풀이

STEP II 2011, 문제 13

확률과 통계 • 왜도의 척도

> **STEP II 2011, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/11-s2-q13

핵심 아이디어

왜도(skewness)는 분포가 한쪽으로 치우친 정도를 나타내며, 여기서 다루는 세 가지 척도는 이를 서로 다른 세 가지 방식으로 포착한다. 3차 중심적률(third central moment)을 통해, 두 십분위수(decile) 사이에서 중앙값(median)이 놓인 위치를 통해, 그리고 평균(mean)과 중앙값의 차이를 통해서이다. 구간 $[0,1]$에서 밀도가 $2x$인 분포에서는 모든 것을 명시적으로 계산할 수 있다. 적률(moment)은 한 줄짜리 적분이고, 누적분포함수(cumulative distribution function)는 x^2이므로 그 역함수(inverse function)는 단순히 제곱근이다. 실제로 해야 할 일은 거듭제곱근(surd)이 들어간 식을 신중하게 계산하는 것과, 마지막에 이들을 깔끔하게 비교하는 방법을 찾는 것뿐이다.

풀이

왜도가 측정하는 것. 왜도는 분포가 중심에 대해 얼마나 **비대칭**(asymmetry)인지를 측정한다. 대칭인 분포의 왜도는 0이다. 더 긴 꼬리(tail)가 오른쪽에 있으면 왜도는 양수이고, 더 긴 꼬리가 왼쪽에 있으면 음수이다.

소문항 (i)

μ가 상수라는 점을 기억하면서 세제곱을 전개하고 기댓값의 선형성(linearity of expectation)을 이용한다.

$$\begin{aligned}\mathrm{E}((X-\mu)^3) &= \mathrm{E}(X^3) - 3\mu\,\mathrm{E}(X^2) + 3\mu^2\,\mathrm{E}(X) - \mu^3 \\ &= \mathrm{E}(X^3) - 3\mu(\sigma^2+\mu^2) + 3\mu^3 - \mu^3 \\ &= \mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3,\end{aligned}$$

여기서 $\mathrm{E}(X) = \mu$와 $\mathrm{E}(X^2) = \sigma^2 + \mu^2$(분산(variance)의 일반적인 공식을 정리한 것)를 이용하였다. 양변을 σ^3으로 나누면 γ에 대해 구하려던 식을 얻는다.

밀도가 $2x$인 분포. $k \geq 0$에 대하여

$$\mathrm{E}(X^k) = \int_0^1 x^k \cdot 2x \,\mathrm{d}x = \frac{2}{k+2}.$$

따라서 $\mu = \frac{2}{3}, \mathrm{E}(X^2) = \frac{1}{2}, \mathrm{E}(X^3) = \frac{2}{5}$이다. 그러므로

$$\sigma^2 = \tfrac{1}{2} - \tfrac{4}{9} = \tfrac{1}{18}, \qquad \sigma = \frac{1}{3\sqrt{2}}, \qquad \sigma^3 = \frac{1}{27 \cdot 2\sqrt{2}} = \frac{1}{54\sqrt{2}}.$$

분자는

$$\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3 = \frac{2}{5} - 3 \cdot \frac{2}{3} \cdot \frac{1}{18} - \frac{8}{27} = \frac{2}{5} - \frac{1}{9} - \frac{8}{27} = \frac{54 - 15 - 40}{135} = -\frac{1}{135}.$$

따라서

$$\gamma = -\frac{1}{135} \cdot 54\sqrt{2} = -\frac{54\sqrt{2}}{135} = -\frac{2\sqrt{2}}{5},$$

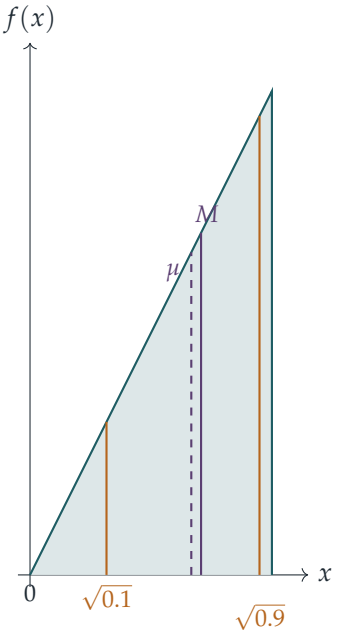
이므로 구하려던 결과가 성립한다. 음의 부호는 타당하다. 밀도가 $x = 1$ 근처에 몰려 있어서, 왼쪽으로 0을 향해 긴 꼬리가 남기 때문이다.

소문항 (ii)

$0 \leq x \leq 1$에 대하여 누적분포함수는

$$\mathrm{F}(x) = \int_0^x 2t \,\mathrm{d}t = x^2,$$

이고, 이 함수는 $[0, 1]$에서 증가하므로 $0 \leq p \leq 1$에 대하여 그 역함수는 $\mathrm{F}^{-1}(p) = \sqrt{p}$이다.



십분위 왜도. D의 분자와 분모에 $\sqrt{10}$을 곱하면 거듭제곱근이 하나만 남고 모두 사라진다.

$$D = \frac{\sqrt{\frac{9}{10}} - 2\sqrt{\frac{1}{2}} + \sqrt{\frac{1}{10}}}{\sqrt{\frac{9}{10}} - \sqrt{\frac{1}{10}}} = \frac{3 - 2\sqrt{5} + 1}{3 - 1} = \frac{4 - 2\sqrt{5}}{2} = 2 - \sqrt{5},$$

여기서 $\sqrt{10} \cdot \sqrt{\frac{1}{2}} = \sqrt{5}$임을 이용하였다. 이것이 보여야 했던 것이다.

