풀이

STEP III 1987, 문제 16

확률과 통계 • 순서통계량과 푸아송분포

STEP III 1987, 16번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/87-s3-q16

핵심 아이디어

소문항 (i)에서 사건 "$A \le k$이고 $B \ge 1-k$"를 직접 다루기는 까다롭지만, 그 여사건(complement)은 두 단순한 사건 "모든 X_i가 k보다 크다"와 "모든 X_i가 $1-k$보다 작다"의 합사건(union)이다. 그다음은 포함-배제 원리(inclusion–exclusion)가 해결해 준다. 소문항 (ii)에서 당첨자 수는 n이 매우 크고 p가 매우 작은 이항분포(binomial distribution)를 따르는데, 이것이 바로 푸아송 분포(Poisson distribution)가 등장하는 상황이다. 따라서 당첨자가 한 명 또는 두 명일 푸아송 확률을 최대화한 다음, 근사(approximation) 때문에 답이 60%를 넘어갈 수는 없다는 것을 확인한다.

풀이

소문항 (i)

E를 사건 $A > k$, F를 사건 $B < 1-k$라 하자. 그러면 $A > k$는 모든 X_i가 k보다 크다는 뜻이고, $B < 1-k$는 모든 X_i가 $1-k$보다 작다는 뜻이다. 독립성(independence)에 의해

$$\mathrm{P}(E) = (1-k)^n, \qquad \mathrm{P}(F) = (1-k)^n, \qquad \mathrm{P}(E \cap F) = (1-2k)^n,$$

이다. 마지막 식이 성립하는 것은 $E \cap F$가 모든 X_i가 길이 $1-2k > 0$인 구간 $(k, 1-k)$에 놓인다는 뜻이기 때문이다.

구하려는 사건은 $E \cup F$의 여사건이므로, 포함-배제 원리에 의해

$$p_n = 1 - \mathrm{P}(E \cup F) = 1 - 2(1-k)^n + (1-2k)^n.$$

$0 < 1-2k < 1-k < 1$이므로 두 거듭제곱은 모두 0으로 수렴하고, 따라서

$$p_n \to 1 \quad \text{단, } n \to \infty.$$

참고. k를 아무리 작게 잡더라도, 관측값이 충분히 많으면 어떤 관측값은 0으로부터 k 이내에, 또 다른 관측값은 1로부터 k 이내에 놓인다는 것이 거의 확실해진다. 다시 말해 표본 최솟값(sample minimum)은 0으로, 표본 최댓값(sample maximum)은 1로 (확률적으로) 수렴하므로, 표본 범위(sample range) $B - A$는 전체 구간의 길이인 1로 수렴한다. 예를 들어 A와 B가 미지의 균등분포(uniform distribution)의 양 끝점에 대한 자연스러운 추정량(estimator)인 것도 이 때문이다.

소문항 (ii)

$N = 4\,000\,000$이라 하고 X를 당첨자 수라 하자. 독자들은 각각 확률 p로 서로 독립적으로 당첨되므로 $X \sim \mathrm{B}(N, p)$이다. $q = 1 - p$라 쓰자. 자리를 지키는 것은 정확히 $X = 1$ 또는 $X = 2$일 때이며, 그 확률은

$$K(p) = Npq^{N-1} + \binom{N}{2} p^2 q^{N-2}.$$

푸아송 분포로 바라보기. N이 매우 크고 p가 작으므로 X는 근사적으로 평균이 $\lambda = Np$인 푸아송 분포를 따르고, 따라서

$$K(p) \approx h(\lambda), \qquad \text{여기서} h(\lambda) = \mathrm{e}^{-\lambda}\left(\lambda + \frac{\lambda^2}{2}\right).$$

미분하면

$$h'(\lambda) = \mathrm{e}^{-\lambda}\left(1 + \lambda - \lambda - \frac{\lambda^2}{2}\right) = \mathrm{e}^{-\lambda}\left(1 - \frac{\lambda^2}{2}\right),$$

이고, 이는 $0 < \lambda < \sqrt{2}$에서 양수, $\lambda > \sqrt{2}$에서 음수이다. 따라서 h는 $\lambda = \sqrt{2}$에서 최댓값을 가진다.

$$h(\sqrt{2}) = \mathrm{e}^{-\sqrt{2}}(\sqrt{2} + 1) \approx 0.2431 \times 2.4142 \approx 0.5869.$$

가장 좋은 선택은 $p = \sqrt{2}/N \approx 3.5 \times 10^{-7}$이고, 그렇게 하더라도 살아남을 확률은 약 58.7%에 불과하다. 이것이 문제가 요구하는 추론이다.

