풀이

STEP III 1998, 문제 14

확률과 통계 • 불편추정량

STEP III 1998, 14번 문제

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

step.maths.org/questions/98-s3-q14

이 문제는 *Beyond STEP*의 13.3 절 **독일 전차 문제**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch13/german-tanks.html

핵심 아이디어

Z_2는 독립(independent)인 조각들의 합이므로, 그 평균(mean)과 분산(variance)은 균등(uniform) 확률변수 하나의 평균과 분산에서 곧바로 나온다. Z_1은 최댓값(maximum)에 의존하는데, 최댓값을 다루는 자연스러운 방법은 분포함수(distribution function)를 이용하는 것이다. X_i 가운데 가장 큰 값이 m 이하인 것은 **모든** X_i가 m 이하일 때와 정확히 같다. 이것은 확률의 곱으로 바뀌고, 이를 미분하면 밀도(density)를 얻는다.

풀이

Z_2의 평균

각 X_i는 $[0, N]$에서 밀도 $1/N$을 가지므로

$$\mathrm{E}(X_i) = \frac{N}{2}, \qquad \mathrm{E}(X_i^2) = \frac{1}{N}\int_0^N x^2\,\mathrm{d}x = \frac{N^2}{3}, \qquad \mathrm{Var}(X_i) = \frac{N^2}{3} - \frac{N^2}{4} = \frac{N^2}{12}.$$

기댓값의 선형성(linearity of expectation)에 의해

$$\mathrm{E}(Z_2) = \frac{2}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{2}{n}\cdot n\cdot\frac{N}{2} = N.$$

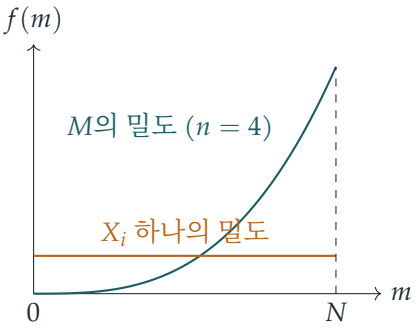
최댓값의 분포

$M = \max\{X_1, \ldots, X_n\}$이라 하자. $0 \le m \le N$일 때, 사건 $M \le m$은 모든 i에 대하여 $X_i \le m$일 때 정확히 일어나고, X_i들은 독립이므로

$$\mathrm{P}(M \le m) = \prod_{i=1}^{n} \mathrm{P}(X_i \le m) = \left(\frac{m}{N}\right)^n.$$

이를 미분하면 M의 밀도는

$$f(m) = \frac{n\, m^{n-1}}{N^n}, \qquad 0 \le m \le N.$$



최댓값은 N 근처에 몰려 있으며, 그래서 불편(unbiased) 추정량이 되는 데 작은 보정 인수만 있으면 된다. 그 적률(moment)은 다음과 같다.

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n\, m^{n-1}}{N^n}\, \mathrm{d}m = \frac{n}{N^n} \cdot \frac{N^{n+1}}{n+1} = \frac{n}{n+1} N,$$
$$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n\, m^{n-1}}{N^n}\, \mathrm{d}m = \frac{n}{n+2} N^2.$$

Z_1의 평균

$$\mathrm{E}(Z_1) = \frac{n+1}{n}\, \mathrm{E}(M) = \frac{n+1}{n} \cdot \frac{n}{n+1} N = N.$$

따라서 요구한 대로 두 추정량(estimator)의 평균은 모두 N이다.

