풀이

# STEP I 1990, 문제 16

확률과 통계 • 삼각분포와 검정

**STEP I 1990, 16번 문제**

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/90-s1-q16

## 핵심 아이디어

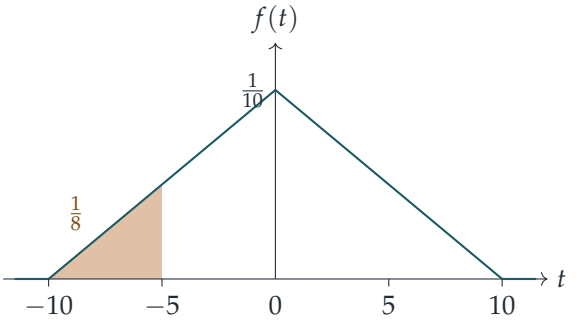
확률밀도함수(density)가 삼각형 모양이므로, 필요한 확률은 모두 작은 삼각형의 넓이이고, 누적분포함수(cumulative distribution function)가 모든 일을 해 준다. 두 번째로 주목할 점은 오전 8시 무렵에 올 수 있는 버스는 8시 버스뿐이라는 것이다. 7시 버스는 7.10까지 이미 떠났고, 9시 버스는 8.50 이전에는 올 수 없다. 따라서 아침에 "버스를 탄다"는 것은 사실상 "8시 버스가 오기 전에 정류장에 도착한다"는 뜻이다. 그러면 마지막 부분은 모수가 하나인 이항 모형(binomial model)과 유의성 검정(significance test)의 문제가 된다.

## 풀이

### $T$의 확률밀도함수

확률밀도함수는 $t = -10$에서 0으로부터 선형으로 증가하여 $t = 0$에서 최댓값에 이르고, 같은 비율로 감소하여 $t = 10$에서 0이 된다. 따라서 $[-10, 10]$에서 $f(t) = k(10 - |t|)$ 꼴이고 그 밖에서는 0이다. 전체 넓이는 밑변이 20이고 높이가 $10k$인 삼각형의 넓이이므로 $\frac{1}{2} \cdot 20 \cdot 10k = 1$이고, $k = \frac{1}{100}$을 얻는다.

$$f(t) = \begin{cases} \dfrac{10+t}{100}, & -10 \le t \le 0, \\ \dfrac{10-t}{100}, & 0 \le t \le 10, \\ 0, & \text{그 외.} \end{cases}$$




나중에 쓰기 위해 구해 두면, 왼쪽 절반에서 누적분포함수는 밑변이 $a+10$이고 높이가 $f(a)$인 삼각형의 넓이이다.

$$\mathrm{P}(T<a)=\tfrac{1}{2}(a+10)\cdot\frac{a+10}{100}=\frac{(a+10)^2}{200},\qquad -10\le a\le 0,$$

또 $t=0$에 대한 대칭성에 의해 $0\le a\le 10$일 때 $\mathrm{P}(T>a)=\dfrac{(10-a)^2}{200}$이다. 그림에서 색칠한 부분의 넓이는 $\mathrm{P}(T<-5)=\frac{25}{200}=\frac{1}{8}$이다.

### 오전 8.10까지 버스 타기

버스가 올 때까지 정류장에서 기다린다고 가정한다. 8시에 오기로 된 버스는 오전 8시로부터 $T$분 뒤에 도착하며, $-10\le T\le 10$이다. 그 앞 버스는 7.10까지 이미 떠났고, 다음 버스는 8.50 이전에는 올 수 없다. 따라서 8.10까지 버스를 타는 것은 정확히 8시 버스가 도착하기 전에 정류장에 도착하는 경우이며, 이때 8.10이라는 시한은 자동으로 만족된다.

오전 8시로부터 $a$분 뒤에 정류장에 도착한다면, $T>a$일 때 버스를 탄다($T=a$일 확률은 0이다). 세 경우는 다음과 같다.

$$\begin{aligned}
7{\cdot}55,\ a=-5:&\quad \mathrm{P}(T>-5)=1-\tfrac{25}{200}=\tfrac{7}{8},\\
7{\cdot}58,\ a=-2:&\quad \mathrm{P}(T>-2)=1-\tfrac{64}{200}=\tfrac{17}{25},\\
8{\cdot}02,\ a=2:&\quad \mathrm{P}(T>2)=\tfrac{64}{200}=\tfrac{8}{25}.
\end{aligned}$$

정류장에 7.55에 도착할 확률은 0.5, 7.58에 도착할 확률은 $0.5\times 0.8=0.4$, 8.02에 도착할 확률은 $0.5\times 0.2=0.1$이다. 내가 늦잠을 자는 것은 버스와 아무 관계가 없으므로, 전확률 법칙(law of total probability)에 의해

$$\begin{aligned}
\mathrm{P}(8.10\text{까지 버스를 탐}) &= 0.5\cdot\tfrac{7}{8}+0.4\cdot\tfrac{17}{25}+0.1\cdot\tfrac{8}{25}\\
&= 0.4375+0.272+0.032=\frac{1483}{2000}=0.7415.
\end{aligned}$$

### 새 시계가 틀렸을까?

먼저 시계가 정확하여 매일 아침 7:55에 정류장에 도착한다고 하자. 8시 버스와 9시 버스의 도착 시각을 정시로부터 몇 분 뒤인지로 나타낸 것을 각각 $T_8$, $T_9$라 하자. 이 둘은 독립(independent)이고, 각각 확률밀도함수 $f$를 가진다. 오전 9시가 지나서도 기다리고 있는 것은 정확히 8시 버스를 놓치고($T_8 < -5$) 그리고 9시 버스가 늦게 오는($T_9 > 0$) 경우이다. 따라서 어느 하루 아침에

$$p = \mathrm{P}(T_8 < -5)\,\mathrm{P}(T_9 > 0) = \tfrac{1}{8} \cdot \tfrac{1}{2} = \tfrac{1}{16}.$$

> **생각해 보기** 검정은 어느 방향으로 해야 할까? 5번은 "너무 많은" 것일까, "너무 적은" 것일까?

