풀이

STEP III 2016, 문제 12

확률과 통계 • 체비쇼프 부등식

> **STEP III 2016, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/16-s3-q12

핵심 아이디어

체비쇼프(Chebyshev) 부등식(Chebyshev's inequality)은 평균에서 **멀리** 떨어질 확률의 상한을 준다. 따라서 어떤 값이 평균에 **가깝다**는 명제는 여사건(complement)에서 얻는다. 각 소문항에서 할 일은 알맞은 확률변수(random variable)를 찾고, 그 평균과 표준편차(standard deviation)를 구한 다음, $k\sigma$가 우리가 관심을 두는 거리와 정확히 같아지도록 k를 고르는 것이다. 소문항 (ii)에서 e^n 옆에 놓인 합 $\sum n^r/r!$은 푸아송 분포(Poisson distribution)의 흔적이다.

풀이

소문항 (i)

$100n$번 던졌을 때 앞면이 나온 횟수를 X라 하자. 각 시행이 독립(independent)이고 앞면이 나올 확률이 각각 0.2라고 가정하면 $X \sim \mathrm{B}(100n, 0.2)$이므로

$$\mu = 100n \times 0.2 = 20n, \qquad \sigma^2 = 100n \times 0.2 \times 0.8 = 16n, \qquad \sigma = 4\sqrt{n}.$$

사건 $16n \le X \le 24n$은 정확히 $|X - 20n| \le 4n$이므로

$$\alpha = \mathrm{P}(|X - \mu| \le 4n) = 1 - \mathrm{P}(|X - \mu| > 4n).$$

$k\sigma = 4n$, 즉 $k \cdot 4\sqrt{n} = 4n$이 되기를 원하므로 $k = \sqrt{n}$이고, 이 값은 양수이다. 체비쇼프 부등식에 의해

$$\mathrm{P}(|X - \mu| > 4n) = \mathrm{P}(|X - \mu| > \sqrt{n}\,\sigma) \le \frac{1}{(\sqrt{n})^2} = \frac{1}{n},$$

이고, 따라서

$$\alpha \ge 1 - \frac{1}{n},$$

이 되어 원하는 결과를 얻는다. (조건 $n > 1$은 이 부등식이 의미 있는 내용을 담도록 할 뿐이다. $n = 1$이면 부등식은 $\alpha \geq 0$이 된다.)

소문항 (ii)

> **생각해 보기** 확률이 $n^r/r!$ 꼴로 생긴 확률변수는 무엇인가?

$Y \sim \mathrm{Po}(n)$이면 $r = 0, 1, 2, \ldots$에 대하여 $\mathrm{P}(Y = r) = \mathrm{e}^{-n} n^r / r!$이므로

$$\mathrm{e}^{-n}\left(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!}\right) = \mathrm{P}(Y \leq 2n).$$

따라서 증명할 부등식은 $\mathrm{e}^n > 0$으로 나누면 다음과 동치이다.

$$\mathrm{P}(Y \leq 2n) \geq 1 - \frac{1}{n}.$$

푸아송 분포의 평균과 분산(variance)은 모두 n이므로 $\mu = n, \sigma = \sqrt{n}$이다.

> **생각해 보기** 체비쇼프 부등식은 양쪽 꼬리를 다루지만 $Y \leq 2n$은 한쪽 사건이다. 둘을 어떻게 맞출 수 있는가?

$Y \geq 0$이므로 항상 $Y - n \geq -n$이고, 따라서 $Y - n < -n$은 일어날 수 없다. 그러므로

$$|Y - n| > n \iff Y - n > n \iff Y > 2n.$$

$k\sigma = n$이 되도록 $k = \sqrt{n}$으로 잡으면 체비쇼프 부등식에 의해

$$\mathrm{P}(Y > 2n) = \mathrm{P}(|Y - n| > \sqrt{n}\,\sigma) \leq \frac{1}{n},$$

이므로 $\mathrm{P}(Y \leq 2n) = 1 - \mathrm{P}(Y > 2n) \geq 1 - \frac{1}{n}$이다. 양변에 e^n을 곱하면

$$1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!} \geq \left(1 - \frac{1}{n}\right)\mathrm{e}^n,$$

이 되어 결과가 증명된다.

검산

이 부등식들은 참이지만 여유가 많다. (i)에서 $n = 2$일 때 정확한 이항분포(binomial distribution) 확률은 $\alpha \approx 0.868$로, 하한 0.5보다 훨씬 크다. $n = 10$이면 $\alpha \approx 0.9986$으로, 하한 0.9와 비교된다. (ii)에서 $n = 2$이면 좌변은 $1 + 2 + 2 + \frac{4}{3} + \frac{2}{3} = 7$이고, $\frac{1}{2}\mathrm{e}^2 \approx 3.69$이다. 두 부등식을 (i)에서는 2부터 30까지, (ii)에서는 1부터 59까지의 모든 n에 대하여 정확히 확인하였다.

흔한 실수

- k를 고를 때 표준편차 $4\sqrt{n}$ 대신 분산 $16n$을 사용하는 것.
- 체비쇼프 부등식이 구간 **바깥**에 있을 확률의 상한을 준다는 것을 잊고, α를 얻으려면 여사건을 써야 한다는 점을 놓치는 것.
- (ii)에서 아래쪽 꼬리 $Y < 0$이 아무것도 기여하지 않는 이유를 설명하지 않는 것.

STEP 너머

체비쇼프 부등식은 **집중 부등식**(concentration inequality)의 가장 간단한 예이며, 이것을 m번의 독립 시행의 평균에 적용하면 큰 수의 약한 법칙(weak law of large numbers)을 얻는다. 즉 퍼짐의 정도가 $1/\sqrt{m}$처럼 줄어든다. 소문항 (ii)는 또한 지수함수 급수(exponential series)에서 $n^{2n}/(2n)!$ 다음부터의 나머지 합이 많아야 e^n/n이라는 것을 말해 주는데, 이는 확률로 증명한 해석학적 사실이다. 체르노프 상한(Chernoff bounds) 같은 더 정밀한 도구를 쓰면 실제 확률이 지수적으로 1에 가깝다는 것을 보일 수 있다. 더 알고 싶다면 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)을 참고하라.

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.