풀이

STEP II 2024, 문제 11

확률과 통계 • 그룹 검사의 기댓값

> **STEP II 2024, 11번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2024년 STEP 시험지 묶음(zip 파일) 가운데 STEP II 시험지에 있다:
>
> www.ocr.org.uk/Images/725753-2024-past-papers.zip

핵심 아이디어

여러 사람의 시료를 합쳐 검사하는 방식(pooling)은 그룹 검사가 대개 음성으로 나올 때에만 이득이 되며, 한 사람당 검사 횟수의 기댓값(expected value)은 $1+\frac{1}{k}-(1-p)^k$이 된다. 이 값이 1 이하가 되는 조건을 묻는 것은 결국 $(1-p)$와 $k^{-1/k}$를 비교하는 문제가 되고, 바로 이 때문에 소문항 (i)에서 $x^{1/x}$를 묻는 것이다. 즉 p가 클 때 가장 좋은 그룹 크기는 $k^{1/k}$를 가장 크게 만드는 정수이다.

풀이

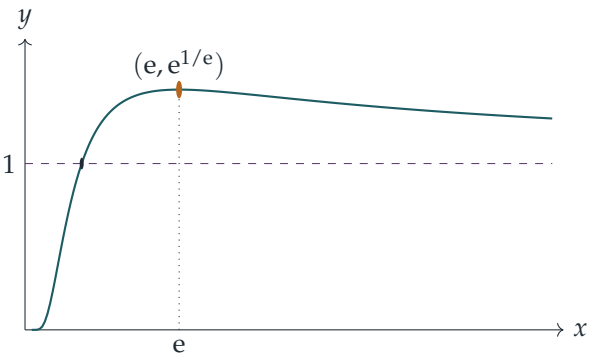
소문항 (i)

$x>0$에 대하여 $y=x^{1/x}=\mathrm{e}^{(\ln x)/x}$라 하면 $\ln y=\dfrac{\ln x}{x}$이다. 미분하면

$$\frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{1-\ln x}{x^2}.$$

$y>0$이므로 $\frac{\mathrm{d}y}{\mathrm{d}x}$의 부호는 $1-\ln x$의 부호와 같다. 즉 $0<x<\mathrm{e}$에서 양수이고, $x=\mathrm{e}$에서 0이며, $x>\mathrm{e}$에서 음수이다. 따라서 극점(turning point)은 하나뿐이며, 그것은 $(\mathrm{e},\mathrm{e}^{1/\mathrm{e}})\approx(2.72,\,1.44)$에서의 극대(maximum)이다.

그래프의 양 끝을 보자. $x\to 0^+$일 때 $(\ln x)/x\to-\infty$이므로 $y\to 0$이고, $x\to\infty$일 때 $(\ln x)/x\to 0$이므로 y는 위쪽에서 1로 다가간다($x>1$이면 $\ln y>0$이기 때문이다). 곡선은 $(1,1)$을 지난다.



양의 정수 n에 대하여, 이 함수는 $(0, \mathrm{e}]$에서 증가하고 $[\mathrm{e}, \infty)$에서 감소하므로 $1^{1/1} < 2^{1/2}$이고 $3^{1/3} > 4^{1/4} > 5^{1/5} > \cdots$이다. 따라서 최댓값은 $2^{1/2}$ 또는 $3^{1/3}$이다. 두 수를 각각 여섯 제곱하면 $2^3 = 8, 3^2 = 9$이므로 $3^{1/3} > 2^{1/2}$이고

$$\max_n n^{1/n} = 3^{1/3}.$$

소문항 (ii)

$q = 1 - p$라 하자. k명으로 이루어진 그룹 하나를 생각한다. 독립성(independence)에 의해 그룹 안의 누구도 효소를 갖지 않을 확률은 q^k이고, 오직 이 경우에만 합친 시료가 음성으로 나온다. 따라서 이 그룹이 사용하는 검사 횟수는

$$\begin{cases} 1\text{회 검사} & \text{확률} q^k, \\ 1+k\text{회 검사} & \text{확률} 1-q^k, \end{cases}$$

이고, 검사 횟수의 기댓값은 $1 + k(1 - q^k)$이다. 그룹은 $r = N/k$개이고 기댓값은 더할 수 있으므로, 전체 검사 횟수의 기댓값은

$$E = \frac{N}{k}(1 + k(1 - q^k)) = N\left(1 + \frac{1}{k} - (1-p)^k\right).$$

소문항 (iii)

소문항 (ii)에 의해 $E \le N$일 필요충분조건은

$$\frac{1}{k} \le (1-p)^k \iff 1 - p \ge k^{-1/k} = \frac{1}{k^{1/k}},$$

이다. 여기서 양변의 양의 k제곱근을 취하였다(양변이 모두 양수이고, $t \mapsto t^{1/k}$는 증가함수이다).

