풀이

STEP III 2005, 문제 12

확률과 통계 • 불편추정량의 결합

> **STEP III 2005, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/05-s3-q12

핵심 아이디어

모든 것은 독립(independent)인 확률변수(random variable)들의 선형결합(linear combination)에 관한 두 가지 사실에 기대고 있다. 기댓값(expectation)은 더해지고, 분산(variance)은 계수의 제곱을 곱하여 더해진다. 실제 랩 타임이 **어떤 값이든** 기댓값이 올바르게 나와야 한다는 조건은 계수 비교(matching coefficients)를 뜻하며, 이로써 각 조건이 상수들에 대한 방정식이 된다. 그러면 X는 경주 시간의 합리적인 추정값(estimate)이 되고, Y는 두 가지 계측 방법 사이의 불일치로부터 미지의 σ를 추정하는 수단이 된다.

풀이

실제 랩 타임을 $\mu_1, \ldots, \mu_4$라 하고, 실제 경주 시간을 $\mu = \mu_1 + \mu_2 + \mu_3 + \mu_4$라 하자. 다음과 같이 둔다.

$$S = T_1 + T_2 + T_3 + T_4.$$

그러면 $\mathrm{E}(S) = \mu_1 + \cdots + \mu_4 = \mu$이고, $T_1, \ldots, T_4$가 서로 독립이므로(계측자들이 서로 독립이므로) $\mathrm{Var}(S) = 4\sigma^2$이다. 또한 $\mathrm{E}(T) = \mu$, $\mathrm{Var}(T) = \sigma^2$이며, S는 나머지 네 계측자에만 의존하므로 T는 S와 독립이다.

X 구하기

$X = aT + bS$에 대하여

$$\mathrm{E}(X) = a\mu + b\mu = (a+b)\mu.$$

이 값은 실제 랩 타임이 무엇이든 μ와 같아야 하므로, 특히 어떤 $\mu \neq 0$에 대해서도 그러하다. 따라서

$$a + b = 1.$$

T와 S가 독립이므로

$$\mathrm{Var}(X) = a^2 \mathrm{Var}(T) + b^2 \mathrm{Var}(S) = \sigma^2(a^2 + 4b^2) = \sigma^2(a^2 + 4(1-a)^2).$$

이차식 $a^2 + 4(1-a)^2 = 5a^2 - 8a + 4$는 최고차항의 계수가 양수이므로, 도함수 $10a - 8$이 0이 되는 $a = \frac{4}{5}$에서 최솟값을 가진다. (완전제곱식(completing the square)으로 나타내면 $5a^2 - 8a + 4 = 5(a - \frac{4}{5})^2 + \frac{4}{5}$이다.) 따라서 $a = \frac{4}{5}$, $b = \frac{1}{5}$이고

$$X = \tfrac{4}{5}T + \tfrac{1}{5}(T_1 + T_2 + T_3 + T_4), \qquad \mathrm{Var}(X) = \tfrac{4}{5}\sigma^2.$$

이는 이치에 맞는다. T는 S보다 네 배 믿을 만하므로(분산이 $4\sigma^2$에 대해 σ^2), 네 배의 가중치를 받는다.

Y 구하기

$Y = cT + dS$에 대하여 $\mathrm{E}(Y^2) = \mathrm{Var}(Y) + (\mathrm{E}(Y))^2$을 이용하면

$$\mathrm{E}(Y^2) = \sigma^2(c^2 + 4d^2) + (c+d)^2\mu^2.$$

생각해 보기 실제 랩 타임이 무엇이든 이 값이 어떻게 σ^2과 같을 수 있을까?

상수 c, d는 μ에 의존할 수 없지만, μ는 임의의 양수가 될 수 있다. 이 식은 $\sigma^2(c^2 + 4d^2)$에 μ^2의 상수배를 더한 것이므로, μ가 변해도 σ^2과 같게 유지되려면 μ^2의 계수가 0이 되어야 한다. 즉 $(c+d)^2 = 0$이므로 $d = -c$이다. 그러면 $\sigma^2(c^2 + 4c^2) = \sigma^2$, 곧 $c^2 = \frac{1}{5}$이어야 한다. $c = 1/\sqrt{5}$로 택하면

$$Y = \frac{T - (T_1 + T_2 + T_3 + T_4)}{\sqrt{5}}$$

이다(반대 부호를 택하면 $-Y$가 되는데, Y^2만 중요하므로 이것도 똑같이 성립한다). 따라서 Y는 두 계측값 사이의 불일치를 나타낸다. 그 기댓값은 0이고, 그 제곱은 σ^2의 불편(unbiased) 추정값이다.

