풀이

STEP I 2000, 문제 13

확률과 통계 • 조건부 확률

> **STEP I 2000, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/00-s1-q13

핵심 아이디어

모든 것은 부모 각각에게서 하나씩 받는 유전자(gene)를 추적하는 일로 귀결된다. (i)에서 두 AB 부모의 각 자녀는 같은 작은 확률로 AA 또는 BB가 되고, 네 자녀는 서로 독립(independent)이므로 배열의 수를 세어 곱하면 된다. (ii)는 "거꾸로 가는" 조건부 확률(conditional probability)이다. 결과(자녀)를 알고 있을 때 원인(아버지)을 알고 싶은 것이다. 이것이 바로 베이즈 정리(Bayes' theorem)이며, 아버지의 가능한 세 유전자형(genotype)에 대한 수형도(tree diagram)로 정리할 수 있다.

풀이

소문항 (i)

부모가 모두 AB이므로 각자 A 또는 B를 확률 $\frac{1}{2}$로 물려주며, 두 부모는 서로 독립적으로 작용한다. 자녀 한 명에 대하여

$$\mathrm{P}(AA) = \tfrac{1}{2}\cdot\tfrac{1}{2} = \tfrac{1}{4}, \qquad \mathrm{P}(BB) = \tfrac{1}{4}, \qquad \mathrm{P}(AB) = \tfrac{1}{2},$$

이다. 여기서 마지막 확률은 두 순서(어머니에게서 A, 아버지에게서 B를 받는 경우와 그 반대의 경우)를 모두 센 것이다.

서로 다른 자녀들은 서로 독립적으로 유전자를 물려받는다. 네 자녀 중 둘이 AA, 둘이 BB인 특정한 순서, 예를 들어 AA, AA, BB, BB의 확률은 $\left(\frac{1}{4}\right)^4$이다. 네 자녀 중 어느 두 명이 AA인지 고르는 방법의 수는 $\binom{4}{2} = 6$이다. 따라서

$$\mathrm{P}(\text{두 명은}AA,\ \text{두 명은}BB) = 6\times\left(\tfrac{1}{4}\right)^4 = \frac{6}{256} = \frac{3}{128},$$

이고, 이것이 증명하려던 결과이다.

소문항 (ii)

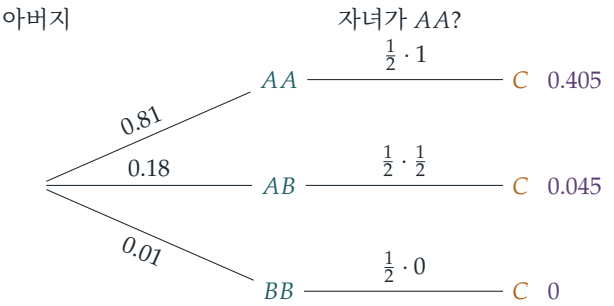
F_{AA}, F_{AB}, F_{BB}를 각각 나의 아버지가 AA, AB, BB인 사건이라 하고, C를 내가 AA인 사건이라 하자. 아버지는 어머니의 유전자형과 독립적으로 선택된, 집단의 무작위 구성원이라고 보자. 그러면

$$\mathrm{P}(F_{AA}) = 0.81, \qquad \mathrm{P}(F_{AB}) = 0.18, \qquad \mathrm{P}(F_{BB}) = 0.01.$$

생각해 보기 왜 0.18이 곧바로 답이 되지 않을까?

추가 정보, 즉 내가 AA라는 사실이 있기 때문이다. 이 사실은 BB인 아버지를 완전히 배제하고, AA인 아버지는 언제나 A를 물려주므로 AA인 아버지의 가능성을 높인다. 따라서 C를 조건으로 하여 계산해야 한다.

내가 AA이려면 어머니에게서 A를, 아버지에게서도 A를 받아야 한다. 어머니는 AB이므로 아버지가 누구든 확률 $\frac{1}{2}$로 A를 준다. 아버지는 유전자형에 따라 확률 1, $\frac{1}{2}$, 0으로 A를 준다. 수형도는 다음과 같다.



전확률 법칙(law of total probability)에 의하여

$$\begin{aligned}\mathrm{P}(C) &= 0.81 \cdot \tfrac{1}{2} \cdot 1 + 0.18 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} + 0.01 \cdot \tfrac{1}{2} \cdot 0 \\ &= 0.405 + 0.045 + 0 = 0.45.\end{aligned}$$

그러면 베이즈 정리에 의하여

$$\mathrm{P}(F_{AB} \mid C) = \frac{\mathrm{P}(F_{AB})\,\mathrm{P}(C \mid F_{AB})}{\mathrm{P}(C)} = \frac{0.045}{0.45} = \frac{1}{10}.$$

따라서 나의 아버지가 AB일 확률은 $\frac{1}{10}$이다.

검산

- (i)에서 네 자녀의 유전자형 개수에 대한 확률을 모두 더하면 1이 되어야 한다. $\binom{4}{2}$라는 인수는 $\left(\frac{1}{4} + \frac{1}{2} + \frac{1}{4}\right)^4$의 다항 전개(multinomial expansion)에 나타나는 것과 같은 인

수이며, $3/128 \approx 0.0234$는 그 전개의 한 항이다. 여덟 개의 유전자가 유전되는 2^8가지 방법을 모두 정확히 나열해 보아도 $3/128$을 얻는다.

- (ii)에서 답은 사전확률(prior) 0.18에서 0.1로 줄었다. BB인 아버지가 배제되고 AA인 아버지가 유리해졌으니 당연한 결과이다. 10^6개의 가족을 시뮬레이션한 결과 0.0998을 얻었다.

흔한 실수

- (i)에서 인수 $\binom{4}{2} = 6$을 빠뜨려 $1/256$을 얻는 실수.
- (ii)에서 0.18이라고 답하거나, $\mathrm{P}(F_{AB} \mid C)$ 대신 $\mathrm{P}(C \mid F_{AB}) = \frac{1}{4}$를 계산하는 실수.
- 어머니의 인수 $\frac{1}{2}$를 일부 가지에서만 빠뜨리는 실수. 이 인수는 결국 약분되지만, 일관되게 사용했을 때만 그렇다.

STEP 너머

집단에서의 비율 81%, 18%, 1%는 $p = 0.9$, $q = 0.1$일 때의 p^2, $2pq$, q^2이다. 이것이 하디-바인베르크 평형(Hardy–Weinberg equilibrium)으로, 무작위 교배(random mating)에서는 세대에서 세대로 유지된다. 관찰된 결과로부터 숨은 원인을 추론하는 것은 베이즈 통계학(Bayesian statistics)의 핵심이며, 같은 계산이 유전 상담과 친자 확인 검사의 바탕이 된다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 조건부확률과 베이즈 정리를 이와 같은 많은 예제와 함께 철저하게 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.