생각해 보기 어떤 k가 가장 큰 p를 허용하는가?

주어진 k에 대하여 조건은 $p \le 1 - k^{-1/k}$이고, 이 상한은 $k^{1/k}$가 클수록 크다. 소문항 (i)에 의해 모든 정수 $k > 1$에 대하여 $k^{1/k} \le 3^{1/3}$이므로, 어떤 k를 택하든 $p \le 1 - 3^{-1/3}$이어야 한다. 역으로 N은 3의 배수이므로 $k = 3$은 허용되는 그룹 크기이고, $k = 3$일 때 조건은 정확히 $p \le 1 - 3^{-1/3}$이다(등호일 때 $E = N$이다). 따라서 가장 큰 값은

$$p = 1 - 3^{-1/3} \approx 0.307.$$

이 값이 $\frac{1}{4}$보다 큼을 보이자.

$$1 - 3^{-1/3} > \tfrac{1}{4} \iff 3^{-1/3} < \tfrac{3}{4} \iff \tfrac{1}{3} < \tfrac{27}{64} \iff 64 < 81,$$

가운데 단계에서는 세제곱을 하였다(양수에 대해 세제곱은 대소 관계를 보존한다). 마지막 명제가 참이므로 $1 - 3^{-1/3} > \frac{1}{4}$이다.

소문항 (iv)

이항 전개(binomial expansion)에 의해

$$(1-p)^k = 1 - kp + \binom{k}{2} p^2 - \cdots.$$

pk가 작으면 $\binom{k}{2}p^2 < \frac{1}{2}(pk)^2$은 그보다 훨씬 더 작고, 그 뒤의 항들도 마찬가지이므로 $(1-p)^k \approx 1 - pk$이다. 이를 소문항 (ii)의 식에 대입하면

$$E \approx N\left(1 + \frac{1}{k} - (1 - pk)\right) = N\left(\frac{1}{k} + pk\right).$$

(실제로 베르누이 부등식(Bernoulli's inequality) $(1-p)^k \ge 1 - pk$에 의해 이 근사는 E를 결코 과소평가하지 않으며, 전개를 한 항 더 늦게 끊으면 오차가 많아야 $\frac{1}{2}(pk)^2 N$임을 알 수 있다.)

$p = 0.01, k = 10$이면 $pk = 0.1$로 작은 값이고,

$$E \approx N\left(\frac{1}{10} + 0.1\right) = 0.2N,$$

이므로, 이 방식은 모든 사람을 따로 검사할 때 필요한 N회 검사의 약 20%만 필요로 한다.

검산

- $p = 0.01, k = 10$일 때 정확히 계산하면 $E/N = 1.1 - 0.99^{10} \approx 1.1 - 0.9044 = 0.1956$, 즉 약 19.6%로, 근사의 오차 한계 $\frac{1}{2}(0.1)^2 = 0.005$ 이내에 있다.
- $p = 1 - 3^{-1/3}, k = 3$일 때 $(1-p)^3 = \frac{1}{3}$이므로 $E = N(1 + \frac{1}{3} - \frac{1}{3}) = N$으로, 정확히 경계에 놓인다.

- 소문항 (ii)의 타당성: $p \to 0$이면 $E \to N/k$(그룹당 한 번 검사)이고, $p \to 1$이면 $E \to N(1+\frac{1}{k})$(모든 그룹을 재검사)이다. 이 공식은 작은 그룹에 대해 모든 경우를 나열하는 방식으로 컴퓨터를 이용해서도 확인하였다.

흔한 실수

- 소문항 (i)의 답을 $e^{1/e}$로 쓰는 것: n은 정수여야 하므로 $2^{1/2}$와 $3^{1/3}$을 비교해야 한다.
- 소문항 (iii)에서 $k = 3$만 확인하고, 왜 다른 그룹 크기로는 더 큰 p를 허용할 수 없는지 설명하지 않는 것.
- 양성인 그룹에서 합친 시료에 대한 검사 자체를 빠뜨려, $k+1$회가 아니라 k회 검사로 세는 것.

STEP 너머

이것은 Dorfman의 그룹 검사(group testing) 방식으로, 제2차 세계 대전 중 병사들의 혈액 시료를 선별하기 위해 고안되었으며 COVID-19 검사에서 다시 쓰이게 되었다. $\frac{1}{k} + pk$를 최소화하면 $k \approx 1/\sqrt{p}$(따라서 여기서는 $k = 10$이 최적이다)와 약 $2\sqrt{p}\,N$회의 검사를 얻는다. 정보 이론(information theory)에서 연구하는 더 영리한 적응적(adaptive) 방식은 이보다 더 나은 결과를 낸다. 그 바탕이 되는 확률론에 대해서는 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)를 참고하라.

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.