주어진 경주의 경우

$T = 220$, $S = 220.5$이면

$$X = \tfrac{4}{5}(220) + \tfrac{1}{5}(220.5) = 176 + 44.1 = 220.1,$$
$$Y = \frac{220 - 220.5}{\sqrt{5}} = -\frac{0.5}{\sqrt{5}}, \qquad Y^2 = \frac{0.25}{5} = 0.05.$$

따라서 $\sigma^2 \approx 0.05$로 추정하며, 이로부터

$$\mathrm{Var}(X) = \tfrac{4}{5}\sigma^2 \approx \tfrac{4}{5}(0.05) = 0.04, \qquad \text{표준편차} X \approx 0.2 \text{ 초}.$$

그러므로 실제 경주 시간의 추정값은 220.1 ± 0.2초이다. 즉 표준편차(standard deviation) 하나만큼 양쪽으로 잡으면, 실제 시간은 대략 구간

$$[219.9,\ 220.3] \text{ 초}$$

안에 있다. 더 높은 신뢰도를 원한다면 표준편차 두 배만큼 양쪽으로 잡아 $[219.7,\ 220.5]$초를 얻을 수 있다. 오차가 정규분포(normal distribution)를 따른다면 이는 약 95% 구간이 된다. 다만 σ^2은 Y^2의 값 하나로부터 추정한 것이므로, 구간의 폭 자체도 대략적인 추정일 뿐임을 염두에 두어야 한다.

검산

- $a = \frac{4}{5}$일 때 $\mathrm{Var}(X) = \frac{4}{5}\sigma^2$은 $\mathrm{Var}(T) = \sigma^2$과 $\mathrm{Var}(S) = 4\sigma^2$보다 모두 작다. 서로 독립인 두 측정값을 결합하면 마땅히 이렇게 되어야 한다.
- $X = 220.1$은 두 측정값 사이에 있으며, 더 믿을 만한 쪽인 $T = 220$에 훨씬 가깝다.
- 200 000회의 경주($\sigma = 0.7$)로 시뮬레이션한 결과, X의 평균은 실제 시간과 0.001 이내로 차이가 났고, $\mathrm{Var}(X) \approx 0.393$으로 $\frac{4}{5}(0.49) = 0.392$와 비슷했으며, Y^2의 평균은 0.491로 $\sigma^2 = 0.49$와 비슷했다.

흔한 실수

- $\mathrm{Var}(S) = \sigma^2$ 또는 $16\sigma^2$으로 잡는 것. 이 값은 독립인 네 분산의 합인 $4\sigma^2$이다.
- $\mathrm{E}(Y^2)$에서 $(\mathrm{E}\,Y)^2$ 항을 빠뜨리는 것. 바로 이 항 때문에 $d = -c$가 된다.
- $\frac{4}{5}$을 곱하지 않고 $Y^2 = 0.05$를 그대로 X의 분산으로 쓰는 것.

STEP 너머

X는 두 추정값의 선형결합 가운데 분산이 최소인 불편 결합이다. 일반적인 규칙은 서로 독립인 각 추정값에 그 분산의 역수만큼 가중치를 주는 것이며, 이는 메타분석(meta-analysis)에서 쓰이는 역분산 가중(inverse-variance weighting)과 최소제곱법(least squares)에 관한 가우스(Gauss)–마르코프(Markov) 정리(Gauss–Markov theorem)의 바탕이 된다. Y^2은 σ^2의 불편 추정량(unbiased estimator)으로, 제수가 $n - 1$인 표본분산(sample variance)과 비슷하다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 합의 기댓값

과 분산을 꼼꼼하게 다루며, 추정량과 신뢰구간(confidence interval)은 대학의 첫 통계학 강의에서 본격적으로 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.