풀이

STEP II 2005, 문제 13

확률과 통계 • 푸아송 분포와 이항분포

> **STEP II 2005, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/05-s2-q13

핵심 아이디어

표본으로 뽑은 각 쪽은 오류가 두 개 미만이거나 그렇지 않거나 둘 중 하나이고, 각 쪽은 서로 독립(independent)적으로 행동하므로 Y는 n번의 독립시행에서 성공한 횟수를 센다. 즉 Y는 이항분포(binomial distribution)를 따른다. 그 이후의 모든 내용은 λ가 작을 때 어떤 일이 일어나는지에 관한 것이며, 이를 위한 도구는 $\mathrm{e}^{-\lambda}$(또는 $\ln(1+\lambda)$)의 매클로린 급수(Maclaurin series)이다. 이때 상쇄되지 않는 첫 번째 항만 남긴다.

풀이

Y의 분포

특정한 한 쪽에 있는 오류의 개수를 X라 하면 $X \sim \mathrm{Po}(\lambda)$이다. $X \leq 1$인 쪽을 **양호**하다고 하자. 그러면

$$\mathrm{P}(\text{양호}) = \mathrm{P}(X=0) + \mathrm{P}(X=1) = \mathrm{e}^{-\lambda} + \lambda\mathrm{e}^{-\lambda} = (1+\lambda)\mathrm{e}^{-\lambda} = p.$$

표본은 서로 다른 n개의 쪽으로 이루어진다. 서로 다른 쪽의 오류 개수는 독립이고 모두 같은 분포 $\mathrm{Po}(\lambda)$를 따르므로, 어떤 n개의 쪽을 고르든 사건 "j번째 쪽이 양호하다"들은 서로 독립이고 각각의 확률은 p이다. (조건 $n \ll N$은 표본을 뽑을 쪽이 충분히 많다는 것을 보장해 줄 뿐이다. 독립성은 모형에서 나오는 것이므로 이 조건이 필요하지는 않다.)

따라서 Y는 성공 확률이 p인 n번의 독립시행에서 성공한 횟수이다. 양호한 쪽 k개와 불량한 쪽 $n-k$개로 이루어진 특정한 조합의 확률은 $p^k q^{n-k}$이고, 어느 k개의 쪽이 양호한지 고르는 방법은 $\binom{n}{k}$가지이므로

$$\mathrm{P}(Y=k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, 1, \ldots, n,$$

이다. 여기서 $p = (1+\lambda)\mathrm{e}^{-\lambda}$, $q = 1-p$이며, 이것이 보이려던 것이다. 즉 $Y \sim \mathrm{B}(n, p)$이다.

소문항 (i)

$\mathrm{e}^{-\lambda} = 1 - \lambda + \frac{1}{2}\lambda^2 - \frac{1}{6}\lambda^3 + \cdots$로 전개하고 $1+\lambda$를 곱하면

$$\begin{aligned} p = (1+\lambda)\mathrm{e}^{-\lambda} &= (1 - \lambda + \tfrac{1}{2}\lambda^2 - \tfrac{1}{6}\lambda^3 + \cdots) + (\lambda - \lambda^2 + \tfrac{1}{2}\lambda^3 - \cdots) \\ &= 1 - \tfrac{1}{2}\lambda^2 + \tfrac{1}{3}\lambda^3 - \cdots. \end{aligned}$$

λ의 일차항이 상쇄되므로, λ가 작을 때

$$q = 1 - p = \tfrac{1}{2}\lambda^2 - \tfrac{1}{3}\lambda^3 + \cdots \approx \tfrac{1}{2}\lambda^2$$

이다. 무시한 항들은 $\frac{1}{2}\lambda^2$보다 λ 정도의 비율만큼 작기 때문이다.

생각해 보기 $\mathrm{e}^{-\lambda}$의 첫 항만 보면 그럴 것 같은데, 왜 $q \approx \lambda$가 아닐까?

$\mathrm{e}^{-\lambda}$에서 나온 $-\lambda$가 $1+\lambda$를 곱해서 생긴 $+\lambda$와 정확히 상쇄되기 때문이다. 한 쪽이 불량하려면 오류가 적어도 두 개 있어야 하고, $\mathrm{P}(X \geq 2) \approx \mathrm{P}(X = 2) = \frac{1}{2}\lambda^2\mathrm{e}^{-\lambda} \approx \frac{1}{2}\lambda^2$이다. 이는 (i)을 독립적으로 확인해 주는 좋은 근거이다.

