풀이

STEP II 2000, 문제 12

확률과 통계 • 정규확률변수의 합

> **STEP II 2000, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/00-s2-q12

핵심 아이디어

헨리(Henry)가 지각하는 것은 정확히 **전체** 이동 시간이 09시 전까지 남은 시간을 넘을 때이다. 따라서 각 경로마다 서로 독립(independent)인 두 정규분포(normal distribution) 이동 시간을 더해 하나의 정규확률변수로 만든다. 표준화(standardising)하면 "지각"은 "표준정규변수가 어떤 수보다 크다"는 사건이 되고, Φ가 증가함수이므로 표 없이도 두 경로를 비교할 수 있다. 나머지는 전확률 법칙(law of total probability)과 베이즈 정리(Bayes' theorem)이다.

풀이

두 경로 비교하기

각 경로에서 기차 이동과 그 이후의 이동은 서로 독립이라고 가정한다.

킹스크로스(King's Cross). 기차 이동 시간을 $X_1 \sim N(55, 25)$, 그 이후의 이동 시간을 $Y_1 \sim N(30, 144)$라 하자. 두 변수는 독립인 정규확률변수이므로 그 합도 정규분포를 따르며, 평균과 분산(variance)은 각각 더해진다.

$$T_1 = X_1 + Y_1 \sim N(85,\ 169), \qquad \text{표준편차} 13.$$

기차는 07시 15분에 출발하므로 헨리에게는 105분이 있고, $T_1 > 105$일 때 지각한다. $Z \sim N(0,1)$이라 하면

$$\mathrm{P}(\text{지각} \mid \text{킹스크로스}) = \mathrm{P}\left(Z > \frac{105-85}{13}\right) = 1 - \Phi\left(\tfrac{20}{13}\right) = A.$$

리버풀 스트리트(Liverpool Street). 마찬가지로 $X_2 \sim N(65, 16)$와 $Y_2 \sim N(25, 9)$가 독립

일 때 $T_2 = X_2 + Y_2$이므로

$$T_2 \sim N(90,\, 25), \qquad \text{표준편차5.}$$

이 기차는 07시 20분에 출발하여 100분이 남으므로

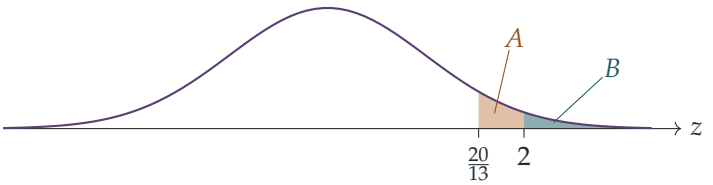
$$\mathrm{P}(\text{지각} \mid \text{리버풀 스트리트}) = \mathrm{P}\left(Z > \frac{100-90}{5}\right) = 1 - \Phi(2) = B.$$

생각해 보기 표 없이 A와 B를 어떻게 비교할 수 있을까?

Φ는 양의 확률밀도(density)를 갖는 누적분포함수(cumulative distribution function)이므로 순증가한다. $\frac{20}{13} < \frac{26}{13} = 2$이므로 $\Phi(\frac{20}{13}) < \Phi(2)$이고, 따라서

$$A = 1 - \Phi\left(\tfrac{20}{13}\right) > 1 - \Phi(2) = B.$$

그러므로 요구한 대로, 헨리는 킹스크로스행 기차를 탈 때 지각할 가능성이 더 크다. (흥미롭게도 이 경로는 평균적으로 **더 빠르다**. 그런데도 불리한 까닭은 킹스크로스에서 이어지는 지하철 구간의 변동이 매우 크기 때문이다.)



표준정규곡선의 색칠된 꼬리 부분: A는 색칠된 영역 전체이고, B는 2 너머의 더 진한 부분이다.

