풀이

STEP II 2020, 문제 12

확률과 통계 • 편향된 주사위와 부등식

> **STEP II 2020, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2020년 STEP 시험지 묶음(zip 파일) 가운데 STEP II 시험지에 있다:
>
> www.ocr.org.uk/Images/696589-2020-past-papers.zip

핵심 아이디어

확률의 합은 1이어야 하므로 치우침(bias)들은 $\sum \varepsilon_i = 0$을 만족한다. 이 사실을 적어 두면, $\sum(\frac{1}{n}+\varepsilon_i)^2$을 전개할 때 일차항이 모두 사라지고 눈에 띄게 음이 아닌 식만 남는다. 그러면 (ii)는 모습만 바꾼 (i)이다. 양수들을 그 합으로 나누면 어떤 주사위의 확률들이 되기 때문이다.

풀이

$p_i = \mathrm{P}(X=i) = \frac{1}{n}+\varepsilon_i$라 하자. $\sum_{i=1}^{n} p_i = 1$이므로

$$1 = n\cdot\frac{1}{n} + \sum_{i=1}^{n}\varepsilon_i \quad\Longrightarrow\quad \sum_{i=1}^{n}\varepsilon_i = 0.$$

또한 각 $p_i \geq 0$이므로 모든 i에 대하여 $\varepsilon_i \geq -\frac{1}{n}$이다. 아래의 합은 모두 $i=1,\ldots,n$에 대한 합이다.

소문항 (i)

두 번의 굴림은 독립(independent)이므로 두 번 모두 i가 나올 확률은 p_i^2이고, 서로 다른 i에 대한 이 사건들은 서로 배반(disjoint)이다. 따라서

$$\begin{aligned}\mathrm{P}(\text{같은 눈이 두 번 나옴}) &= \sum p_i^2 = \sum\left(\frac{1}{n^2}+\frac{2\varepsilon_i}{n}+\varepsilon_i^2\right)\\ &= \frac{1}{n}+\frac{2}{n}\sum\varepsilon_i+\sum\varepsilon_i^2 = \frac{1}{n}+\sum_{i=1}^{n}\varepsilon_i^2.\end{aligned}$$

공정한 주사위에서는 모든 $\varepsilon_i = 0$이고 확률은 $\frac{1}{n}$이다. 치우친 주사위에서는 ε_i가 모두 0인 것은 아니므로 $\sum \varepsilon_i^2 > 0$이다. 따라서 치우친 주사위는 공정한 주사위보다 연속한 두 번의 굴림에서 같은 눈이 나올 가능성이 더 크다.

소문항 (ii)

먼저 이중합을 다시 쓰자. 이 합은 $j < i$인 모든 쌍에 대하여 $x_i x_j$를 더한 것, 즉 서로 다른 첨자의 순서 없는 쌍 각각에 대하여 정확히 한 번씩 더한 것이다. $S = \sum x_i$를 제곱하면 각 x_i^2이 한 번씩, 그리고 그러한 쌍이 각각 두 번씩 나오므로

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j = \frac{1}{2}\left(S^2 - \sum x_i^2\right).$$

생각해 보기 x_i가 그저 임의의 양수일 때, 주사위는 어디에서 나오는가?

$p_i = x_i / S$라 하자. 이들은 양수이고 합이 1이므로 어떤 n면 주사위의 확률들이다. (i)에 의하여 $\sum p_i^2 = \frac{1}{n} + \sum \varepsilon_i^2 \geq \frac{1}{n}$이다(등호는 주사위가 공정할 때, 즉 모든 x_i가 같을 때에만 성립한다). 양변에 $S^2 > 0$을 곱하면

$$\sum x_i^2 \geq \frac{S^2}{n}.$$

따라서

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j = \frac{1}{2}\left(S^2 - \sum x_i^2\right) \leq \frac{1}{2}\left(S^2 - \frac{S^2}{n}\right) = \frac{n-1}{2n}\left(\sum_{i=1}^{n} x_i\right)^2,$$

이것이 구하는 부등식(inequality)이다. 등호는 모든 x_i가 같을 때 성립한다.

