풀이

STEP III 2000, 문제 12

확률과 통계 • 푸아송 근사

> **STEP III 2000, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/00-s3-q12

핵심 아이디어

자신의 복권이 당첨된다는 것을 알고 나면, 남은 무작위성은 다른 참가자들의 복권뿐이다. 다른 참가자 각각은 독립적으로(independently) 확률 $\frac{1}{N}$로 당첨 번호와 일치하므로, 다른 당첨자의 수는 이항분포(binomial distribution)를 따른다. 참가자가 많고 성공 확률이 작으면 이 이항분포는 푸아송 분포(Poisson distribution)에 가까워지며, 주어진 극한 $\left(1-\frac{a}{n}\right)^n \approx \mathrm{e}^{-a}$가 바로 이것을 가능하게 한다.

풀이

정확한 확률

당첨 복권을 가지고 있다고 하자. 나머지 $2N-1$명의 참가자 가운데 당첨 번호를 가진 사람의 수를 X라 하면, 전체 당첨자 수는 $W = 1 + X$이다.

기계는 각 참가자의 번호를 다른 모든 사람의 번호와 독립적으로, 그리고 추첨과도 독립적으로 균등하게(uniformly) 고른다. 따라서 당첨 번호가 (무엇이든) 주어지고 자신의 복권이 주어졌을 때, 다른 $2N-1$명의 참가자 각각은 서로 독립적으로 확률 $\frac{1}{N}$로 당첨 번호와 일치한다. 그러므로

$$X \sim \mathrm{B}\left(2N-1,\ \tfrac{1}{N}\right).$$

당첨자가 세 명 이하라는 것은 $X \le 2$라는 뜻이므로, 정확한 확률은

$$\mathrm{P}(W \le 3) = \sum_{k=0}^{2} \binom{2N-1}{k} \left(\frac{1}{N}\right)^k \left(1-\frac{1}{N}\right)^{2N-1-k}.$$

이를 모두 풀어 쓰면 다음과 같다.

$$\left(1-\frac{1}{N}\right)^{2N-1}+\frac{2N-1}{N}\left(1-\frac{1}{N}\right)^{2N-2}+\frac{(2N-1)(2N-2)}{2N^2}\left(1-\frac{1}{N}\right)^{2N-3}.$$

근사

고정된 각 $k\in\{0,1,2\}$에 대하여 거듭제곱을 다음과 같이 나눈다.

$$\left(1-\frac{1}{N}\right)^{2N-1-k}=\left[\left(1-\frac{1}{N}\right)^{N}\right]^{2}\left(1-\frac{1}{N}\right)^{-(1+k)}.$$

주어진 사실에 $a=1$, $n=N$을 대입하면 대괄호 안은 약 e^{-1}이다. 마지막 인수는 1에 매우 가까운 수의 고정된 작은 거듭제곱이므로 약 1이다. 따라서 각 거듭제곱은 약 e^{-2}이다.

이항계수에 대해서는, k가 많아야 2이고 N이 크므로

$$\binom{2N-1}{k}\frac{1}{N^k}=\frac{(2N-1)(2N-2)\cdots(2N-k)}{k!\,N^k}\approx\frac{(2N)^k}{k!\,N^k}=\frac{2^k}{k!}.$$

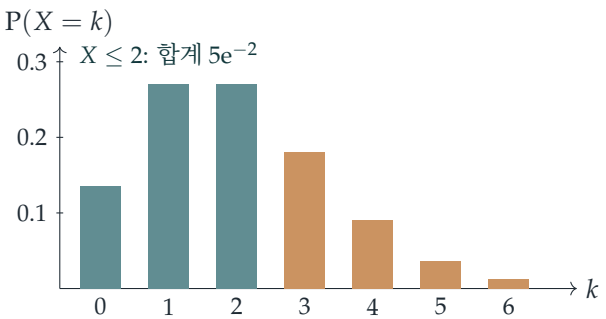
이것들을 종합하면 $k=0,1,2$에 대한 항은 각각 약 e^{-2}, $2\mathrm{e}^{-2}$, $2\mathrm{e}^{-2}$이므로, $\mathrm{e}^2\approx 7.39$를 이용하면

$$\mathrm{P}(W\le 3)\approx \mathrm{e}^{-2}\left(1+2+\frac{2^2}{2!}\right)=5\mathrm{e}^{-2}\approx 0.677\approx\frac{2}{3},$$

이다. 이것이 구하는 결과이다.

? 생각해 보기 X는 왜 평균이 2인 푸아송 확률변수처럼 행동하는가?

위의 계산은 이항분포의 표준적인 푸아송 극한이다. 시행 횟수 $2N-1$은 크고 성공 확률 $\frac{1}{N}$은 작으며, 둘의 곱은 2에 가깝게 유지된다. 이러한 상황에서 $\lambda=np$라 하면 $\binom{n}{k}p^k(1-p)^{n-k}\to \mathrm{e}^{-\lambda}\lambda^k/k!$이고, 여기서는 $\lambda\to 2$이다.



