풀이

STEP I 1997, 문제 12

확률과 통계 • 중심극한정리

> **STEP I 1997, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/97-s1-q12

핵심 아이디어

이차방정식(quadratic equation)을 직접 풀어서 X를 T의 간단한 함수로 나타낸다. 그러면 X에 관한 모든 질문은 분포를 완전히 알고 있는 T에 관한 질문으로 바뀐다. 깔끔하게 단순화할 수 있는 점은 $1-T$ 역시 $[0,1]$에서 균등분포(uniform distribution)를 따른다는 것이다. 마지막 부분은 서로 독립(independent)인 여러 복사본의 합에 관한 것이므로, 중심극한정리(central limit theorem)를 사용하라는 신호이다.

풀이

X의 분포

근의 공식(quadratic formula)에 의해 $x^2+2x+T=0$의 근은 $x=-1\pm\sqrt{1-T}$이다. $0\le T\le 1$이므로 이 근들은 실수이고, 더 큰 근은 플러스 부호를 택한 것이다.

$$X=-1+\sqrt{1-T}.$$

T가 $[0,1]$ 전체를 움직이면 X는 $[-1,0]$ 전체를 움직인다.

$U=1-T$라 하자. T가 $[0,1]$에서 균등분포를 따르므로 U도 그렇다. 구간을 중점에 대해 뒤집어도 균등한 밀도는 바뀌지 않기 때문이다. 따라서 $U\sim \mathrm{U}[0,1]$에 대하여 $X=\sqrt{U}-1$이다.

$X>-\frac{1}{3}$일 확률

$X+1=\sqrt{U}$이고 양변이 모두 음이 아니므로

$$X>-\tfrac{1}{3} \iff \sqrt{U}>\tfrac{2}{3} \iff U>\tfrac{4}{9}.$$

양변이 모두 음이 아니므로 제곱해도 부등식(inequality)의 방향은 유지된다. 따라서

$$\mathrm{P}\left(X > -\tfrac{1}{3}\right) = \mathrm{P}\left(U > \tfrac{4}{9}\right) = 1 - \tfrac{4}{9} = \frac{5}{9}.$$

(같은 말로, $T < \frac{5}{9}$이다.)

평균과 분산

$U \sim \mathrm{U}[0,1]$의 밀도는 $[0,1]$에서 1이므로

$$\mathrm{E}(\sqrt{U}) = \int_0^1 u^{1/2}\,\mathrm{d}u = \frac{2}{3}, \qquad \mathrm{E}((\sqrt{U})^2) = \mathrm{E}(U) = \frac{1}{2}.$$

그러므로 평균(mean)은

$$\mathrm{E}(X) = \mathrm{E}(\sqrt{U}) - 1 = \frac{2}{3} - 1 = -\frac{1}{3}.$$

상수를 빼도 분산(variance)은 바뀌지 않으므로

$$\mathrm{Var}(X) = \mathrm{Var}(\sqrt{U}) = \mathrm{E}(U) - (\mathrm{E}\sqrt{U})^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18},$$

이고, 이것이 구하려던 결과이다.

생각해 보기 이 적률(moment)들을 구하는 데 X의 확률밀도함수(probability density function)가 필요한가?

필요하지 않다. 임의의 함수 g에 대하여 $\mathrm{E}(g(U)) = \int_0^1 g(u)\,\mathrm{d}u$이므로 U를 가지고 직접 계산할 수 있다. 그래도 밀도를 구하고 싶다면, $-1 \le x \le 0$일 때 $\mathrm{P}(X \le x) = \mathrm{P}(U \le (x+1)^2) = (x+1)^2$이므로 X는 $[-1,0]$에서 밀도 $2(x+1)$을 가진다. 이 구간에서 $x \cdot 2(x+1)$을 적분해도 $-\frac{1}{3}$을 얻는다.

800개 근의 합

각 시행이 독립이므로 X_i들도 독립이고, 각각의 평균은 $-\frac{1}{3}$, 분산은 $\frac{1}{18}$이다. 따라서

$$\mathrm{E}(Y) = 800 \cdot \left(-\tfrac{1}{3}\right) = -\frac{800}{3}, \qquad \mathrm{Var}(Y) = 800 \cdot \frac{1}{18} = \frac{400}{9},$$

이며, 여기서 분산을 더하는 단계에 독립성이 쓰인다. Y의 표준편차(standard deviation)는 $\frac{20}{3}$이다.

800은 충분히 크므로, 중심극한정리에 의해 Y는 근사적으로 $\mathrm{N}\left(-\frac{800}{3}, \frac{400}{9}\right)$을 따른다. 구하려는 것은

$$\mathrm{P}(Y \le K) \approx \Phi\left(\frac{K + \frac{800}{3}}{20/3}\right) = 0.08.$$

표에서 $\Phi(1.405) \approx 0.92$이므로, 대칭성에 의해 $\Phi(-1.405) \approx 0.08$이다. 따라서

$$\frac{K + \frac{800}{3}}{20/3} \approx -1.405,$$
$$K \approx -\frac{800}{3} - 1.405 \times \frac{20}{3} \approx -266.67 - 9.37 \approx -276.$$

? 생각해 보기 연속성 수정이 필요하지 않은가?

필요하지 않다. 각 X_i가 연속확률변수(continuous random variable)이므로 Y도 연속확률변수이고, 따라서 연속성 수정(continuity correction)은 필요 없다. 연속성 수정은 이항분포(binomial distribution)와 같은 이산분포를 정규분포로 근사할 때에만 쓴다.

검산

- 근이 맞는지 확인한다. $T = 0$일 때 방정식은 $x(x+2) = 0$이고, 더 큰 근은 $0 = -1 + \sqrt{1}$이다.
- $\mathrm{E}(X) = -\frac{1}{3}$은 $[-1, 0]$ 안에 있고, 밀도 $2(x+1)$은 0 쪽으로 기울어져 있으므로 평균이 중점 $-\frac{1}{2}$보다 큰 것이 타당하다.
- T의 값 400 000개로 시뮬레이션한 결과 $\mathrm{P}(X > -\frac{1}{3}) \approx 0.555$, 평균 ≈ -0.3337, 분산 ≈ 0.0556을 얻었다. Y를 직접 4000번 시뮬레이션한 결과 $\mathrm{P}(Y \leq -276.03) \approx 0.083$을 얻었으며, 이는 0.08과 잘 맞는다.

흔한 실수

- 근의 공식에서 마이너스 부호를 택하는 것. 그러면 더 작은 근이 나온다.
- $\mathrm{Var}(Y) = 800^2 \mathrm{Var}(X)$라고 쓰는 것. 이것은 서로 독립인 800개 복사본의 합의 분산이 아니라 $800X_1$의 분산이다.
- -1.405 대신 $+1.405$를 쓰는 것. 확률 0.08은 아래쪽 꼬리에 해당하므로 K는 평균보다 작아야 한다.

STEP 너머

균등분포를 따르는 U에 대하여 $X = g(U)$로 쓰는 기법은 거꾸로도 작동한다. F가 임의의 연속인 분포함수(distribution function)이면 $F^{-1}(U)$의 분포함수는 F이다. 이 **역변환법**(inverse transform method)은 컴퓨터가 균등 난수로부터 확률변수를 시뮬레이션하는 방법이다. 여기서 블랙박스처럼 사용한 중심극한정리는 대학의 첫 확률론 강의에서 적률생

성함수(moment generating function)나 특성함수(characteristic function)를 이용해 증명한다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문서)이 두 아이디어를 모두 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.