풀이

STEP III 2010, 문제 1

순수수학 • 평균과 분산

> **STEP III 2010, 1번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/10-s3-q1

핵심 아이디어

A와 B는 처음 n개 수의 평균(mean)과 분산(variance)이고, C와 D는 수를 하나 더 추가했을 때의 평균과 분산이다. 분산은 "제곱의 평균에서 평균의 제곱을 뺀 것"의 꼴로 다룰 때 가장 쉬운데, 바로 소문항 (ii)가 이 꼴을 준다. 이를 이용하면 D는 x_{n+1}에 대한 이차식(quadratic)으로 나오고, 두 부등식(inequality)은 모두 이 이차식에 관한 명제가 된다.

풀이

풀이 전체에서 간단히 $x = x_{n+1}$로 쓴다.

소문항 (i)

정의에 의해 $\sum_{k=1}^{n} x_k = nA$이므로

$$C = \frac{1}{n+1}\left(\sum_{k=1}^{n} x_k + x\right) = \frac{nA + x}{n+1}.$$

소문항 (ii)

제곱을 전개하고 $\sum_{k=1}^{n} x_k = nA$와 $\sum_{k=1}^{n} 1 = n$을 이용하면

$$\begin{aligned} B &= \frac{1}{n}\sum_{k=1}^{n}(x_k^2 - 2Ax_k + A^2) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \frac{2A}{n}\cdot nA + \frac{1}{n}\cdot nA^2 \\ &= \frac{1}{n}\sum_{k=1}^{n} x_k^2 - A^2, \end{aligned}$$

이므로 요구된 식이 성립한다.

소문항 (iii)

소문항 (ii)는 어떤 수의 목록에 대해서도 성립하므로, 이를 (평균이 C인) $n+1$개의 수 $x_1, \ldots, x_{n+1}$에 적용하면

$$D = \frac{1}{n+1}\sum_{k=1}^{n+1} x_k^2 - C^2.$$

다시 소문항 (ii)에서 $\sum_{k=1}^{n} x_k^2 = n(B+A^2)$이므로 $\sum_{k=1}^{n+1} x_k^2 = n(B+A^2) + x^2$이다. C에 대해 소문항 (i)을 이용하면

$$D = \frac{nB + nA^2 + x^2}{n+1} - \frac{(nA+x)^2}{(n+1)^2}.$$

양변에 $(n+1)^2$을 곱하고 전개하면

$$\begin{aligned}(n+1)^2 D &= (n+1)(nB + nA^2 + x^2) - (n^2A^2 + 2nAx + x^2)\\ &= n(n+1)B + (n^2 + n - n^2)A^2 - 2nAx + (n+1-1)x^2\\ &= n(n+1)B + n(x^2 - 2Ax + A^2)\\ &= n(n+1)B + n(x-A)^2.\end{aligned}$$

따라서

$$\boxed{D = \frac{n}{n+1}B + \frac{n}{(n+1)^2}(x_{n+1} - A)^2.}$$

첫 번째 부등식. $n+1$을 곱하면

$$(n+1)D = nB + \frac{n}{n+1}(x_{n+1} - A)^2 \;\geq\; nB,$$

인데, 마지막 항이 제곱에 음이 아닌 수를 곱한 것이기 때문이다. 이는 x_{n+1}의 모든 값에 대해 성립하며, 등호는 정확히 $x_{n+1} = A$일 때 성립한다.

> **생각해 보기** 음수의 제곱근을 걱정하지 않고 $D < B$를 x_{n+1}에 대한 조건으로 바꾸려면 어떻게 해야 할까?

