풀이

STEP I 1989, 문제 14

확률과 통계 • 조건부 분포

> **STEP I 1989, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/89-s1-q14

핵심 아이디어

모든 것은 하나의 확률 집합, 곧 씨앗이 폭 1리그(league)인 각 띠에 떨어질 확률에 달려 있다. 씨앗은 공원 지대나 농지에 떨어질 때에만 나무로 자라므로, 나무의 수는 이항분포(binomial distribution)를 따른다. 나무가 자라는 거리 Y는 육지에 떨어졌다는 **조건**(condition) 아래의 L일 뿐이다. 그 확률밀도함수(probability density function)는 f를 육지 띠로 잘라 낸 뒤 전체 넓이가 1이 되도록 늘린 것이다. 마지막 부분은 가설검정(hypothesis test)이다. 이항분포를 따르는 개수로서, 농지에 나무가 23그루 있다는 것은 놀랄 만큼 많은가?

풀이

각 띠의 확률

L의 누적분포함수(cumulative distribution function)는 $0 \le x \le 4$에서

$$\mathrm{F}(x) = \int_0^x \left(\tfrac{1}{2} - \tfrac{1}{8}t\right)\mathrm{d}t = \frac{x}{2} - \frac{x^2}{16},$$

이므로 $\mathrm{F}(1) = \frac{7}{16}$, $\mathrm{F}(2) = \frac{3}{4}$, $\mathrm{F}(3) = \frac{15}{16}$, $\mathrm{F}(4) = 1$이다. 따라서

$$\mathrm{P}(\text{공원}) = \mathrm{P}(0 \le L \le 1) = \tfrac{7}{16}, \qquad \mathrm{P}(\text{호수}) = \tfrac{3}{4} - \tfrac{7}{16} = \tfrac{5}{16},$$
$$\mathrm{P}(\text{농지}) = \mathrm{P}(2 \le L \le 3) = \tfrac{15}{16} - \tfrac{3}{4} = \tfrac{3}{16}, \qquad \mathrm{P}(\text{바다}) = \tfrac{1}{16}.$$

이다. (L은 연속확률변수이므로 경계에서의 확률은 0이고, 따라서 경계가 어느 띠에 속하는지는 상관없다.) 씨앗이 나무로 자랄 확률은

$$p = \tfrac{7}{16} + \tfrac{3}{16} = \tfrac{5}{8}.$$

이다. 한 해에 자라는 나무의 수를 T라 하자. 96개의 씨앗은 서로 독립적으로 날려 가고 각각 확률 $\frac{5}{8}$로 자라므로 $T \sim \mathrm{B}(96, \frac{5}{8})$이고, 해마다 새로 자라는 열병나무(fever tree)의 평균 개수는

$$\mathrm{E}(T) = 96 \times \tfrac{5}{8} = 60.$$

이다.

소문항 (i)

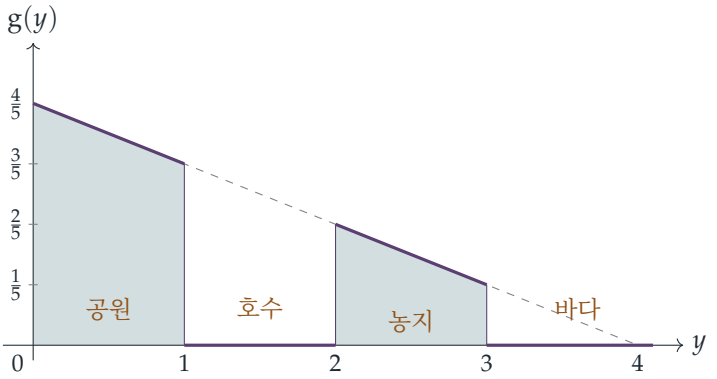
Y는 씨앗이 육지에 떨어졌다는 조건 아래에서 L의 분포를 따른다. 육지를 $A = [0,1] \cup [2,3]$이라 쓰자. $y \in A$에 대하여

$$\mathrm{P}(y \le Y \le y + \delta y) = \frac{\mathrm{P}(y \le L \le y + \delta y)}{\mathrm{P}(L \in A)} \approx \frac{\mathrm{f}(y)\,\delta y}{5/8},$$

이고, Y는 A 바깥의 값을 결코 갖지 않는다. 그러므로 Y의 확률밀도함수는

$$\mathrm{g}(y) = \begin{cases} \frac{8}{5}(\frac{1}{2} - \frac{1}{8}y) = \dfrac{4-y}{5} & \text{조건: } 0 \le y \le 1 \text{ 또는} 2 \le y \le 3, \\ 0 & \text{그 밖의 경우.} \end{cases}$$

이다. 검산으로, 넓이는 $\frac{8}{5}(\frac{7}{16} + \frac{3}{16}) = 1$이다. 그래프는 같은 하강 직선의 두 조각이다. 공원 지대 위에서는 $\frac{4}{5}$에서 $\frac{3}{5}$로 내려가고, 호수 위에서는 아무것도 없으며, 농지 위에서는 $\frac{2}{5}$에서 $\frac{1}{5}$로 내려간다.



점선은 비교를 위해 그린 $\frac{8}{5}\mathrm{f}$이다. Y의 확률밀도함수는 육지 위에서만 이 선을 따른다.

