풀이

STEP II 2002, 문제 12

확률과 통계 • 정규근사와 푸아송 근사

STEP II 2002, 12번 문제

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

step.maths.org/questions/02-s2-q12

핵심 아이디어

이 문제에는 세 겹의 무작위성이 차례로 들어 있다. 동전 하나 안에서 나오는 앞면, 하루 안의 동전들, 그리고 한 주 안의 날들이다. 각 층은 이항(binomial) 개수이고, 그 "성공" 확률(probability)은 바로 아래 층에서 온다. 주어진 공식은 맨 아래 층을 정규분포(normal distribution)로 근사하고(연속성 수정(continuity correction)을 적용한다), 가운데 층을 푸아송 분포(Poisson distribution)로 근사한 다음 다시 정규분포로 근사하며, 맨 위 층은 정확한 이항분포로 그대로 두어서 만들어진다. 세 층이 보이고 나면, 할 일은 공식의 각 식이 어떤 근사에서 나왔는지 알아내고, 주어진 수치에서 각 근사가 쓸 만한지 따져 보는 것이다.

풀이

하루, 동전 하나

하루와 동전 하나를 고정한다. X를 그 동전을 M번 던져서 나온 앞면의 개수라 하자. 각 던지기는 독립(independent)이고 앞면이 나올 확률이 각각 p이므로

$$X \sim \mathrm{B}(M, p), \qquad \mathrm{E}(X) = Mp, \qquad \mathrm{Var}(X) = Mp(1-p).$$

M이 클 때 X를 정규확률변수 $X' \sim \mathrm{N}(Mp,\, Mp(1-p))$로 근사한다. X는 정숫값을 가지므로 "앞면이 m개 미만"은 $X \le m-1$을 뜻하고, 연속성 수정을 하면 이는 $X' < m - \frac{1}{2}$가 된다. 따라서

$$r = \mathrm{P}(X < m) = \mathrm{P}(X \le m-1) \approx \Phi\left(\frac{m - \frac{1}{2} - Mp}{\sqrt{Mp(1-p)}}\right) = \Phi\left(\frac{2m - 1 - 2Mp}{2\sqrt{Mp(1-p)}}\right).$$

문제와 비교하면, 이는 $h \approx Lr$임을 말해 준다.

하루, 모든 동전

Y를 그날 L개의 동전 가운데 앞면이 m개 미만으로 나온 동전의 개수라 하자. 동전들은 독립적으로 던져지고 모두 똑같이 행동하므로, 각 동전은 독립적으로 확률 r로 "성공"하고

$$Y \sim \mathrm{B}(L, r), \qquad \mathrm{E}(Y) = Lr = h, \qquad \mathrm{Var}(Y) = Lr(1-r) = h(1-r).$$

생각해 보기 q의 공식에서 분모에는 $\sqrt{h}$가 있다. 왜 표준편차(standard deviation)가 $\sqrt{h(1-r)}$가 아니라 $\sqrt{h}$일까?

이 공식은 r가 작은 경우, 곧 동전 하나에서 앞면이 m개 미만으로 나오는 것은 드문 사건이고 L은 큰 경우를 위해 만들어졌다. 이 영역에서는 $1-r \approx 1$이므로 분산(variance)이 평균과 거의 같고, 이는 정확히 푸아송 근사 $Y \approx \mathrm{Po}(h)$의 상황이다. h도 어느 정도 크다면 $\mathrm{Po}(h)$는 다시 근사적으로 $\mathrm{N}(h,h)$이다. (같은 말로, $\mathrm{B}(L,r)$를 곧바로 정규분포로 근사하고 $1-r$를 1로 바꾸어도 된다.)

그러므로 $Y' \sim \mathrm{N}(h,h)$로 둔다. "동전이 l개 초과"는 $Y \geq l+1$을 뜻하고, 연속성 수정을 하면 $Y' > l + \frac{1}{2}$가 된다. 따라서

$$q = \mathrm{P}(Y > l) \approx 1 - \Phi\left(\frac{l+\frac{1}{2}-h}{\sqrt{h}}\right) = \Phi\left(\frac{h-l-\frac{1}{2}}{\sqrt{h}}\right) = \Phi\left(\frac{2h-2l-1}{2\sqrt{h}}\right),$$

여기서 표준정규분포의 대칭성(symmetry) $1-\Phi(z) = \Phi(-z)$를 썼다. 이것이 주어진 q이다.

모든 날

하루 동안 앞면이 m개 미만으로 나온 동전이 l개보다 많으면 그날을 **좋은 날**이라 부르자. 날마다 좋은 날일 확률은 똑같이 q이고, 서로 다른 날에는 서로 다른 던지기가 관련되므로 날들은 독립이다. 따라서 K일 가운데 좋은 날의 수는 $\mathrm{B}(K,q)$를 따르고

$$\mathrm{P}(\text{좋은 날이 정확히 } k\text{일}) = \binom{K}{k} q^k (1-q)^{K-k},$$

이며, q는 위와 같이 근사한다. 이것이 구하려던 결과이다.

수치 계산

여기서는 $Mp = 60$, $Mp(1-p) = 24$이므로

$$\frac{2m-1-2Mp}{2\sqrt{Mp(1-p)}} = \frac{96-1-120}{2\sqrt{24}} = \frac{-25}{9.798} \approx -2.55,$$

이고, $r \approx \Phi(-2.55) \approx 0.0054$, $h = 500r \approx 2.68$이다. 그러면

$$\frac{2h-2l-1}{2\sqrt{h}} \approx \frac{5.36-9}{3.27} \approx -1.11, \qquad q \approx \Phi(-1.11) \approx 0.133,$$

이고, 근사값은 $\binom{7}{2}q^2(1-q)^5 \approx 21 \times 0.0178 \times 0.490 \approx 0.18$이다.