지각 횟수의 기댓값

한 번의 이동에서 헨리가 킹스크로스행 기차를 타는 사건을 K라 하여 $\mathrm{P}(K) = p$라 하고, 지각하는 사건을 D라 하자. 전확률 법칙에 의해

$$\mathrm{P}(D) = \mathrm{P}(D \mid K)\,\mathrm{P}(K) + \mathrm{P}(D \mid K')\,\mathrm{P}(K') = pA + (1-p)B.$$

M번의 이동이 서로 독립이라고 가정하면 지각 횟수는 모수가 M과 $\mathrm{P}(D)$인 이항분포(binomial distribution)를 따르므로

$$L = M(pA + (1-p)B) = M(B + p(A-B)).$$

$A - B > 0$이고 $0 \le p \le 1$이므로 식 $B + p(A-B)$는 p에 대해 증가하며, $p = 0$일 때 B에서 $p = 1$일 때 A까지 변한다. 따라서 가능한 모든 p에 대하여

$$BM \le L \le AM,$$

이고, 왼쪽 등호는 $p = 0$일 때만, 오른쪽 등호는 $p = 1$일 때만 성립한다. 말로 하면, L/M은 A와 B의 가중평균(weighted average)이므로 둘 사이에 있어야 한다.

p 추정하기

지각한 경우 가운데 킹스크로스행 기차를 탔던 경우의 비율은 $\mathrm{P}(K \mid D)$의 추정값이며, M이 크므로 관측된 $\frac{3}{5}$를 이 확률과 같다고 본다. 베이즈 정리에 의해

$$\mathrm{P}(K \mid D) = \frac{\mathrm{P}(D \mid K)\,\mathrm{P}(K)}{\mathrm{P}(D)} = \frac{pA}{pA + (1-p)B}.$$

이것을 $\frac{3}{5}$와 같다고 놓으면

$$\begin{aligned} 5pA &= 3pA + 3(1-p)B \\ 2pA + 3pB &= 3B, \end{aligned}$$

이므로 추정값은

$$\hat{p} = \frac{3B}{2A + 3B}.$$

검산

- 수치로는 $A \approx 0.0620$, $B \approx 0.0228$이므로 킹스크로스 경로는 약 6%, 리버풀 스트리트 경로는 약 2%의 확률로 지각한다. 각 경로에서 4×10^5번의 이동을 모의실험한 결과는 소수점 아래 셋째 자리까지 일치했다.
- 이 값들을 쓰면 $\hat{p} \approx 0.355$이다. 다시 대입하면 $pA/(pA + (1-p)B) = 0.6$이 정확히 성립하며, 이 p로 전체 과정을 모의실험하면 지각의 약 60%가 킹스크로스 경로에서 나왔다.
- 마지막 답의 타당성 점검: 두 경로의 신뢰도가 같다면($A = B$) 지각은 기차 선택만을 반영하므로, 지각의 $\frac{3}{5}$가 킹스크로스 경로라는 것은 $p = \frac{3}{5}$를 뜻한다. 공식은 $3B/5B = \frac{3}{5}$를 준다. 킹스크로스는 신뢰도가 더 낮은 경로이므로 지각한 날 가운데 실제보다 많이 나타나고, 따라서 p는 $\frac{3}{5}$보다 작아야 하는데, 실제로 그렇다.

흔한 실수

- 분산 대신 표준편차(standard deviation)를 더하는 실수: 킹스크로스 경로의 전체 시간의 분산은 $25 + 144 = 169$이므로 표준편차는 $5 + 12 = 17$이 아니라 13이다.
- 두 기차 모두에 똑같이 105분의 여유를 적용하는 실수: 리버풀 스트리트행 기차는 5분 늦게 출발한다.

- 곧바로 $p = \frac{3}{5}$로 놓는 실수. $\frac{3}{5}$는 $\mathrm{P}(K)$가 아니라 $\mathrm{P}(K \mid D)$이며, 두 경로의 신뢰도가 같지 않으므로 둘은 다르다.

STEP 너머

여기 나온 두 아이디어는 대학 첫 확률론 강의 내내 되풀이해 등장한다. 독립인 정규확률변수의 합이 정규분포를 따른다는 사실은 적률생성함수(moment generating function)로 가장 깔끔하게 증명되며, 이는 중심극한정리(central limit theorem)의 바탕이 되는 정규분포족의 안정성의 특수한 경우이다. 마지막 소문항은 축소판 통계적 추론(statistical inference)이다. 우리는 결과(지각)를 관측하고 베이즈 정리를 이용해 숨은 원인(기차 선택)을 추론했다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 두 주제를 모두 다루며, Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에는 바로 이런 유형의 베이즈 문제가 훨씬 많이 실려 있다.

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.