풀이

STEP I 2004, 문제 13

확률과 통계 • 균등분포 확률변수의 최댓값

> **STEP I 2004, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/04-s1-q13

핵심 아이디어

서로 독립(independent)인 여러 수의 최댓값(maximum)이 x 이하인 것은 그 수들 **하나하나가 모두** x 이하인 것과 정확히 같다. 이렇게 하면 다루기 까다로운 "최댓값"이 똑같은 확률들의 곱으로 바뀌고, X의 분포함수(distribution function)를 단 한 줄로 얻는다. 나머지는 모두 여기서 따라 나온다. 미분하여 확률밀도함수(density)와 기댓값(expectation)을 구하고, 분포함수를 0.05와 같다고 놓아 검정(test)의 임계값(critical value)을 찾는다.

풀이

소문항 (i)

세 수를 U_1, U_2, U_3이라 하자. 이들은 서로 독립이고 각각 $[0,1]$에서 균등분포(uniform distribution)를 따르므로, $0 \le x \le 1$에 대하여 $\mathrm{P}(U_i \le x) = x$이다. $X = \max(U_1, U_2, U_3)$이므로

$$X \le x \iff U_1 \le x,\ U_2 \le x \text{ 그리고} U_3 \le x.$$

독립성에 의하여

$$F(x) = \mathrm{P}(X \le x) = x^3 \qquad (0 \le x \le 1).$$

특히 $\mathrm{P}(X \le 0.8) = 0.8^3 = 0.512$이므로 요구한 결과가 성립한다.

미분하면 X의 확률밀도함수는 $[0,1]$에서 $f(x) = F'(x) = 3x^2$이므로

$$\mathrm{E}(X) = \int_0^1 x \cdot 3x^2 \,\mathrm{d}x = \left[\tfrac{3}{4}x^4\right]_0^1 = \frac{3}{4}.$$

이는 타당하다. 세 수 가운데 가장 큰 수는 한 수의 평균 $\frac{1}{2}$보다 꽤 위에 있어야 하기 때문이다.

소문항 (ii)

이제 서로 독립인 N개의 수가 있고, 각각 $[0, a]$에서 균등분포를 따르므로 $0 \le x \le a$에 대하여 $\mathrm{P}(U_i \le x) = x/a$이다. 같은 논법으로

$$\mathrm{P}(X \le x) = \left(\frac{x}{a}\right)^N \qquad (0 \le x \le a),$$

이고, $x \ge a$이면 $\mathrm{P}(X \le x) = 1$이다. X는 연속확률변수이므로 $\mathrm{P}(X < x)$도 이와 같다.

임계값. 이 검정은 $x < c$일 때 H_0을 기각하며, 유의수준(significance level) 5%라는 것은 H_0이 참일 때($a = 1$) 기각할 확률이 0.05라는 뜻이다.

$$\mathrm{P}(X < c \mid a = 1) = c^N = 0.05, \qquad \text{따라서} \qquad c = 0.05^{1/N}.$$

x의 값이 작으면 $a < 1$이라는 증거가 되므로, 기각역(rejection region)은 $x < c$이다.

> **?** **생각해 보기** "$x \le 0.8$이면 언제나 기각된다"는 조건을 어떻게 N에 대한 부등식으로 바꿀 수 있을까?

모든 $x \le 0.8$이 기각되는 것은 그러한 x가 모두 c보다 작을 때, 즉 $0.8 < c$일 때와 정확히 같다. $t > 0$에서 $t \mapsto t^N$은 증가함수이므로, 이는 다음과 동치이다.

$$0.8^N < c^N = 0.05.$$

따라서 $0.8^N < 0.05$를 만족하는 가장 작은 정수 N을 구해야 한다. (N이 커질수록 c가 1에 가까워지므로, N이 클수록 검정이 더 쉽게 기각한다는 점에 주의하자.)

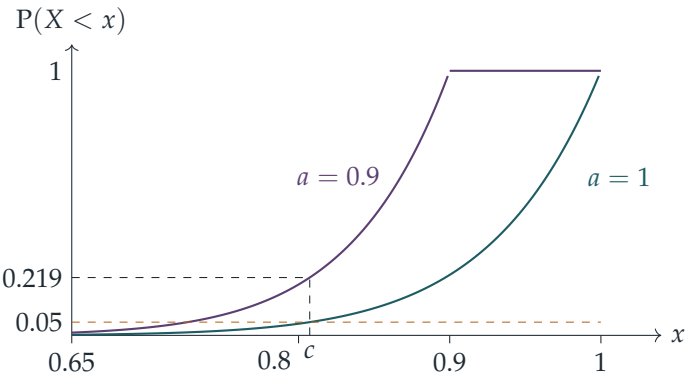
$2^{10} \approx 10^3$ 이용하기. $0.8 = 2^3/10$으로 쓰면 $0.8^N = 2^{3N}/10^N$이다.