> **생각해 보기** 푸아송 분포는 근사일 뿐이다. 실제 이항분포 확률이 0.6을 슬그머니 넘어설 수는 없을까?

그럴 수 없다. $K(p)$를 직접 위로 유계시킬 수 있다. $q \le 1$이고 $\binom{N}{2} \le \frac{1}{2}N^2$이므로

$$K(p) = q^{N-2}\left(Npq + \binom{N}{2}p^2\right) \le q^{N-2}\left(\lambda + \frac{\lambda^2}{2}\right).$$

이제 모든 p에 대해 $1 - p \le \mathrm{e}^{-p}$이므로(직선 $1 - p$는 볼록(convex)한 곡선 e^{-p}의 $p = 0$에서의 접선(tangent)이다), $\mu = (N-2)p$라 할 때 $q^{N-2} \le \mathrm{e}^{-(N-2)p} = \mathrm{e}^{-\mu}$이다. $c = N/(N-2)$라 두면 $c > 1$이고 $\lambda = c\mu$이다. 그러면

$$K(p) \le \mathrm{e}^{-\mu}\left(c\mu + \frac{c^2\mu^2}{2}\right) \le c^2\,\mathrm{e}^{-\mu}\left(\mu + \frac{\mu^2}{2}\right) = c^2 h(\mu) \le c^2 h(\sqrt{2}).$$

여기서 $c^2 = (1 - \frac{2}{N})^{-2} < 1.000002$이므로, p를 어떻게 선택하든

$$K(p) < 1.000002 \times 0.58694 < 0.587 < 0.6.$$

따라서 p를 어떻게 정하더라도 자리를 지킬 확률은 60%보다 작으며, 이로써 요구된 결과가 증명된다.

검산

- 소문항 (i)에서 $n = 1$이면 $p_1 = 1 - 2(1-k) + (1-2k) = 0$이다. 이는 옳다. $k < \frac{1}{2}$일 때 관측값 하나가 $\le k$이면서 동시에 $\ge 1-k$일 수는 없기 때문이다. $n = 2$이면 $p_2 = 2k^2$를 얻는데, 이는 두 점 중 하나씩이 양 끝의 띠에 각각 떨어질 확률(순서는 상관없음)이다.
- $n = 10, k = 0.3$으로 한 시뮬레이션(simulation)에서는 0.944를 얻었는데, 공식의 값은 0.9436이다.
- $\lambda = Np$의 촘촘한 격자에서 정확한 이항분포 확률 $K(p)$를 계산하면 $\lambda \approx 1.414$에서 최댓값 0.58694를 얻으며, 이는 푸아송 값과 소수점 아래 다섯째 자리까지 일치한다.

흔한 실수

- $p_n = \mathrm{P}(A \le k)\,\mathrm{P}(B \ge 1-k)$라고 쓰는 것. 같은 표본의 최솟값과 최댓값은 독립이 아니다.
- 이항분포 확률을 p에 대해 직접 최대화하는 것. 할 수는 있지만 계산이 번거롭다. $\lambda = Np$로 다루면 한 줄짜리 미분 문제가 된다.
- 0.587이 0.6에 상당히 가까워서 어느 정도 정당화가 필요하다는 점을 알아차리지 못한 채 푸아송 근사에서 멈추는 것.

STEP 너머

소문항 (i)은 **순서통계량**(order statistics)과 확률수렴(convergence in probability)을 처음 맛보는 문제이다. 표본의 최솟값과 최댓값은 받침(support)의 양 끝으로 수렴한다. 소문항 (ii)는 푸아송의 "희귀 사건의 법칙(law of rare events)"이며, 여기서 사용한 부등식은 p가 작을 때 이항분포와 푸아송 분포가 총변동 거리(total variation)의 의미에서 매우 가깝다는 일반적인 사실의 간단한 예이다. 두 아이디어 모두 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 그리고 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에서 더 자세히 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.