풀이

STEP I 1991, 문제 15

확률과 통계 • 정규근사

> **STEP I 1991, 15번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/91-s1-q15

핵심 아이디어

각 X_j의 평균(mean)은 0, 분산(variance)은 1이고 각 던지기는 독립(independent)이므로, S_n의 평균은 0, 분산은 n이다. n이 클 때는 중심극한정리(central limit theorem)에 의해 S_n을 근사적으로 정규분포(normal distribution)를 따른다고 볼 수 있다. 마지막 부분은 비슷해 보이지만 함정이 있다. 모든 Y_k에 **같은** 첫 번째 던지기 X_1이 들어 있으므로 Y_k들은 독립이 아니며, 이 하나의 공통된 던지기가 결국 전체 합을 지배하게 된다.

풀이

각 X_j는 ± 1의 값을 각각 확률 $\frac{1}{2}$로 가지므로

$$\mathrm{E}(X_j) = 0, \qquad \mathrm{E}(X_j^2) = 1, \qquad \mathrm{Var}(X_j) = 1.$$

이다. 각 던지기는 독립이므로 $\mathrm{E}(S_n) = 0$이고 $\mathrm{Var}(S_n) = n$이다. 중심극한정리에 의해 n이 클 때

$$S_n \approx \mathrm{N}(0,\, n).$$

이다. (같은 말로, 앞면의 개수를 H라 하면 $H \sim \mathrm{B}(n, \frac{1}{2})$이고 $S_n = H - (n - H) = 2H - n$이다.)

소문항 (i)

$n = 10\,000$일 때 S_n의 표준편차(standard deviation)는 $\sqrt{10\,000} = 100$이므로

$$\mathrm{P}(S_n > 100) \approx \mathrm{P}\left(Z > \frac{100}{100}\right) = 1 - \Phi(1) \approx 0.16.$$

이다. 조금 더 신중하게 하고 싶다면 H에 연속성 보정(continuity correction)을 적용한다. $S_n > 100$은 $2H - 10\,000 > 100$, 즉 $H \geq 5051$을 뜻한다. H의 평균은 5000, 표준편차는 50이므로

$$\mathrm{P}(H \geq 5051) \approx \mathrm{P}\left(Z > \frac{5050.5 - 5000}{50}\right) = 1 - \Phi(1.01) \approx 0.156.$$

이다. 어느 쪽이든 답은 약 0.16이다.

소문항 (ii)

표준화(standardising)하면

$$\mathrm{P}(S_n > 0.01n) \approx \mathrm{P}\left(Z > \frac{0.01n}{\sqrt{n}}\right) = 1 - \Phi\left(0.01\sqrt{n}\right).$$

이다. 이 값은 n이 커질수록 감소하며, $\Phi(0.01\sqrt{n}) > 0.99$일 때 정확히 0.01보다 작다. 표에서 $\Phi(2.326) = 0.99$이므로

$$0.01\sqrt{n} > 2.326 \quad \Longleftrightarrow \quad \sqrt{n} > 232.6 \quad \Longleftrightarrow \quad n > 54\,100 \text{ (근사적으로)}.$$

이어야 한다. 따라서 이러한 n 중 가장 작은 값은 대략 54 000이다. (2.33을 대신 쓰면 약 54 300이 나온다. 이 정도의 근사에서는 5.4×10^4에 가까운 답이면 모두 맞는 답이다.)

새로운 점수 방식

$2 \leq k \leq n$에 대하여 첫 번째 던지기와 k번째 던지기에는 네 가지 경우가 있다.

첫 번째 던지기	k번째 던지기	점수 Y_k	$X_1 + X_k$
H	H	2	$1+1=2$
T	T	-2	$-1-1=-2$
H	T	0	$1-1=0$
T	H	0	$-1+1=0$

모든 경우에 마지막 두 열이 일치하므로, 요구한 대로 $Y_k = X_1 + X_k$이다.

평균과 분산. 선형성(linearity)에 의해 $\mathrm{E}(Y_k) = \mathrm{E}(X_1) + \mathrm{E}(X_k) = 0$이다. $k \geq 2$이므로 X_1과 X_k는 서로 다른 던지기이고, 따라서 독립이다. 그러므로

$$\mathrm{Var}(Y_k) = \mathrm{Var}(X_1) + \mathrm{Var}(X_k) = 2.$$

이다. (직접 계산하면, Y_k는 2, -2, 0을 각각 확률 $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$로 가지므로 $\mathrm{E}(Y_k^2) = 2$이다.)

