풀이

STEP II 2005, 문제 12

확률과 통계 • 베이즈 정리

> **STEP II 2005, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/05-s2-q12

핵심 아이디어

이 문제는 공정한 동전을 사전확률(prior)로 하는 베이즈 정리(Bayes' theorem)이다. 동전이 공정하므로, 답변이 주어졌을 때 앞면일 확률이 $\frac{1}{2}$이 되는 것은 동전이 앞면이든 뒷면이든 그 답변이 나올 가능성이 **똑같을** 때이고, 오직 그때뿐이다. 따라서 각 소문항에서는 두 가능도(likelihood) $\mathrm{P}(\text{답변} \mid H)$와 $\mathrm{P}(\text{답변} \mid T)$를 비교하기만 하면 되고, 이것을 구할 때는 쌍둥이 중 누가 접속해 있는지에 대해 조건부(conditioning)로 계산한다.

풀이

동전이 앞면, 뒷면이 나오는 사건을 각각 H, T라 하면 $\mathrm{P}(H) = \mathrm{P}(T) = \frac{1}{2}$이다. 안나(Anna)가 접속해 있는 사건을 A, 벨라(Bella)가 접속해 있는 사건을 B라 하고, $\mathrm{P}(A) = p$, $\mathrm{P}(B) = q = 1 - p$라 하자. 누가 접속해 있는지는 동전과 아무 관계가 없으므로 A와 B는 H와 독립(independent)이다. 소문항 (ii)와 (iii)에서는 동전을 한 번만 던지고, 두 이메일 모두 같은 쌍둥이, 즉 우리가 이메일을 쓸 때 접속해 있는 쪽이 답한다고 본다. (다른 해석에 대해서는 끝의 참고를 보라.)

$t = ap + bq$라 두자. 이것은 답변 하나가 참일 전체 확률이다.

임의의 답변 기록 E에 대해, $\mathrm{P}(H) = \mathrm{P}(T)$와 베이즈 정리로부터

$$\mathrm{P}(H \mid E) = \frac{\frac{1}{2}\mathrm{P}(E \mid H)}{\frac{1}{2}\mathrm{P}(E \mid H) + \frac{1}{2}\mathrm{P}(E \mid T)} = \frac{\mathrm{P}(E \mid H)}{\mathrm{P}(E \mid H) + \mathrm{P}(E \mid T)}.$$

이다. 이 값이 $\frac{1}{2}$일 필요충분조건은 $\mathrm{P}(E \mid H) = \mathrm{P}(E \mid T)$이다. 이 관찰 하나가 거의 모든 일을 해 준다.

소문항 (i)

답변이 "예"인 사건을 Y라 하자. 동전이 앞면이면 "예"는 참이다. 누가 접속해 있는지에 대해 조건부로 계산하면

$$\mathrm{P}(Y \mid H) = pa + qb = t.$$

동전이 뒷면이면 "예"는 거짓이므로

$$\mathrm{P}(Y \mid T) = p(1-a) + q(1-b) = (p+q) - (ap+bq) = 1 - t.$$

따라서

$$\mathrm{P}(H \mid Y) = \frac{t}{t + (1-t)} = t = ap + bq.$$

그러므로 $\mathrm{P}(H \mid Y) = \frac{1}{2}$일 필요충분조건은 $ap + bq = \frac{1}{2}$, 즉 $2(ap+bq) = 1$이며, 이것이 보이고자 한 것이다.

소문항 (ii)

"예" 다음에 "아니오"가 나오는 사건을 E라 하자. 안나가 접속해 있다고 하자. 안나의 두 답변은 서로 독립이고, 동전이 어느 쪽이 나왔든 둘 중 정확히 하나만 참이다.

$$\mathrm{P}(E \mid H \cap A) = a(1-a) \qquad \text{(참인 "예", 그다음 거짓)},$$
$$\mathrm{P}(E \mid T \cap A) = (1-a)a \qquad \text{(거짓, 그다음 참인 "아니오")}.$$

이 둘은 같다. 마찬가지로 $\mathrm{P}(E \mid H \cap B) = b(1-b) = \mathrm{P}(E \mid T \cap B)$이다. 그러므로

$$\mathrm{P}(E \mid H) = pa(1-a) + qb(1-b) = \mathrm{P}(E \mid T),$$

이고, 앞의 관찰에 의해 p, a, b의 값에 관계없이 $\mathrm{P}(H \mid E) = \frac{1}{2}$이다.

> **생각해 보기** 여기서는 왜 아무 조건 없이도 정확히 $\frac{1}{2}$이 나올까?

서로 모순되는 답변은 동전이 무엇이 나왔든 쌍둥이가 한 번은 참을 말하고 한 번은 거짓을 말했다는 뜻이다. 이 조합은 앞면일 때와 뒷면일 때 가능성이 똑같으므로 동전에 대한 정보를 전혀 담고 있지 않고, 결국 사전확률 $\frac{1}{2}$만 남는다.