피어슨(Pearson) 왜도. 중앙값은 $\mathrm{F}(M) = \frac{1}{2}$를 만족하므로 $M = \frac{1}{\sqrt{2}}$이다. $\frac{1}{\sigma} = 3\sqrt{2}$를 이용하면

$$P = 3\left(\frac{2}{3} - \frac{1}{\sqrt{2}}\right) \cdot 3\sqrt{2} = 9\sqrt{2} \cdot \frac{2}{3} - 9 = 6\sqrt{2} - 9.$$

수치로 계산하면 $D \approx -0.236$, $P \approx -0.515$, $\gamma \approx -0.566$이므로 대소 순서는 맞아 보인다. 이제 이를 증명하자.

> ? **생각해 보기** 계산기 없이 $2 - \sqrt{5}$와 $6\sqrt{2} - 9$ 같은 수를 어떻게 비교할 수 있을까?

양변이 모두 양수가 되도록 정리한 다음 제곱한다. 양수에 대해서는 $a < b$일 필요충분조건이 $a^2 < b^2$이기 때문이다. 아래의 각 단계는 거꾸로도 성립하므로, 마지막 명제가 참이면 처음 명제가 증명된다.

$D > P$. $2 - \sqrt{5} > 6\sqrt{2} - 9$, 즉 $\sqrt{5} + 6\sqrt{2} < 11$임을 보여야 한다. 양변이 모두 양수이므로

이는 다음과 동치이다.

$$\begin{aligned}(\sqrt{5}+6\sqrt{2})^2 = 77+12\sqrt{10} < 121 &\iff 12\sqrt{10} < 44\\ &\iff 3\sqrt{10} < 11\\ &\iff 90 < 121,\end{aligned}$$

여기서 마지막 단계는 다시 두 양수를 제곱한 것이다. $90 < 121$은 참이므로 $D > P$이다.

$P > \gamma$. $6\sqrt{2} - 9 > -\frac{2\sqrt{2}}{5}$임을 보여야 한다. 5를 곱하고 정리하면 이는 $32\sqrt{2} > 45$이다. 양변이 모두 양수이므로 이는 $32^2 \cdot 2 > 45^2$, 즉 $2048 > 2025$와 동치이며, 이는 참이다. 따라서 $P > \gamma$이다.

그러므로 $D > P > \gamma$이다. 세 척도는 모두 음수이며 분포가 왼쪽으로 치우쳐 있다는 점에서는 일치하지만, 그 정도에 대해서는 서로 다르다.

검산

- (i)의 항등식은 임의의 대칭 분포에 대해 0을 준다. 예를 들어 X가 같은 확률로 ± 1의 값을 가지면 $\mu = 0, \sigma = 1, \mathrm{E}(X^3) = 0$이다.
- 중점법(midpoint rule)으로 $\mathrm{E}\big((X-\mu)^3\big)/\sigma^3$을 수치 계산하면 -0.56569가 나오며, 이는 $-2\sqrt{2}/5$와 일치한다.
- U가 $[0,1]$에서 균등분포(uniform distribution)를 따를 때 $X = \sqrt{U}$(이는 $\mathrm{F}(x) = x^2$를 가진다)를 모의실험하면, 그림에서처럼 표본 중앙값은 0.707에 가깝고 십분위수는 0.316과 0.949에 가깝게 나온다.
- 평균 $\frac{2}{3}$은 중앙값 $\frac{1}{\sqrt{2}} \approx 0.707$의 왼쪽에 있는데, 이는 왼쪽으로 치우친 분포의 전형적인 특징이며 $P < 0$인 이유이기도 하다.

흔한 실수

- $(\frac{1}{18})^{3/2}$ 대신 $\sigma^3 = (\frac{1}{18})^3$으로 쓰는 실수. 세제곱하는 것은 분산이 아니라 표준편차(standard deviation)이다.
- 한쪽 변이 음수일 수도 있는데 부등식을 제곱하는 실수. 먼저 양변이 모두 양수가 되도록 정리해야 한다.
- γ의 공식이나 P의 분모에서 $\sigma^2 = \frac{1}{18}$이 필요한 자리에 $\mathrm{E}(X^2) = \frac{1}{2}$를 쓰는 실수.

STEP 너머

$\mathrm{E}\big((X-\mu)^3\big)$은 3차 **중심적률**이며, (i)의 전개는 적률과 **누율**(cumulant) 사이의 관계의 출발점이다. 누율은 독립인 확률변수(independent random variable)들의 합에 대해 깔끔하게 더해진다. D와 같은 분위수(quantile) 기반 척도는 **로버스트**(robust)하다는 이유로 통계학에서 높이 평가된다. 극단적인 관측값 하나는 십분위수를 거의 움직이지 않지만, 3차 적률은 크게 바꿀 수 있다. Grimmett과 Welsh의 *Probability: An Introduction*(확률론 입문)은 적률과 적률생성함수(moment generating function)를 알맞은 수준에서 전개한다.

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.