> **생각해 보기** 각 근사는 모두 표준적인 것이다. 그 가운데 어느 것을 믿을 만한 조건 밖에서 쓰고 있는가?

세 층을 차례로 살펴보자.

- **맨 위 층.** $\mathrm{B}(7, q)$ 단계는 (q가 주어지면) 정확하므로 그 자체로는 오차를 만들지 않는다. 그러나 $k = 2$이므로 q의 오차를 제곱으로 키운다.
- **맨 아래 층.** $M = 100$이고 $Mp = 60$, $M(1-p) = 40$이면 이항분포의 정규근사를 흔히 쓰는 범위에 넉넉히 들어간다. 그러나 여기서는 이를 꼬리(tail) 쪽으로 표준편차의 2.55배만큼 들어간 곳에서 써서 약 0.005인 확률을 구하고 있다. 정규근사는 분포의 대부분에서 **절대**오차(absolute error)로는 좋지만, 꼬리에서는 **상대**오차(relative error)가 클 수 있고, 특히 $p \neq \frac{1}{2}$이어서 이항분포가 치우쳐(skewed) 있을 때 그렇다. r의 상대오차는 모두 그대로 $h = Lr$로 넘어가고, q는 h에 민감하다.
- **가운데 층.** $\mathrm{B}(500, r)$를 $\mathrm{Po}(h)$로 바꾸는 것은 아주 훌륭하다. r가 매우 작고 L이 크기 때문이다. 그러나 그다음에 $h \approx 2.7$인 $\mathrm{Po}(h)$를 정규분포로 근사하는 것은 정당화되지 않는다. 평균이 3보다 작은 푸아송 분포는 눈에 띄게 치우쳐 있고, 흔히 쓰는 경험 법칙은 평균이 적어도 약 10(어떤 이는 15나 20)이기를 요구한다.

그러므로 근사가 정확하리라고 기대하지는 않는다. 의심스러운 단계는 $\mathrm{B}(100, 0.6)$의 꼬리에 대한 정규근사와 평균이 작은 푸아송 분포에 대한 정규근사이다.

실제로 r의 정확한 값은 $\mathrm{P}(X \leq 47) \approx 0.00576$으로 정규근사 추정값보다 약 7% 크므로, h는 약 2.88이어야 한다. 그러면 정확한 $\mathrm{P}(Y > 4)$는 0.133이 아니라 약 0.165이다. 최종 확률은 0.18이 아니라 약 0.23이다. (이 특정한 수치에서는 $\mathrm{Po}(2.68)$의 정규근사가 우연히 푸아송 분포 자체와 거의 같은 꼬리 확률을 주므로(0.133 대 0.134), 피해의 대부분은 맨 아래 층에서 생긴다. 이런 행운을 미리 기대할 수는 없었을 것이다.)

검산

- 연속성 수정의 방향이 맞다. "m개 미만"은 $X \leq m - 1$이므로 $m - \frac{1}{2}$에서 멈추고, "l개 초과"는 $Y \geq l + 1$이므로 $l + \frac{1}{2}$에서 시작한다. 주어진 공식에도 정확히 $2m - 1$과 $2l + 1$이 들어 있다.
- 수치로는 $\Phi(-2.5516) = 0.005362$, $h = 2.681$, $q = \Phi(-1.1108) = 0.1333$이고, 결과는 0.1825이다. 정확한 이항분포 합으로 계산하면 $r = 0.005761$, $q = 0.1645$(2×10^4일에 대한 모의실험으로 0.164를 얻어 확인했다)이고 최종 답은 0.2314이다.

흔한 실수

- $m - \frac{1}{2}$ 대신 m을 쓰는 것(또는 $l + \frac{1}{2}$ 대신 l을 쓰는 것). 그러면 공식이 맞지 않는다.
- $\sqrt{h(1-r)}$를 쓴 뒤 $\sqrt{h}$에 이르지 못하는 것. 문제의 공식은 사실상 푸아송 근사를 쓰고 있기 때문에 인수 $1-r$를 버린 것이고, 이를 밝혀 써야 한다.
- "$Mp > 5$이고 $M(1-p) > 5$"만으로 정확도를 판단하는 것. 이 규칙은 분포의 중심에 관한 것이지, 0.005인 꼬리 확률에 관한 것이 아니다.

STEP 너머

이 문제는 근사를 여러 번 이어 쓸 때 오차가 어떻게 전파되는지, 그리고 꼬리에서 절대오차와 상대오차가 어떻게 다른지에 관한 것이다. 대학에서는 중심극한정리(central limit theorem)가 베리-에센 정리(Berry–Esseen theorem)로 더 정밀해지는데, 이 정리는 정규근사의 절대오차에 상한을 준다. 꼬리 확률에는 체르노프 부등식(Chernoff bounds)이나 큰 편차 이론(large-deviation theory)처럼 다른 도구가 필요하며, 이들은 꼬리 확률이 얼마나 빨리 줄어드는지를 기술한다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 푸아송 근사와 정규근사, 그리고 체르노프 유형의 부등식을 명료하게 다룬다. Grimmett과 Welsh의 *Probability: An Introduction*(확률론: 입문)도 함께 보기에 좋다.

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.