(n)$은 n에 따라 증가한다.

(b) $n = 2$일 때 전개하고 $\cos^2 x = \frac{1}{2}(1+\cos 2x)$를 이용하면

$$\int_0^\pi (1-\cos x)^2\,\mathrm{d}x = \int_0^\pi \left(1 - 2\cos x + \tfrac{1}{2} + \tfrac{1}{2}\cos 2x\right)\mathrm{d}x = \tfrac{3}{2}\pi,$$

이다. $\cos x$와 $\cos 2x$는 모두 $[0,\pi]$에서 적분하면 0이기 때문이다. 따라서

$$\mu(2) = \pi - \tfrac{1}{4}\cdot\tfrac{3}{2}\pi = \frac{5\pi}{8}.$$

소문항 (ii)에서 $m(2) = \arccos(1-\sqrt{2})$이다. $5\pi/8$과 $m(2)$는 모두 코사인이 순감소하는 구간 $[0,\pi]$에 있으므로

$$\mu(2) < m(2) \iff \cos\tfrac{5\pi}{8} > \cos m(2) = 1 - \sqrt{2}.$$

? **생각해 보기** $\cos\frac{5\pi}{8}$의 정확한 값은 어떻게 구할까?

$\cos\frac{5\pi}{8} = \cos(\frac{\pi}{2} + \frac{\pi}{8}) = -\sin\frac{\pi}{8}$으로 쓰고, 반각 공식(half-angle formula) $\sin^2\frac{\pi}{8} = \frac{1}{2}(1-\cos\frac{\pi}{4}) = \frac{2-\sqrt{2}}{4}$를 이용한다. $\sin\frac{\pi}{8} > 0$이므로

$$\cos\tfrac{5\pi}{8} = -\tfrac{1}{2}\sqrt{2-\sqrt{2}}.$$

따라서 $-\frac{1}{2}\sqrt{2-\sqrt{2}} > 1-\sqrt{2}$, 즉 $\sqrt{2}-1 > \frac{1}{2}\sqrt{2-\sqrt{2}}$임을 보이면 된다. 양변이 모두 양수이므로 이는 양변을 제곱하고 4를 곱한 부등식과 동치이다.

$$\begin{aligned} 4(3-2\sqrt{2}) > 2-\sqrt{2} &\iff 10 > 7\sqrt{2} \\ &\iff 100 > 98, \end{aligned}$$

마지막 단계에서는 두 양수를 제곱하였다. 이 부등식은 참이고 모든 단계가 동치였으므로 $\cos\frac{5\pi}{8} > 1-\sqrt{2}$이며, 따라서 $\mu(2) < m(2)$이다.

검산

- $n = 1$이면 $Y = X_1$은 $\pi/2$에 대하여 대칭이고, 실제로 $\mu(1) = \pi - \frac{1}{2}\int_0^\pi (1-\cos x)\,\mathrm{d}x = \pi/2 = m(1)$이다.

- 수치로 보면 $\mu(2) = 5\pi/8 \approx 1.9635$이고 $m(2) = \arccos(1-\sqrt{2}) \approx 1.9979$이므로 평균이 중앙값보다 약간 작다. 1부터 7까지의 각 n에 대하여 최댓값 10^5개를 컴퓨터로 모의실험한 결과는 두 공식과 0.01 이내로 일치하였고(예를 들어 $\mu(3) \approx 2.1598$, $m(3) \approx 2.1986$), 두 수열 모두 증가하였다.
- $n = 2$일 때 Y의 밀도함수는 $\frac{1}{2}(1-\cos t)\sin t$로, 왼쪽으로 치우친(skewed to the left) 분포이며(0 쪽으로 긴 꼬리가 있다), 왼쪽으로 치우친 분포는 보통 평균이 중앙값보다 작다.

흔한 실수

- $\mathrm{P}(Y \le t) = [F(t)]^n$의 근거를 밝히지 않는 것. "모두 t 이하"라는 동치 관계와 독립성이 둘 다 필요하다.
- 부분적분을 하면서 경계항(boundary term) $\left[tG(t)\right]_0^{\pi} = \pi$를 빠뜨리는 것. 공식 맨 앞의 π가 바로 여기서 나온다.
- 코사인 값으로 각의 크기를 비교하면서 코사인이 $[0, \pi]$에서 **감소**한다는 점을 짚지 않아 부등호 방향을 거꾸로 쓰는 것.

STEP 너머

표본의 최댓값은 **순서통계량**(order statistic)의 한 예이며, 소문항 (i)의 방법을 쓰면 어떤 표본 최댓값의 분포든 구할 수 있다. 항등식 $\mathrm{E}[Y] = \pi - \int_0^{\pi} G(t)\,\mathrm{d}t$는 음이 아닌 확률변수(random variable)에 대한 일반적인 꼬리 적분 공식(tail-integral formula) $\mathrm{E}[Y] = \int_0^{\infty} \mathrm{P}(Y > t)\,\mathrm{d}t$의 한 경우로, 대학 확률론 강의 초반에 배우게 된다. 두 아이디어 모두 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)에서 자세히 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.