풀이

STEP I 1995, 문제 14

확률과 통계 • 비율의 추정

> **STEP I 1995, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/95-s1-q14

핵심 아이디어

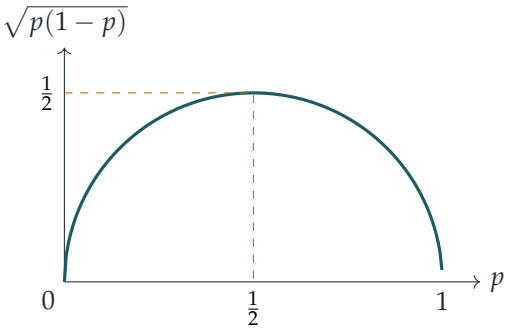
표본비율(sample proportion)은 이항(binomial) 개수를 n으로 나눈 것이다. n이 크면 표본비율은 평균(mean)이 p이고 분산(variance)이 $p(1-p)/n$인 정규분포(normal distribution)를 근사적으로 따른다. 문제는 분산이 바로 우리가 모르는 p에 의존한다는 점이다. 소문항 (i)이 이 문제를 해결한다. 소문항 (i)은 표준편차(standard deviation)가 가질 수 있는 가장 나쁜 값을 알려 주므로, 최악의 경우에 맞추어 정한 표본 크기(sample size)는 p가 무엇이든 통한다. 이어서 소문항 (iii)은 p에 대해 대략적인 정보를 이미 알고 있을 때 무엇이 달라지는지 묻는다.

풀이

소문항 (i)

완전제곱식(completing the square)을 만들면 $f(p) = p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2$이다. 따라서 $[0,1]$에서 $f(p) \le \frac{1}{4}$이고, 등호는 $p = \frac{1}{2}$일 때만 성립한다. 제곱근 함수는 증가함수이므로

$$\max_{0 \le p \le 1} \sqrt{p(1-p)} = \sqrt{\tfrac{1}{4}} = \tfrac{1}{2}, \qquad \text{최댓값을 갖는 점: } p = \tfrac{1}{2}.$$



소문항 (ii)

X를 표본에 포함된 여성의 수라고 하자. 뽑힌 각 사람이 여성일 확률은 p이다. 모집단이 충분히 크다고 가정하여 n번의 선택이 사실상 독립(independent)이라고 본다(아주 큰 모집단에서의 비복원추출(sampling without replacement)은 복원추출(sampling with replacement)과 거의 같다). 그러면

$$X \sim \mathrm{B}(n,p), \qquad \hat{p} = \frac{X}{n}, \qquad \mathrm{E}(\hat{p}) = p, \qquad \mathrm{Var}(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

n이 크면 중심극한정리(central limit theorem), 즉 이항분포의 정규근사에 의해

$$\hat{p} \approx \mathrm{N}\left(p,\ \frac{p(1-p)}{n}\right), \qquad \text{따라서} \qquad Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx \mathrm{N}(0,1).$$

우리가 원하는 것은 $\mathrm{P}(|\hat{p}-p| > 0.01) < 0.01$, 즉

$$\mathrm{P}\left(|Z| > \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) < 0.01.$$

대칭성에 의해 이 1%는 양쪽 꼬리(tail)에 0.5%씩 나뉜다. $\Phi(2.576) = 0.995$이므로 다음이 필요하다.

$$\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}} > 2.576 \quad \Longleftrightarrow \quad \sqrt{n} > 257.6\sqrt{p(1-p)}.$$

생각해 보기 p를 모르는데, p가 들어 있는 부등식으로부터 어떻게 n을 정할 수 있을까?

바로 여기서 소문항 (i)이 쓰인다. p가 무엇이든 $\sqrt{p(1-p)} \le \frac{1}{2}$이다. 따라서 $\sqrt{n} > 257.6 \times \frac{1}{2} = 128.8$이면 모든 p에 대해 부등식이 성립한다. 이 조건은 필요조건이기도 하다. $p = \frac{1}{2}$일 때(그리고 여성의 비율이 그 근처일 때) 이 상계가 실제로 달성되기 때문이다. 그러므로

$$n > 128.8^2 = 16\,589.44,$$

이어야 하고, 필요한 표본 크기는

$$\boxed{n = 16\,590} \qquad (\text{약} 16\,600).$$

더 정밀한 백분율점(percentage point) $z = 2.5758$을 쓰면 $n > 16\,587.2$를 얻는다. 어느 쪽이든 답은 "약 16 600"이며, 세 번째 유효숫자(significant figure)는 사용한 표에 따라 달라진다.

소문항 (iii)

일반적으로 필요한 조건은

$$n > \left(\frac{2.576}{0.01}\right)^2 p(1-p) \approx 66\,360\, p(1-p).$$

값 16 590은 최악의 경우인 $p = \frac{1}{2}$에서 나왔다. p가 $\frac{1}{2}$에서 충분히 멀리 떨어져 있다는 것을 안다면 $p(1-p)$에 대해 더 작은 상계를 쓸 수 있다.