시계가 잘못 맞춰져 있으면 $p$는 어느 쪽으로든 움직일 수 있다. 시계가 $d$분 빠르다고 하자. 그러면 나는 실제 시각 $7.55 - d$에 정류장에 도착하고, 시계는 실제 시각 $9.00 - d$에 9:00을 가리키므로, "오전 9시"가 지나서도 기다릴 확률은

$$g(d) = \mathrm{P}(T_8 < -5 - d)\,\mathrm{P}(T_9 > -d).$$

가 된다. 시계가 빠르면($d > 0$) 더 일찍 도착하므로 8시 버스를 놓치는 일이 드물어 $g(d) < \frac{1}{16}$이다. 시계가 조금 느리면 반대가 된다(예를 들어 $g(-2.5) \approx 0.079$). 따라서 정직한 검정은 양측 검정(two-sided test)이다.

새 시계로 100일의 아침 가운데 오전 9시가 지나서도 기다린 날의 수를 $X$라 하자. 서로 다른 버스의 도착 시각은 독립이고 나는 매일 같은 시각에 도착하므로, 각 아침은 독립 시행(trial)이며

$$H_0:\ \text{시계가 정확함}, \quad X \sim \mathrm{B}(100, \tfrac{1}{16}); \qquad H_1:\ \text{시계가 잘못 맞춰짐, 따라서} p \neq \tfrac{1}{16}.$$

$H_0$ 아래에서 기대 횟수는 $\mathrm{E}(X) = 6.25$이고 $\mathrm{Var}(X) = 100 \cdot \frac{1}{16} \cdot \frac{15}{16} \approx 5.86$이므로, 표준편차(standard deviation)는 약 2.42이다. 관측값 5는 평균보다 조금 작다. 정확히 계산하면

$$\mathrm{P}(X \leq 5) = \sum_{k=0}^{5} \binom{100}{k} p^k (1-p)^{100-k} \approx 0.40 \qquad \text{여기서} p = \tfrac{1}{16},$$

이고, 연속성 보정(continuity correction)을 한 정규근사(normal approximation)는 $\Phi((5.5 - 6.25)/2.42) = \Phi(-0.31) \approx 0.38$을 준다. 이는 0.025(유의수준 5% 양측 검정의 아래쪽 꼬리)는 물론이고 통상적인 어떤 유의수준(significance level)과도 거리가 멀다. 따라서 $H_0$를 기각하지 않는다. **새 시계가 잘못 맞춰져 있다는 증거는 없다.**

이것이 시계가 정확하다는 증거와 같은 것은 아니다. 실제로 확률밀도함수가 대칭이므로(두 버스의 역할을 맞바꾸면) $g(d) = g(-5 - d)$이고, 따라서 5분 느린 시계는 정확한 시계와 똑같이 $p = \frac{1}{16}$을 준다. 이런 종류의 자료로는 아침을 아무리 많이 관찰해도 이 두 경우를 구별할 수 없다.

## 검산

- 확률밀도함수의 넓이는 1이고, 삼각형 공식은 $f$의 수치적분 결과와 일치한다. 예를 들어 $\mathrm{P}(T < -2) = 0.32$이다.
- 답 0.7415는 최선의 경우 $\frac{7}{8}$(항상 7.55에 도착)과 최악의 경우 0.32(항상 8.02에 도착) 사이에 있으며, 마땅히 그래야 한다. $10^6$일의 아침을 시뮬레이션한 결과는 0.741이었다.
- $10^6$개의 쌍 $(T_8, T_9)$을 시뮬레이션한 결과는 $\frac{1}{16} = 0.0625$에 대해 0.0622였고, 정확한 이항분포 꼬리확률 $\mathrm{P}(X \leq 5)$의 값은 0.4002이다.

## 흔한 실수

- 9시가 지나서도 기다리려면 8시 버스를 놓쳐야 한다는 점을 잊는 것. 답 $p = \frac{1}{2}$(단지 "9시 버스가 늦게 온다")는 첫 번째 인수를 무시한 것이다.
- "8.10까지 버스를 탄다"를 $T \leq 10$을 따로 확인해야 하는 것으로 다루는 것. 8시 버스는 항상 8.10까지 도착하므로 $T > a$만 중요하다.
- 자료가 오히려 조금 적게 나왔는데 "너무 많다"에 대한 단측 검정(one-tailed test)을 하는 것. 잘못된 시계는 $p$를 어느 방향으로든 움직일 수 있다.

## STEP 너머

삼각형 모양의 확률밀도함수는 정확히 서로 독립인 두 균등 확률변수(uniform random variable)의 합의 분포이며, 이는 합성곱(convolution)의 첫 번째 예이다. 대학 1학년 확률론 강의에서는 이 방법으로 합의 확률밀도함수를 구한다. 마지막 부분은 통계적 식별 가능성(identifiability)을 조금 맛보게 해 준다. 대칭성 $g(d) = g(-5 - d)$는 서로 다른 두 모수 값이 똑같은 자료를 만들어 낸다는 뜻이므로, 어떤 검정도 이 둘을 구별할 수 없다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 합성곱과 확률변수의 합을 명쾌하게 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.