X의 극한분포는 평균이 2인 푸아송 분포이다. 청록색 막대($X\le 2$, 즉 전체 당첨자가 많아야 세 명)는 전체의 약 3분의 2를 차지한다.

참가자가 직접 번호를 고르는 경우

참가자가 직접 고르면 번호는 더 이상 균등하지 않다. 사람들은 생일(31 이하의 번호), 행운의 숫자, 규칙적인 패턴을 선호한다. 참가자가 번호 i를 확률 p_i로 고른다고 하자. 여기서 $\sum p_i = 1$이고, 당첨 번호는 여전히 균등하게 추첨된다. 그러면 자신이 당첨되었다는 조건 아래에서 자신의 번호가 i일 확률은 $p_i \cdot \frac{1}{N}$에 비례하며, 곧 p_i이다. 따라서 인기 있는 번호는 당첨 복권 가운데 실제보다 많이 나타난다. 자신의 번호가 i이면 다른 당첨자의 수는 $\mathrm{B}(2N-1, p_i)$를 따르고 평균은 $(2N-1)p_i$이다. i에 대해 평균을 내면 다른 당첨자 수의 기댓값(expectation)은 코시–슈바르츠 부등식(Cauchy–Schwarz inequality) $(\sum p_i)^2 \le N \sum p_i^2$에 의해

$$(2N-1)\sum_i p_i^2 \;\ge\; (2N-1)\cdot\frac{1}{N},$$

이고, 등호는 모든 $p_i = \frac{1}{N}$일 때에만 성립한다. 따라서 번호 선택이 균등하지 않으면 평균적으로 더 많은 사람이 당첨을 나누어 가지게 되고, 당첨자가 세 명 이하일 확률은 대체로 **감소**한다. (인기 없는 번호를 일부러 고르는 한 참가자는 평균보다 나은 결과를 얻겠지만, 참가자 전체로는 더 나빠진다.)

당첨자 수의 기댓값

$X \sim \mathrm{B}\left(2N-1, \frac{1}{N}\right)$이므로 그 평균은 $np = (2N-1)\cdot\frac{1}{N}$이다. (같은 이야기로, X를 다른 참가자 한 명마다 하나씩, 각각 평균이 $\frac{1}{N}$인 $2N-1$개의 지시확률변수(indicator variable)의 합으로 쓸 수 있다.) 그러므로

$$\mathrm{E}(W) = 1 + \mathrm{E}(X) = 1 + \frac{2N-1}{N} = 3 - \frac{1}{N} = 3 - N^{-1},$$

이고, 이것이 구하는 결과이다. 이것은 근사가 아니라 정확한 값임에 유의한다.

검산

- $N = 2$와 $N = 3$에 대하여 $2N$명의 참가자에게 번호를 배정하는 모든 경우와 모든 당첨 번호를 나열하면, 정확히 이항분포 식과 정확히 $3 - \frac{1}{N}$을 얻는다.
- 정확한 확률은 $N = 10$일 때 0.7054, $N = 100$일 때 0.6794, $N = 10\,000$일 때 0.6767로, $5\mathrm{e}^{-2} = 0.67668$에 가까워진다.
- $N = 50$으로 시뮬레이션(10^5주)을 하면 정확한 값 0.682에 대해 0.683을 얻었고, 당첨자 수의 평균은 $3 - \frac{1}{50} = 2.98$에 대해 2.981이었다.
- $N = 10$이고 번호의 절반이 나머지보다 두 배 인기가 있을 때, 확률은 0.705에서 0.641로 떨어지고 당첨자 수의 기댓값은 2.9에서 3.11로 올라간다.

흔한 실수

- 시행 횟수를 $2N-1$이 아닌 $2N$으로 쓰는 것: 자신의 복권은 이미 당첨된다는 것을 알고 있으므로 무작위인 것은 다른 참가자들뿐이다. 시행 횟수를 $2N$으로 하면 기댓값이 $3-\frac{1}{N}$이 아니라 $1+2=3$으로 나온다.
- 조건 없는 평균을 쓰는 것: 모든 주를 통틀어 당첨자 수의 기댓값은 $2N\cdot\frac{1}{N}=2$이지만, 자신이 당첨된다는 것을 알면 확실한 당첨자 한 명이 더해지고 나머지는 $2N-1$명이 된다.
- $k=2$인 항을 빠뜨리거나, $\frac{1}{2!}$ 없이 2^2e^{-2}로 쓰는 것.

STEP 너머

여기서의 근사는 희귀 사건의 법칙(law of rare events)이다. 즉 시행 횟수가 많고 성공 확률이 작은 이항분포는 푸아송 분포로 수렴한다. 마지막 소문항은 검사 역설(inspection paradox), 또는 크기 편향 표본추출(size-biased sampling)과 맞닿아 있다. 자신의 복권이 당첨된다는 조건을 걸면 많은 사람이 고른 번호일 가능성이 커지며, 그래서 조건부(conditional) 평균이 2가 아니라 3이 된다. 두 아이디어 모두 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 그리고 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에 잘 설명되어 있다.

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.