풀이

STEP III 2025, 문제 11

확률과 통계 • 최댓값의 분포

STEP III 2025, 11번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2025년 STEP 시험지 묶음(zip 파일) 가운데 STEP III 시험지에 있다:

www.ocr.org.uk/Images/750486-2025-past-papers.zip

이 문제는 *Beyond STEP*의 12.25 절 **여럿 가운데 가장 큰 값**에서 단계별 안내와 함께 자세히 다룬다: https://beyondstep.keithmcnulty.org/ko/ch12/extremes.html

핵심 아이디어

X_i 중 가장 큰 값이 y 이하인 것은 **모든** X_i가 y 이하인 것과 정확히 같은 사건이고, 독립성(independence)에 의해 그 확률은 곱으로 바뀐다. 그다음부터 이 문제는 G로부터 분위수(quantile)를 읽어 내고, n이 클 때 그 분위수가 어떻게 되는지 살펴보는 것이 전부이다. 핵심 역할을 하는 근사(approximation)는 단 하나, $\alpha^{1/n} = \mathrm{e}^{(\ln\alpha)/n} \approx 1 + \frac{\ln\alpha}{n}$이다. n이 크면 $(\ln\alpha)/n$이 작기 때문이다.

풀이

소문항 (i)

$x \geq 0$일 때

$$F(x) = \int_0^x \lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = 1 - \mathrm{e}^{-\lambda x}.$$

$Y = \max(X_1, \ldots, X_n)$이므로 사건 $\{Y \leq y\}$는 모든 i에 대하여 $X_i \leq y$인 사건과 같다. X_i들은 서로 독립이므로 $y \geq 0$일 때

$$G(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(X_1 \leq y) \cdots \mathrm{P}(X_n \leq y) = F(y)^n = \left(1 - \mathrm{e}^{-\lambda y}\right)^n,$$

이고, 이것이 구하려던 결과이다.

소문항 (ii)

Y는 연속확률변수(continuous random variable)이므로 $\mathrm{P}(Y < y) = \mathrm{P}(Y \le y) = G(y)$이다. 따라서 $L(\alpha)$에 대한 조건은 $G(L(\alpha)) = \alpha$이다.

$$\begin{aligned}\left(1 - \mathrm{e}^{-\lambda L}\right)^n = \alpha &\iff 1 - \mathrm{e}^{-\lambda L} = \alpha^{1/n} \qquad (\text{양의 } n\text{제곱근을 취하면}) \\ &\iff \mathrm{e}^{-\lambda L} = 1 - \alpha^{1/n} \\ &\iff L(\alpha) = -\frac{1}{\lambda}\ln\left(1 - \alpha^{1/n}\right).\end{aligned}$$

$0 < \alpha < 1$이면 $0 < \alpha^{1/n} < 1$이므로 이 로그는 정의된다.

$U(\alpha)$의 경우, 조건 $\mathrm{P}(Y > U) = \alpha$는 $G(U) = 1 - \alpha$를 뜻하므로, 위와 같은 과정에서 α를 $1 - \alpha$로 바꾸면

$$U(\alpha) = -\frac{1}{\lambda}\ln\left(1 - (1-\alpha)^{1/n}\right).$$

$\alpha \le \frac{1}{2} \le 1 - \alpha$이고 G가 증가함수이므로 $L(\alpha) \le U(\alpha)$이며, 등호는 $\alpha = \frac{1}{2}$일 때에만 성립한다.

소문항 (iii)

$\alpha^{1/n} = \mathrm{e}^{(\ln\alpha)/n}$으로 쓰자. α를 고정하고 n이 크면 $t = (\ln\alpha)/n$은 작으므로, $\mathrm{e}^t \approx 1 + t$에서

$$1 - \alpha^{1/n} \approx -\frac{\ln\alpha}{n} = \frac{1}{n}\ln\left(\frac{1}{\alpha}\right).$$

$\alpha < 1$이므로 이 값은 양수임에 주의한다. 이를 소문항 (ii)의 결과에 대입하면

$$\lambda L(\alpha) = -\ln\left(1 - \alpha^{1/n}\right) \approx -\ln\left(\frac{1}{n}\ln\frac{1}{\alpha}\right) = \ln(n) - \ln\left(\ln\frac{1}{\alpha}\right).$$

> **생각해 보기** 근삿값에 로그를 취해도 괜찮을까? $1 - \alpha^{1/n}$ 자체가 매우 작은 값이므로, $1 - \alpha^{1/n}$의 오차가 그 값에 비해 클 수도 있지 않을까?

중요한 것은 **상대오차**(*relative error*)이다. $t = (\ln\alpha)/n$이라 하면 $1 - \mathrm{e}^t = -t\left(1 + \frac{t}{2} + \cdots\right)$이므로, 정확한 값과 근삿값의 비는 $1 + O(1/n)$이다. 로그를 취하면 $\lambda L(\alpha)$의 오차는 $\ln(1 + O(1/n))$이고, 이는 0으로 수렴(converge)한다. 따라서 이 근사는 n이 클 때 좋은 근사일 뿐 아니라, 양변의 차이가 실제로 0으로 수렴한다.

소문항 (iv)

Y의 중앙값(median) m은 $G(m) = \frac{1}{2}$를 만족하므로 $m = L\left(\frac{1}{2}\right)$이다(이는 $U\left(\frac{1}{2}\right)$와도 같다). 소문항 (iii)에 의해

$$\lambda m \approx \ln n - \ln(\ln 2).$$

$\ln(\ln 2)$는 고정된 수(약 -0.37)이고 $\ln n \to \infty$이므로, n이 커지면 중앙값은 무한대로 발산한다.

