풀이

STEP I 견본 문제지, 문제 15

확률과 통계 • 확률변수의 합

> **STEP I 견본 문제지, 15번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/spec-s1-q15

핵심 아이디어

C의 낯설어 보이는 분포(distribution)는, 가격의 마지막 자리 숫자가 $0, 1, \ldots, 9$ 중 어느 것이든 같은 확률로 나온다고 할 때, 가격 하나를 가장 가까운 10펜스 단위로 반올림하면서 생기는 오차와 정확히 같다. 그러면 쇼핑객의 전체 오차는 C의 독립(independent)인 복사본 36개의 합이므로 그 평균과 분산을 곧바로 구할 수 있고, 20펜스의 차이가 얼마나 놀라운 일인지는 중심극한정리(Central Limit Theorem)가 알려 준다.

풀이

C의 기댓값과 분산

모든 i에 대하여 $\mathrm{P}(C = i) = \mathrm{P}(C = -i)$이므로 분포는 0에 대하여 대칭이다. 따라서 항 $i\,\mathrm{P}(C = i)$들은 둘씩 짝지어 소거되고,

$$\mathrm{E}(C) = 0.$$

이다. (확률의 합은 실제로 1이다: $9 \cdot \frac{1}{10} + 2 \cdot \frac{1}{20} = 1$.) 평균이 0이므로 분산(variance)은 그냥 $\mathrm{E}(C^2)$이다:

$$\begin{aligned}\mathrm{Var}(C) = \mathrm{E}(C^2) &= 2 \cdot \frac{25}{20} + \frac{2}{10}(1^2 + 2^2 + 3^2 + 4^2) \\ &= \frac{5}{2} + \frac{2 \cdot 30}{10} = \frac{17}{2}.\end{aligned}$$

쇼핑객의 어림셈

모형. 물건 하나에 대하여, 실제 가격에서 반올림한 가격을 뺀 값을 펜스 단위로 C라 하자. 이 값은 가격의 마지막 자리 숫자 d에만 의존한다. d가 $0, 1, \ldots, 9$ 중 어느 것이든 같은 확률로 나온다고 가정한다. 여기서 "무작위로" 산다는 말은 이런 뜻이어야 하지만, 이것은 실제 가격에 대한 하나의 가정이다.

- $d = 0$: 반올림이 일어나지 않으므로 $C = 0$이다.
- $d = 1, 2, 3, 4$: 가격이 d만큼 내려가므로 $C = d$이다.
- $d = 6, 7, 8, 9$: 가격이 $10 - d$만큼 올라가므로 $C = -(10 - d)$, 곧 $-4, -3, -2, -1$이다.
- $d = 5$: 가격이 같은 확률로 내려가거나 올라가므로 $C = 5$ 또는 -5이고, 각각의 확률은 $\frac{1}{2} \cdot \frac{1}{10} = \frac{1}{20}$이다.

$-4, \ldots, 4$의 각 값은 정확히 한 숫자에서 확률 $\frac{1}{10}$로 나오고, ± 5는 각각 확률 $\frac{1}{20}$을 가진다. 이것은 바로 첫 부분의 분포이며, 그래서 처음에 그렇게 설정해 둔 것이다.

전체 오차. 36개 물건의 오차를 $C_1, \ldots, C_{36}$이라 하고,

$$S = C_1 + C_2 + \cdots + C_{36}$$

을 실제 청구액에서 쇼핑객의 어림값을 뺀 값이라 하자. 물건을 무작위로 고르므로 C_j들은 서로 독립이고 각각 C와 같은 분포를 따른다고 본다. 그러면

$$\mathrm{E}(S) = 36 \cdot 0 = 0, \qquad \mathrm{Var}(S) = 36 \cdot \tfrac{17}{2} = 306,$$

이다. 분산을 구할 때 독립성을 사용하였다. 따라서 S의 표준편차(standard deviation)는 $\sqrt{306} \approx 17.5$펜스이다.