극한. $m = n-1$, $T = Y_2 + Y_3 + \cdots + Y_n$이라 하자. 그러면

$$T = \sum_{k=2}^{n} (X_1 + X_k) = mX_1 + S', \qquad \text{여기서} S' = X_2 + X_3 + \cdots + X_n.$$

이다.

> **생각해 보기** 각 Y_k의 평균은 0, 분산은 2로, 마치 상수배한 X_j와 같다. 그런데 왜 소문항 (ii)의 논증을 그대로 가져다 쓸 수 없을까?

소문항 (ii)의 논증에는 각 항이 독립이라는 조건이 필요했는데, Y_k들은 모두 X_1을 공유하므로 독립이 아니기 때문이다. 실제로 $j \neq k$이면 $\mathrm{Cov}(Y_j, Y_k) = \mathrm{Var}(X_1) = 1$이고, $\mathrm{Var}(T) = m^2 + m$이며 $2m$이 아니다. 따라서 X_1의 두 값을 따로 살펴본다. S'은 각각 ± 1인 m개 항의 합이므로 $-m \le S' \le m$임에 주목하자.

- 첫 번째 던지기가 뒷면이면 $X_1 = -1$이고 $T = -m + S' \le 0$이다. 따라서 $T > 0.01m$은 불가능하다.
- 첫 번째 던지기가 앞면이면 $X_1 = 1$이고 $T = m + S'$이므로, $T > 0.01m$은 $S' > -0.99m$일 때 정확히 성립한다. 소문항 (ii)와 같은 정규 근사를 $S' \approx \mathrm{N}(0, m)$으로 적용하면

  $$\mathrm{P}(S' \le -0.99m) \approx \Phi\left(-0.99\sqrt{m}\right) \to 0 \quad \text{단, } m \to \infty.$$

  이다. (근사가 없어도 된다. 체비쇼프 부등식(Chebyshev's inequality)에 의해 이 확률은 많아야 $m/(0.99m)^2 \to 0$이다.)

S'은 X_1과 독립이므로

$$\mathrm{P}(T > 0.01m) = \tfrac{1}{2} \cdot 0 + \tfrac{1}{2}\,\mathrm{P}(S' > -0.99m) \;\to\; \tfrac{1}{2} \quad \text{단, } n \to \infty.$$

이다. 따라서 주어진 명제는 **거짓**이다. 확률은 0이 아니라 $\frac{1}{2}$로 수렴한다. 말로 하면, n이 클 때 합은 사실상 첫 번째 던지기가 앞면일 때 양수이고, 그렇지 않으면 양수가 아니다.

검산

- 소문항 (i): $H \sim \mathrm{B}(10\,000, \frac{1}{2})$에 대한 정확한 이항(binomial) 확률 $\mathrm{P}(H \ge 5051)$은 0.15625로, 연속성 보정을 한 값 0.156과 일치한다.
- 소문항 (ii): $\mathrm{P}(S_n > 0.01n)$의 정확한 이항 확률값은 $n = 50\,000$에서 0.0125, $n = 54\,000$ 근처에서 약 0.0100, $n = 55\,000$에서 0.0094이므로, 경계값은 실제로 54 000에 가깝다.
- 마지막 부분: $n \le 14$에 대해 2^n가지 결과를 모두 나열하면 $\frac{1}{2}\,\mathrm{P}(S' > -0.99m)$와 정확히 일치한다. $n = 14$에서 확률은 이미 0.49994이고, $n = 100$인 시뮬레이션에서는 0.497이 나왔다.

흔한 실수

- $Y_2, \ldots, Y_n$을 독립으로 보고 $\mathrm{Var}(T) = 2(n-1)$이라고 결론 내리는 것. 이렇게 하면 확률이 0으로 수렴하는 것처럼 보인다.
- 점수의 분산으로 n 대신 $\mathrm{Var}(S_n) = n/4$ (앞면 개수의 분산)를 쓰는 것.

STEP 너머

소문항 (ii)는 대수의 약법칙(weak law of large numbers)을 정량적으로 나타낸 것이다. 즉 n이 크면 S_n/n이 0에서 멀리 떨어질 가능성은 매우 작다. 마지막 부분은 각 항이 강하게 종속되어 있을 때 이 법칙이 어떻게 깨지는지를 보여 준다. 여기서는 모든 항이 첫 번째 던지기를 공유하므로, 평균 $T/m = X_1 + S'/m$은 평균값 0이 아니라 확률변수인 X_1로 수렴한다. 종속된 수열에 대한 대수의 법칙(에르고딕 정리(ergodic theorems), 교환 가능한 수열(exchangeable sequences))은 학부 확률론에서 풍부한 주제이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 대수의 약법칙과 체비쇼프 부등식을 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.