13.3 독일 전차 문제

학습지

문제

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

힌트

먼저 혼자 힘으로 문제를 풀어 보자. 막히면 힌트를 **하나씩** 읽고, 힌트를 하나 읽을 때마다 다시 문제로 돌아가자. **잠깐, 직접 해 보기**라고 표시된 힌트가 핵심 단계이다. 각 힌트 뒤에는 풀이를 적을 공간이 있다.

생각해 보기 M이 $[0, N]$ 위의 서로 독립인 균등분포(uniform distribution) 값 n개 중 가장 큰 값이라면, $\mathrm{P}(M \leq m)$은 무엇인가?

1단계 둘 다 불편추정량이다

다루는 내용: Z_1과 Z_2의 평균

생각해 보기 Z_2의 평균은 금방 구할 수 있다. 각 X_i의 평균은 얼마인가?

잠깐, 직접 해 보기 Z_1에 대해서는 최댓값 M의 분포(distribution)를 구한다. $M \leq m$인 것은 언제인가? 이를 미분하여 확률밀도함수(density)를 얻은 다음 $\mathrm{E}(M)$을 구한다.

2단계 두 추정량의 분산

다루는 내용: 분산

생각해 보기 $\mathrm{E}(M^2)$을 구하여 $\mathrm{Var}(M)$을 얻은 다음, 상수배를 하여 $\mathrm{Var}(Z_1)$을 구한다. Z_2에 대해서는 X_i 하나의 분산(variance)이 얼마인가? X_i들은 서로 독립(independent)이다.

3단계 어느 쪽이 더 나은가?

다루는 내용: 어느 쪽을 택할지

> **생각해 보기** 한 분산을 다른 분산으로 나누어 본다. n이 커질 때 이 비는 어떻게 변하는가? 두 추정량(estimator) 중 어느 하나라도 실제로 관찰한 일련번호보다 작아질 수 있는가?

숨은 수학

충분성

최댓값이 왜 이렇게 훨씬 나은가? 가장 큰 값 M이 주어지면, 나머지 값들은 N이 무엇이든 그저 $[0, M]$ 위의 균등분포를 따를 뿐이다. 즉, 나머지 값들은 N에 관해 더 이상 아무런 정보도 주지 않는다. 최댓값은 **충분통계량**(sufficient statistic)이다. C. R. Rao와 David Blackwell의 정리(1945–47)에 따르면, 어떤 불편추정량이든 충분통계량이 주어졌을 때의 조건부 기댓값으로 바꾸면 개선되거나, 적어도 나빠지지는 않는다. 여기서는 그렇게 하면 Z_2가 Z_1이 된다. 나아가 **레만–셰페 정리**(Lehmann–Scheffé theorem)를 쓰면 Z_1이 가능한 가장 작은 분산을 갖는 불편추정량임을 보일 수 있다. 또한 Z_1의 분산은 $\frac{1}{n^2}$처럼 줄어드는데, 이는 대부분의 추정량이 보이는 $\frac{1}{n}$보다 빠르다는 점에도 주목하라. 이런 일이 생기는 까닭은 N이 분포의 끝점이고, 자료가 그 끝점 쪽으로 몰려 쌓이기 때문이다.

이산적인 경우

실제 일련번호는 자연수 $1, 2, \ldots, N$이며, 비복원추출(sampling without replacement)로 뽑힌다. 이때 가장 좋은 불편추정값은 $\hat{N} = m + \frac{m}{k} - 1$인데, 여기서 m은 관측한 일련번호 k개 중 가장 큰 값이다. 즉, 최댓값에 관측값 사이의 평균 간격을 더한 것이다. 주로 변속기와 바퀴의 일련번호에 바탕을 둔 이런 종류의 추정값을 써서, 연합군 통계학자들은 독일의 생산량을 한 달에 약 246대로 추정하였다. 나중에 독일 측 기록에서 확인된 수치는 245대였고, 첩보 보고가 제시한 수치는 1400대였다.

역사 속의 일련번호

이 방법은 그 뒤로도 여러 번 쓰였다. 예를 들어 분석가들은 처음 판매된 몇 개의 일련번호로부터 새로운 소비자 제품이 몇 개나 만들어졌는지를 추정하였다. 제조사들도 교훈을 얻어, 지금은 많은 제조사가 일련번호를 단순한 순서로 읽히지 않도록 암호화한다. Richard Ruggles와 Henry Brodie는 1947년 *Journal of the American Statistical Association*에 전시의 이 연구를 소개하였다.

더 도전하기

1 **평균제곱오차.** 편향추정량 M에 대하여 $\mathrm{E}(M-N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M)-N\big)^2$이다. 이 평균제곱오차(mean squared error)를 계산하고 $\mathrm{Var}(Z_1)$과 비교하라. $\mathrm{E}(cM-N)^2$을 최소로 하는 상수 c를 구하라.

2 **최솟값도.** $(n+1)\min\{X_i\}$도 N의 불편추정량임을 보이고, 그 분산을 구하라. 이 추정량은 쓸 만한가?

3 **이산 공식.** 일련번호 $1,\dots,N$에서 k개를 비복원추출하고, 그 최댓값을 m이라 하자. $\mathrm{P}(\max \le m) = \binom{m}{k}/\binom{N}{k}$임을 보이고, $k=1$일 때 $m + \frac{m}{k} - 1$이 불편추정량임을 확인하라.

4 **전차.** 노획한 전차 5대의 일련번호가 19, 40, 42, 60, 98이라면, 만들어진 전차의 수를 추정하라.

더 읽을거리

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**, *Journal of the American Statistical Association* 42 (1947). 원래의 기록이다.
- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). 편향, 분산, 충분성을 다룬다.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican).

안내에 따른 전체 풀이는 제 13.3절에 있다:
https://beyondstep.keithmcnulty.org/ko/ch13/german-tanks.html

매주 STEP 문제 하나를 끝까지 풀어 본다. The STEP Up은 이메일로 무료로 받아 볼 수 있다: keithmcnulty.substack.com