> ? **생각해 보기** 20펜스는 큰 차이인가, 작은 차이인가?

이 값은 반올림 오차 하나의 크기가 아니라 표준편차와 비교해야 한다. 계산원의 청구액이 어림값보다 20펜스 많으므로 $S = 20$이다. 이는 평균보다

$$\frac{20}{\sqrt{306}} \approx 1.14$$

표준편차만큼 위에 있다. S는 독립이고 같은 분포를 따르는 변수 36개의 합이므로, 중심극한정리에 따라 S는 근사적으로 $\mathrm{N}(0, 306)$을 따른다. (S가 정숫값을 가지므로) 연속성 수정(continuity correction)을 하면

$$\mathrm{P}(S \geq 20) \approx \mathrm{P}\left(Z \geq \frac{19.5}{\sqrt{306}}\right) = \mathrm{P}(Z \geq 1.11) \approx 0.13,$$

이다. 여기서 $Z \sim \mathrm{N}(0,1)$이다. (어느 방향으로든) 적어도 이만큼 큰 오차는 약 26%의 경우에 일어난다.

따라서 쇼핑객은 실수를 의심할 이유가 **없다**. 20펜스의 차이는 반올림 오차의 정상적인 퍼짐 안에, 약 1 표준편차 안에 넉넉히 들어 있으며, 대략 장보기 일곱 번이나 여덟 번에 한 번꼴로 일어난다. 약 2 표준편차보다 큰 차이, 이를테면 35펜스 이상이어야 비로소 실제로 의심할 근거가 된다.

검산

- 가격을 가장 가까운 10펜스 단위로 반올림하면 많아야 5펜스가 바뀌므로, 물건 36개에 대한 최악의 경우는 180펜스이다. 전형적인 오차 약 17펜스는 이보다 훨씬 작은데, 당연히 그래야 한다. 양의 오차와 음의 오차가 대부분 서로 상쇄되기 때문이다.
- 컴퓨터로 확인한 결과, 반올림 오차는 정확히 C의 분포를 따르고 $\mathrm{E}(C) = 0$, $\mathrm{Var}(C) = \frac{17}{2}$이다. 이어서 분수를 써서 36개 복사본의 합성곱(convolution)을 정확히 계산하였다: $\mathrm{Var}(S) = 306$이고 $\mathrm{P}(S \geq 20) \approx 0.1330$이며, 정규근사(normal approximation)로는 0.1325, 2×10^5번의 장보기 모의실험으로는 0.132였다.

흔한 실수

- $\mathrm{Var}(S)$를 $36^2 \mathrm{Var}(C)$로 잡는 것. 이것은 하나의 오차가 36번 반복된 $36C$의 분산이다. 독립인 오차들의 분산은 그냥 더해진다.
- S가 정확히 20일 확률(약 0.012)을 놀라움의 척도로 삼는 것. S의 개별 값은 어느 것이나 꽤 일어나기 어렵다. 중요한 것은 관찰된 오차 이상으로 큰 오차가 나올 확률이다.
- 무엇을 가정하는지 밝히는 것을 잊는 것. 곧 가격의 마지막 자리 숫자가 균등분포(uniformly distributed)를 따르고 물건들의 오차가 서로 독립이라는 가정이다.

STEP 너머

이것이 가설검정(hypothesis test)의 바탕에 있는 추론이다. 아무 잘못도 없다고 가정하고(여기서는 S가 단지 반올림 오차라고 가정하고), 관찰된 결과 이상으로 극단적인 결과가 나올 가능성이 얼마나 되는지 묻는다. 반올림 오차에 대한 같은 분석이 수치해석(numerical analysis)에서도 쓰이는데, 여기서 많은 수를 반올림할 때 생기는 오차는 보통 수의 개수에 정비례하지 않고 그 제곱근처럼 커진다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 확률변수(random variable)의 합과 중심극한정리를 명료하게 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.