(a) 왼손잡이(left-handedness). 사람들 가운데 대략 10%가 왼손잡이이다. $p \le 0.15$라고 안전하게 가정할 수 있다면, $p(1-p)$는 $[0, \frac{1}{2}]$에서 증가하므로 $p(1-p) \le 0.15 \times 0.85 = 0.1275$이다. 이로부터 $n > 8461$을 얻는다. 정확히 $p = 0.1$이라면 $n > 66\,360 \times 0.09 \approx 5970$이므로 약 6000이다. 어느 쪽이든 아무것도 모를 때 필요한 16 590보다 훨씬 작은 표본으로 충분하다. np가 수백에 이르므로 정규근사도 여전히 잘 맞는다.

(b) 백만장자. 이번에는 p가 매우 작아서 아마 0.001에서 0.01 정도일 것이다. 글자 그대로 받아들이면 $p \le 0.01$에서 $p(1-p) \le 0.0099$이고 $n > 657$이다. 그러나 이는 요점을 놓친 것이다.

- p 자체가 약 0.001일 때 오차 0.01은 쓸모가 없다. 추정값 $\hat{p} = 0$만 해도 이미 참값과의 차이가 0.01 이내이기 때문이다. 우리가 정말 원하는 것은 **상대** 정확도(relative accuracy), 예를 들어 오차가 p의 1% 이하가 되는 것이다. 그러면 $n > 2.576^2(1-p)/(p \times 0.01^2)$이 필요하고, $p = 0.001$일 때 이 값은 약 6.6×10^7이다. 엄청나게 더 크다.
- p가 아주 작고 n이 수백에 불과하면 np는 1보다 작다. 그러면 이항분포는 매우 치우쳐 있어서 정규근사는 더 이상 타당하지 않으며, 푸아송 근사(Poisson approximation)가 더 나은 모형이 된다.
- 실제로는 백만장자에게 접근하기 어렵고, 사람들은 자신의 재산에 대해 사실대로 말하지 않을 수도 있으므로 무작위 추출과 정직한 응답 모두 의심스럽다.

따라서 왼손잡이의 경우에는 p를 대략 알면 더 작은 표본을 쓸 수 있다. 백만장자의 경우에는 합리적인 정확도 요구 조건이라면 어떤 것이든 훨씬 더 큰 표본이 필요하거나, 아예 다른 표본추출 방법이 필요하다.

검산

- $p = \frac{1}{2}$, $n = 16\,590$일 때 정규근사로 $\mathrm{P}(|\hat{p} - p| > 0.01) = 2\big(1 - \Phi(2.5762)\big) \approx 0.00999$를 얻으며, 이는 1%보다 아주 조금 작다.
- $n = 16\,590$, $p = \frac{1}{2}$에 대해 정확한 이항확률을 더하면 약 0.0102이다. 이 작은 초과분은 이항분포의 이산성(연속성 수정(continuity correction)을 하지 않은 정규근사)에서 온다. $p = 0.3$이면 정확한 값은 약 0.005로, 소문항 (i)이 예측하듯이 여유 있게 기준 안에 든다.
- $p = 0.1$, $n = 6000$일 때 정확한 이항확률은 약 0.0098로, (iii)(a)의 수치를 확인해 준다.

흔한 실수

- 2.576 대신 1.96이나 2.326을 쓰는 것. 1%는 양쪽 꼬리에 나뉘므로 $\Phi(z) = 0.995$가 필요하다.
- n을 p에 대한 식으로 남겨 두는 것. 문제는 어떤 p에 대해서도 통하는 하나의 수를 원하며, 소문항 (i)이 있는 이유가 바로 이것이다.
- "$\sqrt{p(1-p)} \leq \frac{1}{2}$"로부터 $n > 128.8$을 얻는다고 읽는 것. 128.8보다 커야 하는 것은 n이 아니라 $\sqrt{n}$이다.

STEP 너머

이것은 여론조사에서 인용되는 "오차 한계(margin of error)" 뒤에 있는 표준적인 계산이다. 약 1000명을 대상으로 한 여론조사는 신뢰도 95%에서 대략 ±3퍼센트포인트의 오차 한계를 주는데, 이는 정확히 소문항 (i)의 최악의 경우 논법에 따른 것이다. 대학에서는 이것이 신뢰구간(confidence interval)으로 이어지고, 드문 비율을 어떻게 추정할 것인가 하는 문제로 이어진다. 그런 경우에는 정규근사 대신 푸아송 방법과 정확한 이항 방법이 쓰인다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 중심극한정리와 이러한 근사를 매우 읽기 쉽게 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.