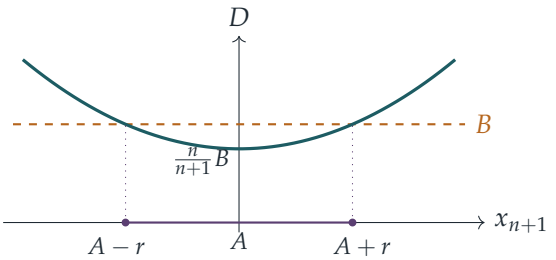
두 번째 부등식. 먼저 $B \geq 0$임에 주목한다. B는 제곱들의 평균이기 때문이다. 따라서 $\sqrt{(n+1)B/n}$은 실수이다. 이제 상자 안의 공식을 이용하면, 아래의 각 단계는 거꾸로도 성

립한다(양변에서 같은 양을 빼고 양수 $(n+1)^2/n$을 곱하기만 하기 때문이다).

$$\begin{aligned} D < B &\iff \frac{n}{(n+1)^2}(x_{n+1}-A)^2 < B - \frac{n}{n+1}B = \frac{B}{n+1} \\ &\iff (x_{n+1}-A)^2 < \frac{(n+1)B}{n} \\ &\iff |x_{n+1}-A| < \sqrt{\frac{(n+1)B}{n}} \\ &\iff A - \sqrt{\frac{(n+1)B}{n}} < x_{n+1} < A + \sqrt{\frac{(n+1)B}{n}}. \end{aligned}$$

세 번째 단계는 양변의 음이 아닌 제곱근을 취하는데, 이는 음이 아닌 수들 사이의 엄격한 부등식을 보존한다. 이것이 요구된 "필요충분조건(if and only if)"이다. ($B = 0$이면 구간은 공집합이고, 실제로 이때는 항상 $D \geq 0 = B$이다.)

아래 그림은 $n = 3, A = 0, B = 1$일 때 D를 x_{n+1}의 함수로 나타낸 것이다. 이는 $x_{n+1} = A$에서 최솟값 $\frac{n}{n+1}B$를 갖는 포물선이며, 정확히 위에서 구한 구간에서 높이 B 아래로 내려간다.



여기서 $r = \sqrt{(n+1)B/n}$이다.

검산

- $n = 1$로 두자. 그러면 $B = 0$이고, 공식은 $D = \frac{1}{4}(x_2 - x_1)^2$을 주는데, 이는 평균이 중점인 두 수 x_1, x_2에 대해 옳은 값이다.
- $x_1, x_2, x_3 = 0, 0, 3$으로 두면 $A = 1, B = 2$이고, 여기에 $x_4 = 1 = A$를 추가한다. 직접 계산하면 $D = \frac{1}{4}(1+1+4+0) = \frac{3}{2} = \frac{3}{4}B$로, 최솟값이다.
- 20 000개의 무작위 목록에 대해 세 결과를 모두 (분수로) 정확히 확인했으며, 여기에는 동치 관계 $D < B \iff (x_{n+1}-A)^2 < (n+1)B/n$도 포함된다. 또한 구간의 양 끝점이 유리수인 216가지 경우에 대해 두 끝점에서 $D = B$임을 확인했다.

흔한 실수

- 소문항 (ii)에서 전개할 때 $\sum_{k=1}^{n} A^2 = nA^2$이고 A^2이 아니라는 것을 잊는 것.

- D를 $\sum x_k^2$로 나타내고 거기서 멈추는 것. 문제는 B, A, x_{n+1}, n으로 나타내기를 요구하므로, 소문항 (ii)를 이용해 합을 바꿔 써야 한다.
- $(x_{n+1} - A)^2 < k$의 제곱근을 취해 $x_{n+1} - A < \sqrt{k}$만 얻고, 구간의 아래쪽 끝을 놓치는 것.

STEP 너머

D에 대한 공식은 **갱신 공식**(*updating formula*)이다. 이 공식을 쓰면 컴퓨터가 모든 데이터를 저장하지 않고도 데이터 스트림의 평균과 분산을 값 하나씩 받아 가며 유지할 수 있는데, 이 아이디어는 Welford 알고리즘(Welford's algorithm)에서 더 정교해진다. 항 $n(x_{n+1} - A)^2/(n+1)^2$은 통계학의 분산분석(analysis of variance)의 바탕이 되는, 분산을 "집단 내" 부분과 "집단 간" 부분으로 분해하는 것의 작은 예이다. 더 자세한 내용은 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)을 보라.

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.