- $N = 13$: $0.8^{13} = \dfrac{2^{39}}{10^{13}} = \dfrac{2^9 \cdot 2^{30}}{10^{13}} \approx \dfrac{512 \times 10^9}{10^{13}} = 0.0512 > 0.05.$
- $N = 14$: $0.8^{14} = \dfrac{2^{42}}{10^{14}} = \dfrac{2^2 \cdot 2^{40}}{10^{14}} \approx \dfrac{4 \times 10^{12}}{10^{14}} = 0.04 < 0.05.$

N이 커질수록 0.8^N은 감소하므로, $N \le 13$인 모든 경우도 조건을 만족하지 못한다. 따라서 가장 작은 값은

$$N = 14.$$

($2^{10} = 1024$이므로 이 근사는 2의 거듭제곱을 약간 작게 어림한다. 실제 값 $0.8^{13} \approx 0.0550$과 $0.8^{14} \approx 0.0440$으로도 같은 결론에 이른다.)



$N = 14$일 때 임계값은 $c = 0.05^{1/14} \approx 0.807$로, 0.8보다 아주 조금 크다.

$a = 0.8$인 경우. 모든 수가 $[0, 0.8]$에 있으므로 반드시 $X \le 0.8 < c$이다. 귀무가설(null hypothesis)은 확률 1로 기각된다.

$a = 0.9$인 경우. 이제 $c \approx 0.807 < 0.9$이므로

$$\mathrm{P}(\text{기각}) = \mathrm{P}(X < c \mid a = 0.9) = \left(\frac{c}{0.9}\right)^{14} = \frac{c^{14}}{0.9^{14}} = \frac{0.05}{0.9^{14}}.$$

여기서 $0.9^2 = 0.81, 0.9^4 = 0.6561, 0.9^8 \approx 0.4305$이므로 $0.9^{14} = 0.9^8 \cdot 0.9^4 \cdot 0.9^2 \approx 0.2288$이다. 따라서

$$\mathrm{P}(\text{기각} \mid a = 0.9) = \frac{0.05}{0.9^{14}} \approx 0.219.$$

즉 관측값이 14개나 있어도, 이 검정은 a가 1이 아니라 0.9라는 사실을 대개 알아채지 못한다.

검산

- 몬테카를로 시뮬레이션(Monte Carlo simulation)(2×10^5회 시행)으로 소문항 (i)에서 $\mathrm{P}(X \le 0.8) \approx 0.512$, $\mathrm{E}(X) \approx 0.750$을 얻었다.
- 일반적으로 $[0, 1]$에서 균등분포를 따르는 n개의 수의 최댓값의 평균은 $\frac{n}{n+1}$이다. $n = 1$이면 당연히 $\frac{1}{2}$이고, $n = 3$이면 $\frac{3}{4}$이다.
- $N = 14$, $c = 0.05^{1/14}$일 때 시뮬레이션 결과 기각 비율은 $a = 1$에서 약 0.0505(검정의 크기(size))였고, $a = 0.9$에서 약 0.218로 $0.05/0.9^{14} = 0.2186$과 일치하였다.

흔한 실수

- 0.8^N 대신 $1 - 0.8^N$을 쓰는 실수. 최댓값이 **작을** 확률은 곱 0.8^N이지, 그 여사건의 확률이 아니다.
- $a = 0.9$에서의 기각 확률을 $(0.8/0.9)^{14}$로 계산하는 실수. 검정은 $x < c$일 때 기각하며, c는 0.8이 아니다.
- 0.0512가 "대략 0.05"라는 이유로 $N = 13$에서 멈추는 실수. 0.05보다 엄격하게 작아야 한다.

STEP 너머

최댓값은 **순서통계량**(order statistic)의 하나이며, $[0, a]$에서의 균등분포에 대해서는 a의 **충분통계량**(sufficient statistic)이다. 즉 표본이 a에 대하여 담고 있는 모든 정보를 최댓값이 담고 있다. 참값 a의 함수로 본 기각 확률을 검정의 **검정력 함수**(power function)라 한다. 여기서는 $a \leq 1$에 대하여 $\min\{1, 0.05/a^N\}$이며, 통계학자들은 검정력 함수를 비교하여 여러 검정 가운데 하나를 고른다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 최댓값과 최솟값의 분포를 다루고, Blitzstein과 Hwang의 *Introduction to Probability*(확률 입문)는 순서통계량을 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.