소문항 (ii)

$\mathrm{P}(Y = n) = p^n$이고, $p^n \geq 1 - \lambda$를 만족하는 가장 큰 n을 구해야 한다. ($0 < \lambda < 1$일 때) 양변이 모두 양수이고 ln은 증가함수이므로, 이는

$$n \ln p \geq \ln(1-\lambda).$$

와 동치이다. 이제 $\ln p = \ln(1+\lambda) - \lambda = (\lambda - \frac{1}{2}\lambda^2 + \cdots) - \lambda \approx -\frac{1}{2}\lambda^2$이고 $\ln(1-\lambda) \approx -\lambda$이다. $\ln p < 0$이므로 이것으로 나누면 부등호의 방향이 바뀐다.

$$n \leq \frac{\ln(1-\lambda)}{\ln p} \approx \frac{-\lambda}{-\frac{1}{2}\lambda^2} = \frac{2}{\lambda}.$$

따라서 그러한 가장 큰 n은 근사적으로 $2/\lambda$이다.

(i)을 이용해도 같은 결과를 얻는다. $p^n = (1-q)^n \approx 1 - nq \approx 1 - \frac{1}{2}n\lambda^2$이고, $1 - \frac{1}{2}n\lambda^2 \geq 1 - \lambda$는 정확히 $n \leq 2/\lambda$일 때 성립한다. $(1-q)^n \approx 1 - nq$라는 단계는 nq가 작아야 하는데, 여기서 $nq \approx \frac{2}{\lambda} \cdot \frac{1}{2}\lambda^2 = \lambda$로 작으므로 이 근사는 자기모순 없이 일관된다.

소문항 (iii)

$Y > 1$이면 $Y > 0$이므로

$$\mathrm{P}(Y > 1 \mid Y > 0) = \frac{\mathrm{P}(Y > 1)}{\mathrm{P}(Y > 0)} = \frac{\mathrm{P}(Y > 0) - \mathrm{P}(Y = 1)}{\mathrm{P}(Y > 0)} = 1 - \frac{npq^{n-1}}{1 - q^n}.$$

(여기서 $n \geq 2$라는 조건이 중요하다. 이 조건 덕분에 $Y > 1$이 일어날 수 있는 사건이 되고, 거듭제곱 q^{n-1}이 적어도 q의 거듭제곱이 된다.) 이제 n을 고정하고 λ가 작다고 하자. 그러면 $q \approx \frac{1}{2}\lambda^2$이 작으므로 $p \approx 1$이고 $1 - q^n \approx 1$이어서

$$\mathrm{P}(Y > 1 \mid Y > 0) \approx 1 - nq^{n-1} \approx 1 - n\left(\frac{\lambda^2}{2}\right)^{n-1},$$

을 얻는다. 마지막 단계에서 (i)을 이용하였다. 이것이 구하려던 결과이다.

검산

- $\lambda = 0.8$인 푸아송 분포(Poisson distribution)를 따르는 $n = 5$개의 쪽으로 이루어진 표본을 200,000번 시뮬레이션한 결과, $Y = 0, \ldots, 5$의 상대도수가 $\mathrm{B}(5, p)$와 0.002 이내로 일치하였다.
- $\lambda = 0.01$일 때 $q = 4.967 \times 10^{-5}$이고, $\frac{1}{2}\lambda^2 = 5 \times 10^{-5}$이다.
- $\lambda = 0.01$일 때 $p^n \geq 0.99$를 만족하는 가장 큰 n의 정확한 값은 202이고, $2/\lambda = 200$이다. (각 급수에서 항을 하나씩 더 남기면 $2/\lambda + \frac{7}{3}$을 얻으므로, $2/\lambda$가 커지는 동안 오차는 유계로 남는다.)
- $\lambda = 0.01$, $n = 3$일 때 정확한 값은 $1 - \mathrm{P}(Y > 1 \mid Y > 0) = 7.40 \times 10^{-9}$이고, $3(\lambda^2/2)^2 = 7.5 \times 10^{-9}$이다.

흔한 실수

- "오류가 두 개 미만"을 $X = 0$만으로, 또는 $X \leq 2$로 받아들이는 것.
- $(1 + \lambda)\mathrm{e}^{-\lambda}$의 전개를 일차항에서 멈추고 $q \approx 0$이라고 결론짓는 것. 살아남는 첫 번째 항은 이차항이다.
- (ii)에서 $\ln p < 0$으로 나눌 때 부등호의 방향을 바꾸는 것을 잊는 것.

STEP 너머

소문항 (ii)는 점근 해석(asymptotic analysis)의 한 예이다. 조건은 곱 $n\lambda^2$에 의해 좌우되므로 n의 자연스러운 척도는 $1/\lambda$이고, p^n 같은 양은 $\mathrm{e}^{-n\lambda^2/2}$처럼 행동한다. 이항분포의 푸아송 근사와 확률론의 여러 "드문 사건" 극한도 같은 논리에 바탕을 두고 있다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 이러한 극한 정리들을 꼼꼼하게 전개한다.

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.