분산

Z_1에 대해서는 먼저

$$\mathrm{Var}(M) = \frac{n}{n+2} N^2 - \frac{n^2}{(n+1)^2} N^2 = \frac{n\left[(n+1)^2 - n(n+2)\right]}{(n+2)(n+1)^2} N^2 = \frac{n}{(n+2)(n+1)^2} N^2,$$

인데, $(n+1)^2 - n(n+2) = 1$이기 때문이다. 여기에 $\left(\frac{n+1}{n}\right)^2$을 곱하면

$$\boxed{\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2} \cdot \frac{n}{(n+2)(n+1)^2} N^2 = \frac{N^2}{n(n+2)}.}$$

Z_2에 대해서는 X_i들이 독립이므로 분산이 더해진다.

$$\boxed{\operatorname{Var}(Z_2) = \frac{4}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{4}{n^2}\cdot n\cdot\frac{N^2}{12} = \frac{N^2}{3n}.}$$

? **생각해 보기** 어느 쪽이 더 작은가? 그리고 항상 같은 쪽이 더 작은가?

두 분산의 비는

$$\frac{\operatorname{Var}(Z_1)}{\operatorname{Var}(Z_2)} = \frac{3n}{n(n+2)} = \frac{3}{n+2}.$$

이 값은 $n = 1$일 때 1이고(이때 $Z_1 = Z_2 = 2X_1$이므로 두 추정량은 실제로 같다), $n \geq 2$인 모든 경우에 1보다 작다.

어느 쪽을 택할 것인가

둘 다 불편추정량이므로 퍼진 정도를 비교한다. 우리는 Z_1을 택한다.

- $n \geq 2$이면 Z_1의 분산이 순(strictly) 더 작고, 그 이점은 n이 커질수록 커진다. $\operatorname{Var}(Z_1)$은 $1/n^2$처럼 줄어들지만 $\operatorname{Var}(Z_2)$는 $1/n$처럼만 줄어든다. 예를 들어 $n = 10$이면 Z_1의 표준편차(standard deviation)는 약 $0.09N$인 반면 Z_2의 표준편차는 약 $0.18N$이다.
- Z_1은 항상 관측된 가장 큰 식별 번호 이상인데, N 역시 반드시 그래야 한다. 반면 Z_2는 실제로 관측한 번호보다 작게 나올 수 있어서(예를 들어 관측값이 $1, 2, 30$이면 $Z_2 = 22$이다), 불가능하다고 이미 알고 있는 추정값을 줄 수 있다.

검산

- $n = 1$일 때 두 공식은 모두 $N^2/3$을 준다. $Z_1 = Z_2 = 2X_1$이고 $\operatorname{Var}(2X_1) = 4\cdot N^2/12$이므로 당연히 그래야 한다.
- $N = 10$으로 200 000번 시행한 모의실험(simulation)에서 두 추정량 모두 표본평균이 10과 0.01 이내로 차이 났고, $n = 4$일 때 표본분산은 4.16과 8.33으로, 이론값 $\frac{100}{24} \approx 4.17$과 $\frac{100}{12} \approx 8.33$과 일치했다.

흔한 실수

- 최댓값의 밀도를 직접 다루려고 하는 것. 먼저 $\mathrm{P}(M \leq m)$을 거쳐야 한다. 독립성 덕분에 이것은 간단한 곱이 된다.
- 상수를 제곱하는 것을 잊는 것. $\operatorname{Var}(cY) = c^2\operatorname{Var}(Y)$이므로 곱하는 인수는 $\left(\frac{n+1}{n}\right)^2$과 $\frac{4}{n^2}$이다.

STEP 너머

이것은 고전적인 "독일 전차 문제(German tank problem)"로, 제2차 세계대전 때 연합군의 통계학자들이 일련번호로부터 독일 전차 생산량을 추정하는 데 사용했으며, 그 정확도는 놀라울 정도였다. 여기서 최댓값은 **충분통계량**(*sufficient statistic*)이다. 즉 최댓값을 알고 나면 표본의 나머지는 N에 대해 더 이상 아무것도 알려 주지 않는다. 그러면 레만–셰페 정리(Lehmann–Scheffé theorem)에 의해 Z_1이 모든 불편추정량 가운데 분산이 가장 작다는 것이 밝혀진다. 충분성과 효율적인 추정량은 대학 1학년 통계학에서 다룬다. 순서통계량(order statistics)을 공부하기 시작하기에는 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)가 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.