평균을 구하자. $\int y(4-y)\,\mathrm{d}y = 2y^2 - \frac{1}{3}y^3$이므로

$$\begin{aligned} \mathrm{E}(Y) &= \frac{1}{5}\left[2y^2 - \tfrac{1}{3}y^3\right]_0^1 + \frac{1}{5}\left[2y^2 - \tfrac{1}{3}y^3\right]_2^3 \\ &= \frac{1}{5}\left(\frac{5}{3}\right) + \frac{1}{5}\left(9 - \frac{16}{3}\right) = \frac{1}{5}\cdot\frac{5}{3} + \frac{1}{5}\cdot\frac{11}{3} = \frac{16}{15}. \end{aligned}$$

이다. 따라서 새 나무는 평균적으로 성에서 남쪽으로 $\frac{16}{15}$리그 떨어진 곳에서 자란다. 이는 공원 지대를 막 지나 호수 안에 있는 지점인데, 문제될 것은 없다. 평균이 반드시 확률변수가 가질 수 있는 값일 필요는 없기 때문이다.

소문항 (ii)

한 해에 농지에서 자라는 나무의 수를 X라 하자. 전통이 지켜진다면 96개의 씨앗이 날려지고, 각각은 독립적으로 확률 $\frac{3}{16}$로 농지에 떨어지므로

$$X \sim \mathrm{B}\left(96, \tfrac{3}{16}\right), \qquad \mathrm{E}(X) = 18, \qquad \mathrm{Var}(X) = 96 \cdot \tfrac{3}{16} \cdot \tfrac{13}{16} = \tfrac{117}{8} \approx 14.6,$$

이고, 표준편차(standard deviation)는 약 3.82이다.

> **생각해 보기** 정확히 무엇을, 어느 방향으로 검정하는가?

귀무가설(null hypothesis)은 전통이 지켜졌다는 것, 곧 씨앗이 96개여서 $X \sim \mathrm{B}\left(96, \frac{3}{16}\right)$라는 것이다. 농지의 나무가 평소보다 **많다**는 데서 자연스럽게 품게 되는 의심은 96개보다 많은 씨앗이 날려졌다는 것이므로, 단측검정(one-tailed test)을 사용하여 귀무가설 아래에서 23그루 **이상**이 나올 가능성이 얼마나 되는지 묻는다.

$np = 18$과 $n(1-p) = 78$이 모두 충분히 크므로, 연속성 수정(continuity correction)을 적용한 정규근사(normal approximation) $X \approx \mathrm{N}(18,\, 14.625)$를 사용한다.

$$\mathrm{P}(X \geq 23) \approx \mathrm{P}\left(Z \geq \frac{22.5 - 18}{3.824}\right) = \mathrm{P}(Z \geq 1.18) \approx 0.12.$$

(정확한 이항분포 값은 0.121이다.) 따라서 평상시 조건에서 농지의 나무가 23그루 이상인 일은 대략 8년에 한 번꼴로 일어난다. 이는 5% 유의수준(significance level)(심지어 10%)에도 한참 못 미친다. 이 관측 결과는 전통이 지켜졌다는 것과 충분히 잘 들어맞으며, 이 증거만으로는 왕의 의심이 정당하다고 할 수 **없다**.

비교하자면, 5% 수준에서도 왕이 합리적인 근거를 가지려면 개수가 적어도 25는 되어야 한다(정확한 값 $\mathrm{P}(X \geq 25) \approx 0.049$). 왕이 공원 지대의 나무 수까지 알고 있다면 전체 나무 수 $T \sim \mathrm{B}\left(96, \frac{5}{8}\right)$를 검정할 수 있을 것이고, 이는 날려진 씨앗의 수에 대해 훨씬 더 좋은 증거가 될 것이다.

검산

- 네 띠의 확률 $\frac{7}{16}$, $\frac{5}{16}$, $\frac{3}{16}$, $\frac{1}{16}$을 더하면 1이고, Y의 확률밀도함수의 전체 넓이도 1이다.
- 중점 규칙으로 수치적분하면 $\mathrm{E}(Y) = 1.06667$을 얻어 $\frac{16}{15}$와 일치한다. 씨앗 400 000개를 시뮬레이션하면(U가 $[0,1]$에서 균등분포를 따를 때 $L = 4 - 4\sqrt{1-U}$를 사용) 62.5%가 육지에 떨어졌고, 그중 평균 낙하 거리는 1.069였다.

- 이항분포의 확률을 정확히 더하면 $\mathrm{P}(X \geq 23) = 0.1213$이고, 이는 정규근사 추정값 0.120에 가깝다.

흔한 실수

- f 자체를 Y의 확률밀도함수로 사용하는 것. 육지 위에서 그 넓이는 $\frac{5}{8}$밖에 되지 않으므로 $\frac{5}{8}$로 나누어야 한다.
- $\mathrm{E}(L\,;\, L \in A) = \int_A y\,\mathrm{f}(y)\,\mathrm{d}y = \frac{2}{3}$을 계산하고 $\mathrm{P}(L \in A)$로 나누는 것을 잊는 것.
- 꼬리 확률 $\mathrm{P}(X \geq 23)$ 대신 $\mathrm{P}(X = 23)$을 검정하는 것. X의 어떤 한 값이든 그 확률은 꽤 작다.

STEP 너머

소문항 (i)은 **조건부 분포**(conditional distribution)의 예이다. 확률변수를 어떤 사건으로 제한하고 다시 정규화하는 것이다. 같은 조작이 확률밀도함수에 대한 베이즈(Bayes) 정리와 통계학 전반에서 쓰이는 절단분포(truncated distribution)의 바탕에 있다. 소문항 (ii)는 가설검정과 p값을 처음 맛보는 것으로, 대학 1학년 과정에서는 네이만–피어슨(Neyman–Pearson) 체계로 이를 엄밀하게 다룬다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 이와 같은 풀이 이야기를 풍부하게 곁들여 조건부 분포를 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.