소문항 (iii)

이번에는 확률이 $\sum p_i^3$이다. 전개하고 다시 $\sum \varepsilon_i = 0$을 이용하면

$$\begin{aligned}\sum p_i^3 &= \sum \left(\frac{1}{n^3} + \frac{3\varepsilon_i}{n^2} + \frac{3\varepsilon_i^2}{n} + \varepsilon_i^3\right) \\ &= \frac{1}{n^2} + \sum_{i=1}^{n}\left(\frac{3\varepsilon_i^2}{n} + \varepsilon_i^3\right) = \frac{1}{n^2} + \sum_{i=1}^{n} \varepsilon_i^2 \left(\frac{3}{n} + \varepsilon_i\right).\end{aligned}$$

공정한 주사위에서 이 값은 $\frac{1}{n^2}$이다.

생각해 보기 세제곱 ε_i^3은 음수일 수 있다. 이것이 제곱항들보다 더 클 수 있을까?

그럴 수 없다. 치우침은 아주 큰 음수가 될 수 없기 때문이다. $p_i \geq 0$에서 $\varepsilon_i \geq -\frac{1}{n}$이므로

$$\frac{3}{n} + \varepsilon_i \geq \frac{2}{n} > 0.$$

따라서 각 항 $\varepsilon_i^2\left(\frac{3}{n}+\varepsilon_i\right)$은 ≥ 0이고, $\varepsilon_i \neq 0$일 때마다 양수인데, 치우친 주사위에서는 적어도 하나의 i에 대하여 이런 일이 일어난다. 그러므로

$$\sum p_i^3 > \frac{1}{n^2},$$

이고, 치우친 주사위는 이번에도 공정한 주사위보다 연속한 세 번의 굴림에서 같은 눈이 나올 가능성이 **더 크다**.

검산

- $n=4$이고 확률이 $\frac{1}{2},\frac{1}{6},\frac{1}{6},\frac{1}{6}$인 경우를 보자. 그러면 $\varepsilon=\frac{1}{4},-\frac{1}{12},-\frac{1}{12},-\frac{1}{12}$이다. 직접 계산하면 $\sum p_i^2=\frac{1}{4}+\frac{3}{36}=\frac{1}{3}$이고, $\frac{1}{4}+\sum\varepsilon_i^2=\frac{1}{4}+\frac{1}{16}+\frac{3}{144}=\frac{1}{3}$이며, 이는 공정한 경우의 $\frac{1}{4}$보다 크다. 세 번 굴리는 경우 $\sum p_i^3=\frac{1}{8}+\frac{3}{216}=\frac{5}{36}\approx 0.139$로, 공정한 경우의 $\frac{1}{16}=0.0625$보다 훨씬 크다.
- 가장 극단적인 주사위는 항상 1이 나오는 주사위이다. 이때 두 확률은 모두 1로, 가능한 가장 큰 값이다.
- (ii)에서 $n=2$이면 부등식은 $x_1x_2\leq\frac{1}{4}(x_1+x_2)^2$이 되는데, 이것은 산술-기하 평균 부등식(AM–GM)이다. (i), (ii), (iii)의 부등식은 수천 개의 무작위 사례에 대하여 컴퓨터로도 확인하였다.

흔한 실수

- $\sum\varepsilon_i=0$임을 잊어서, 부호를 알 수 없는 일차항이 남는 경우.
- (iii)에서 $\varepsilon_i\geq-\frac{1}{n}$을 이용하지 않고 $\sum\varepsilon_i^3$이 "작다"거나 부호가 어느 쪽이든 상관없다고 단정하는 경우. 이 논증을 성립시키는 것이 바로 그 하한이다.
- (ii)에서 part (i)을 $x_i/\sum x_j$가 아니라 x_i 자체에 적용하는 경우. x_i들의 합은 1이 아닐 수 있다.

STEP 너머

$\sum p_i^2$라는 양은 분포의 **충돌 확률**(collision probability)이며, (i)은 이 값이 균등분포(uniform distribution)에서 가장 작다는 것을 말한다. (iii)은 $\sum p_i^3$에 대한 같은 명제이다. 둘 다 $t\mapsto t^k$의 볼록성(convexity), 즉 옌센 부등식(Jensen's inequality)이나 멱평균 부등식(power-mean inequality)의 특수한 경우이며, $\sum p_i^k$의 로그를 취하면 정보 이론과 암호학에서 쓰이는 레니 엔트로피(Rényi entropy)를 얻는다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문서)은 이러한 아이디어를 확률론의 맥락에서 접하기에 좋은

책이다.

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.