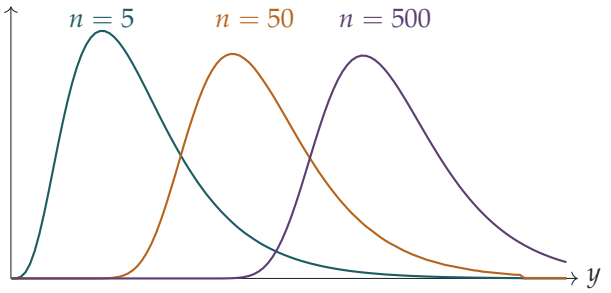
$U(\alpha)$에 대해서는 소문항 (iii)에서 α를 $1-\alpha$로 바꾸면

$$\lambda U(\alpha) \approx \ln n - \ln\left(\ln \frac{1}{1-\alpha}\right).$$

두 식을 빼면 $\ln n$ 항이 소거된다.

$$U(\alpha) - L(\alpha) \approx \frac{1}{\lambda}\left[\ln\left(\ln \frac{1}{\alpha}\right) - \ln\left(\ln \frac{1}{1-\alpha}\right)\right].$$

두 근사의 오차가 모두 0으로 수렴하므로(위의 상자에서 본 것처럼), 구간의 폭은 이 값으로 수렴하며, 이 값은 α와 λ에는 의존하지만 n에는 의존하지 않는다. 즉 n이 커질수록 Y의 분포는 퍼지지 않은 채 $\frac{1}{\lambda}\ln n$처럼 무한대 쪽으로 이동한다.



그림은 $\lambda = 1$일 때 Y의 확률밀도함수(density)를 나타낸다. n이 10배씩 커질 때마다 곡선은 오른쪽으로 약 $\ln 10 \approx 2.3$만큼 이동하고, 그 모양은 일정하게 자리 잡는다.

소문항 (v)

$\alpha = 0.05$로 잡자. 그러면

$$\mathrm{P}\big(L(\alpha) \le Y \le U(\alpha)\big) = 1 - \mathrm{P}(Y < L) - \mathrm{P}(Y > U) = 1 - 2(0.05) = 0.9,$$

이므로, n이 무엇이든 Y는 확률 정확히 0.9로 구간 $[L(0.05), U(0.05)]$에 속한다. n이 클 때, 소문항 (iv)에 의해 이 구간의 폭은 근사적으로

$$\frac{1}{\lambda}\left[\ln(\ln 20) - \ln\left(\ln \frac{1}{0.95}\right)\right].$$

이제 주어진 두 사실을 이용한다.

- $\mathrm{e}^3 \approx 20$이므로 $\ln 20 \approx 3$이고 $\ln(\ln 20) \approx \ln 3 \approx 1.1$이다.
- $t = -0.05$로 두고 $\ln(1+t) \approx t$를 이용하면 $\ln \frac{1}{0.95} = -\ln(1-0.05) \approx 0.05 = \frac{1}{20}$이다. 따라서 $\ln\left(\ln \frac{1}{0.95}\right) \approx \ln \frac{1}{20} = -\ln 20 \approx -3$이다.

그러므로 구간의 폭은 근사적으로

$$\frac{1}{\lambda}(\ln 3 + 3) \approx \frac{4.1}{\lambda} \approx 4\lambda^{-1},$$

이고, 이것이 구하려던 결과이다.

검산

- 소문항 (ii)에서 $L(\alpha)$를 G에 다시 대입하면 정확히 $\left(1-(1-\alpha^{1/n})\right)^n = \alpha$가 된다.
- $\alpha = 0.05$, $n = 10^4$일 때 $\lambda L(\alpha)$의 정확한 값은 8.11330이고 근삿값 $\ln n - \ln\ln 20$은 8.11315이다. $n = 10$일 때는 각각 1.35와 1.21이므로, "충분히 크다"는 조건이 정말 중요하다.
- 소문항 (v)에서 폭의 정확한 극한값은 $4.067/\lambda$로, $4/\lambda$에 충분히 가깝다. 또한 지수분포를 따르는 다섯 확률변수의 최댓값을 컴퓨터로 모의실험한 결과도 소수점 아래 셋째 자리까지 $G(y)$와 일치했다.

흔한 실수

- $G(y) = 1-(e^{-\lambda y})^n$으로 쓰는 실수: 이것은 최댓값이 아니라 **최솟값**의 누적분포함수(cumulative distribution function)이다.
- $P(Y > U) = \alpha$를 $G(U) = 1-\alpha$가 아니라 $G(U) = \alpha$로 놓는 실수.
- 소문항 (v)에서 $\alpha = 0.1$을 택하는 실수: 구간은 양쪽 꼬리에 각각 0.1이 아니라 0.05씩 남겨야 한다.

STEP 너머

이것은 **극값 이론**(*extreme value theory*)을 처음 엿보는 문제이다. 소문항 (iii)은 사실 $\lambda Y - \ln n$이 극한분포(limiting distribution)를 가지며 $P(\lambda Y - \ln n \le x) \to \exp(-e^{-x})$임을 보이는 것이다. 이 분포가 굼벨 분포(Gumbel distribution)로, 기록적인 홍수, 한 해 동안의 가장 강한 돌풍, 가장 큰 보험금 청구액 등을 모형화하는 데 쓰인다. 놀라운 정리에 따르면, 적절히 척도를 조정한 최댓값은 오직 세 가지 분포족 중 하나로만 수렴할 수 있다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 순서통계량(order statistics)을 다루고 있어 시작하기에 좋은 책이다.

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.