풀이

STEP II 2000, 문제 13

확률과 통계 • 포획-재포획 추정

> **STEP II 2000, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/00-s2-q13

이 문제는 *Beyond STEP*의 13.2 절 **들쥐 세기**에서 단계별 안내와 함께 자세히 다룬다: https://beyondstep.keithmcnulty.org/ko/ch13/voles.html

핵심 아이디어

이것은 포획–재포획(capture–recapture) 문제이다. 두 번째 표본은 N마리 가운데 200마리의 들쥐를 무작위로 고른 것이고, 그중 200마리가 표시되어 있으므로, 표본 속 표시된 들쥐의 수는 초기하분포(hypergeometric distribution)를 따른다. p_N이 최대가 되는 곳을 찾을 때는 미분하지 말아야 한다(N의 계승(factorial)은 쉬운 의미로는 미분할 수 없다). 대신 비 p_N / p_{N-1}을 통해 이웃한 항을 비교하면 거의 모든 것이 약분된다. 마지막 항등식은 덤으로 얻어진다. 세 번째 표본에 대한 확률들의 합은 1이어야 하기 때문이다.

풀이

확률 p_N

두 번째 표본을 N마리 들쥐 가운데 200마리를 고르는 것으로 모형화하고, 크기가 200인 모든 부분집합이 똑같은 확률로 뽑힌다고 하자. 이러한 부분집합은 $\binom{N}{200}$개 있다. N마리 가운데 200마리는 표시되어 있고 $N-200$마리는 표시되어 있지 않다. 표시된 들쥐가 정확히 11마리인 표본은 표시된 들쥐 11마리와 표시되지 않은 들쥐 189마리를 고른 것이므로

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

N이 들어 있는 두 이항계수(binomial coefficient)를 계승으로 쓰면

$$\begin{aligned} p_N &= \binom{200}{11} \cdot \frac{(N-200)!}{189!\,(N-389)!} \cdot \frac{200!\,(N-200)!}{N!} \\ &= k\,\frac{((N-200)!)^2}{N!\,(N-389)!}, \qquad \text{여기서} k = \binom{200}{11}\frac{200!}{189!}, \end{aligned}$$

이고, k는 N에 의존하지 않는다. 이것이 요구된 꼴이다. (두 표본을 합하면 서로 다른 들쥐 $200+189=389$마리가 들어 있으므로 $N \geq 389$이어야 한다.)

p_N의 최대화

$N \geq 390$일 때 p_N을 p_{N-1}로 나눈다. $(N-200)! = (N-200)\,(N-201)!$, $N! = N\,(N-1)!$, $(N-389)! = (N-389)\,(N-390)!$을 이용하면 다음을 제외한 모든 것이 약분된다.

$$\frac{p_N}{p_{N-1}} = \frac{(N-200)^2}{N(N-389)}.$$

분자와 분모가 모두 양수이므로

$$\begin{aligned} p_N > p_{N-1} &\iff (N-200)^2 > N(N-389) \\ &\iff N^2 - 400N + 40000 > N^2 - 389N \\ &\iff 40000 > 11N \\ &\iff N < 3636.\overline{36}. \end{aligned}$$

생각해 보기 이것은 수열 $p_{389}, p_{390}, \ldots$의 모양에 대해 무엇을 말해 주는가?

$390 \leq N \leq 3636$이면 비가 1보다 크므로 수열은 p_{3636}까지 순증가한다. $N \geq 3637$이면 비가 1보다 작으므로(등호가 성립하려면 $11N = 40000$이어야 하는데, 이 식은 정수해를 갖지 않는다) 수열은 그 이후로 순감소한다. 따라서 p_N은

$$N = 3636.$$

에서 유일한 최댓값을 갖는다. 이는 $\frac{200\times200}{11}$의 정수 부분이다. 표본에서 표시된 비율 $\frac{11}{200}$이 모집단에서 표시된 비율 $\frac{200}{N}$과 같아야 한다고 생각하면 짐작할 수 있는 값이다.

세 번째 표본

모집단의 크기를 방금 구한 추정값 3636으로 놓자. 두 번째로 풀어 준 뒤에는 원래의 200마리에 새로 표시된 189마리가 더해져 표시된 들쥐가 389마리이고, 표시되지 않은 들쥐는 $3636-389=3247$마리이다. 앞에서와 똑같이 추론하면, 크기가 200인 무작위 표본에 표시

된 들쥐가 정확히 j마리 들어 있을 확률은

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}, \qquad j = 0, 1, \ldots, 200.$$

$389 \geq 200$이고 $3247 \geq 200$이므로 0부터 200까지의 모든 j가 가능하며, 이것들이 일어날 수 있는 모든 결과이다. 이 결과들은 서로 배반이므로 확률의 합은 1이다.

$$\sum_{j=0}^{200} \frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}} = 1.$$

양변에 $\binom{3636}{200}$을 곱하면

$$\sum_{j=0}^{200} \binom{389}{j}\binom{3247}{200-j} = \binom{3636}{200},$$

이 되어 증명이 끝난다.

검산

- 정수 이항계수로 p_N을 정확히 계산하면 여러 N 값(예를 들어 389, 1000, 5000)에서 계승 공식과 일치하고, $389 \leq N < 20000$ 범위를 탐색하면 최댓값이 $N = 3636$에서 나온다.
- 전환점에서 $p_{3636}/p_{3635} = 3436^2/(3636 \times 3246) \approx 1.0003 > 1$이지만 $p_{3637}/p_{3636} < 1$이다.
- 마지막 항등식의 양변을 정수로 정확히 계산하면 서로 같다. 마땅히 그래야 하듯이 $389 + 3247 = 3636$임에 주목하라.

흔한 실수

- N을 연속인 양으로 보고 미분하는 것. 대신 이웃한 항의 비를 이용하고, 부등식을 정확히 판정하라.
- 두 번째로 풀어 준 뒤에는 표시된 들쥐가 200마리가 아니라 389마리라는 것을 잊는 것.
- "$N = 3636.36$에서 최대"라고 말하는 것. 모집단의 크기는 자연수여야 하며, 이웃한 두 정수 가운데 어느 쪽이 더 큰지 확인해야 한다.

STEP 너머

관측된 자료가 나올 확률을 가장 크게 만드는 모수의 값을 고르는 것을 **최대가능도**(maximum likelihood) 방법이라고 하며, 이는 대학 1학년 통계학의 핵심 아이디어이다. 여기서는 이 방법이 포획–재포획에 대한 링컨–피터슨 추정량(Lincoln–Petersen estimator)을 준다. 마지막 항등식은 방데르몽드 항등식(Vandermonde's identity)이며, 여기서 쓴 논법, 즉 같은 것을 두 가지 방법으로 세는 것은 조합론에서 강력한 기법이다. Grimmett와 Welsh의 *Probability: An Introduction* (확률론 입문서)에서 초기하분포와 이러한 논법을 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.