소문항 (iii)

이제 "예" 다음에 "예"가 나오는 사건을 E라 하자. 앞면이면 두 답변 모두 참이고, 뒷면이면 둘 다 거짓이다. 누가 접속해 있는지에 대해 조건부로 계산하면

$$\mathrm{P}(E \mid H) = pa^2 + qb^2,$$
$$\mathrm{P}(E \mid T) = p(1-a)^2 + q(1-b)^2.$$

두 식을 빼고 $(1-a)^2 = 1 - 2a + a^2$을 이용하면

$$\begin{aligned} \mathrm{P}(E \mid H) - \mathrm{P}(E \mid T) &= p(a^2 - (1-a)^2) + q(b^2 - (1-b)^2) \\ &= p(2a-1) + q(2b-1) \\ &= 2(ap+bq) - (p+q) \\ &= 2(ap+bq) - 1. \end{aligned}$$

따라서 $2(ap+bq) = 1$이면 두 가능도가 같고 $\mathrm{P}(H \mid E) = \frac{1}{2}$이며, 이것이 보이고자 한 것이다.

(사실 이 계산은 더 많은 것을 보여 준다. 두 가능도의 차가 정확히 $2(ap+bq) - 1$이므로, $\mathrm{P}(H \mid E) = \frac{1}{2}$은 $2(ap+bq) = 1$일 **때, 그리고 그때에만** 성립하며, 이는 소문항 (i)과 같은 조건이다.)

참고: 다른 해석

만약 이메일마다 그때 접속해 있는 쌍둥이가 독립적으로 새로 정해진다고 보면, 각 답변은 독립적으로 확률 t로 참이다. 이 경우 (ii)에서는 $\mathrm{P}(E \mid H) = t(1-t) = \mathrm{P}(E \mid T)$이고, (iii)에서는 $\mathrm{P}(E \mid H) = t^2$, $\mathrm{P}(E \mid T) = (1-t)^2$이며, 이 둘은 $t = \frac{1}{2}$일 때 같다. 결론은 같다.

검산

- 두 쌍둥이가 항상 참을 말한다면$(a = b = 1)$ $t = 1 \neq \frac{1}{2}$이고, (i)에서 "예"라는 답은 앞면을 확실하게 만든다: $\mathrm{P}(H \mid Y) = t = 1$. 이때 (ii)와 같은 모순된 답변은 불가능한데, 이는 $a(1-a) = b(1-b) = 0$과 일치한다.
- $p = 0.3$, $a = 0.8$, $b = \frac{0.5-0.24}{0.7} \approx 0.371$(따라서 $2(ap+bq) = 1$)로 두고 같은 쌍둥이 모형에서 2×10^5회 시뮬레이션(simulation)한 결과 $\mathrm{P}(H \mid \text{예}, \text{예}) \approx 0.496$을 얻었다. 무작위로 고른 수천 개의 (p, a, b)에 대해 분수로 정확히 열거(enumeration)한 결과, 두 해석 모두에서 세 결과가 모두 확인되었다.
- $p = \frac{1}{2}$, $a = b = 0.9$이면 조건이 성립하지 않고, 두 번의 "예"는 $\mathrm{P}(H \mid E) = \frac{81}{82}$을 주어 $\frac{1}{2}$과 거리가 멀다. 따라서 (iii)의 조건은 정말로 중요하다.

흔한 실수

- 근거 없이 $\mathrm{P}(Y \mid T) = 1 - \mathrm{P}(Y \mid H)$라고 쓰는 것. 여기서는 우연히 참이지만, 그것은 뒷면일 때의 "예"가 거짓이기 때문일 뿐이다. 그 이유를 밝혀야 한다.
- (iii)에서 같은 쌍둥이가 두 답변을 모두 한다고 가정하면서, 동시에 두 답변을 각각 확률

t로 참인 독립적인 답변으로 취급하는 것. 쌍둥이가 주어지면 두 답변은 독립이지만, 조건 없이는 독립이 아니다. 둘 다 누가 접속해 있는지에 의존하기 때문이다.

STEP 너머

여기서의 원리는 단 하나, 사후 오즈(posterior odds)는 사전 오즈에 가능도비(likelihood ratio)를 곱한 것과 같다는 것이다. 두 가설 아래에서 가능성이 똑같은 증거는 아무것도 알려 주지 않는다. 같은 쌍둥이 모형은 간단한 **혼합**(mixture) 모형이다. 답변들은 숨은 변수(누가 접속해 있는지)가 주어지면 독립이지만 전체적으로는 종속이며, 이는 은닉 변수(hidden variable) 모형과 베이즈 모형의 핵심에 있는 아이디어이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 바로 이러한 관점에서 조건부 확률과 베이즈 규칙을 훌륭하게 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.