풀이

STEP II 1993, 문제 15

확률과 통계 • 중심극한정리

STEP II 1993, 15번 문제

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

step.maths.org/questions/93-s2-q15

이 문제는 *Beyond STEP*의 13.6 절 쌓이는 오차에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch13/errors-add-up.html

핵심 아이디어

전체 오차는 수 하나마다 하나씩 생기는 작고 서로 독립인 오차들의 합이다. 따라서 하나의 수에서 생기는 오차의 평균과 분산을 구하고, 이를 모든 수에 대하여 더한 다음, 중심극한정리(Central Limit Theorem)를 이용하여 전체 오차를 근사적으로 정규분포로 다룬다. 두 컴퓨터의 차이는 전적으로 평균에서 나온다. 한 기계의 오차는 평균적으로 서로 상쇄되지만, 다른 기계의 오차는 모두 같은 방향으로 쏠린다.

풀이

수 하나에서 생기는 오차

수를 $a = m + U$로 나타내자. 여기서 $m = \lfloor a \rfloor$은 정수이고 U는 소수 부분(fractional part)이며, $U \sim \mathrm{U}[0,1)$이다. 오차를 (근삿값) $-$ (참값)으로 정의한다. 부호를 어떻게 정하든 오차의 크기에는 영향이 없다.

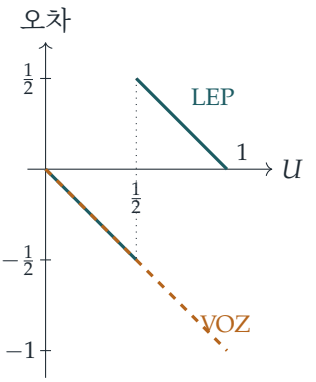
LEP (반올림(*rounding*)). $U < \frac{1}{2}$이면 수가 m으로 바뀌므로 오차는 $-U$이고, $U \geq \frac{1}{2}$이면 $m+1$로 바뀌므로 오차는 $1-U$이다. 따라서 오차는

$$X = \begin{cases} -U, & 0 \leq U < \frac{1}{2}, \\ 1-U, & \frac{1}{2} \leq U < 1, \end{cases}$$

이고, 이는 $(-\frac{1}{2}, \frac{1}{2}]$의 각 값을 일정한 밀도로 취한다. 즉 X는 $[-\frac{1}{2}, \frac{1}{2}]$에서 균등분포(uniform

distribution)를 따른다(그림의 톱니 모양).

VOZ (**버림**(*truncation*)). 수가 m으로 바뀌므로 오차는 $Y = -U$이고, $(-1, 0]$에서 균등분포를 따른다.



길이가 1인 구간에서의 균등분포는 구간이 어디에 있든 분산이 $\frac{1}{12}$이다. 따라서

$$\mathrm{E}(X) = 0, \quad \mathrm{Var}(X) = \tfrac{1}{12}; \qquad \mathrm{E}(Y) = -\tfrac{1}{2}, \quad \mathrm{Var}(Y) = \tfrac{1}{12}.$$

(참고로, $\mathrm{Var}(U) = \int_0^1 u^2 \,\mathrm{d}u - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$이다.)

1500개의 수 더하기

컴퓨터는 근삿값들을 정확하게 더하므로, 답의 오차는 각 수에서 생긴 오차의 합이다. 소수 부분들은 서로 독립이므로 각 오차도 서로 독립이고, 따라서 분산을 더할 수 있다. $n = 1500$개의 수에 대하여

$$\mathrm{Var}(\text{전체 오차}) = \frac{1500}{12} = 125, \qquad \text{표준편차} \sqrt{125} = 5\sqrt{5} \approx 11.18.$$

중심극한정리에 의하여, 독립이고 동일한 분포를 따르는 1500개 오차의 합은 정규분포에 매우 가깝다.

LEP. 전체 오차 E_L은 근사적으로 $\mathrm{N}(0, 125)$를 따르므로

$$\begin{aligned}\mathrm{P}(|E_L| > 15) &\approx \mathrm{P}\left(|Z| > \frac{15}{\sqrt{125}}\right) = \mathrm{P}(|Z| > 1.342) \\ &= 2\big(1 - \Phi(1.342)\big) \approx 2(1 - 0.9102) \approx 0.18.\end{aligned}$$

오차는 정수가 아니라 연속확률변수이므로 연속성 보정(continuity correction)은 필요 없다.

VOZ. 전체 오차 E_V는 근사적으로 $\mathrm{N}(-750, 125)$를 따른다. 오차의 크기가 15 이하가 되려면 $E_V \geq -15$이어야 하는데, 이는 평균보다

$$\frac{-15 - (-750)}{\sqrt{125}} = \frac{735}{11.18} \approx 66$$

표준편차만큼 위에 있는 값이다. 이 확률은 완전히 무시할 수 있을 만큼 작으므로

$$\mathrm{P}(|E_V| > 15) \approx 1.$$

버림은 언제나 아래쪽으로, 평균적으로 $\frac{1}{2}$만큼 오차를 내므로, 1500개의 수를 더하고 나면 VOZ는 약 750만큼 틀리게 된다.

생각해 보기 두 기계의 오차는 분산이 같다. 그런데 왜 답이 이렇게 다를까?

전체 오차의 퍼짐은 약 ± 11로 같지만, 그 **중심**은 다르다. LEP의 오차는 중심이 0이므로 전체 오차는 0 근처에서 오르내린다. VOZ의 오차는 모두 평균이 $-\frac{1}{2}$이므로, 편향(bias)은 n에 비례하여 커지는 반면 무작위적인 퍼짐은 $\sqrt{n}$ 정도로만 커진다. 편향이 쉽게 이긴다.

LEP는 몇 번 더할 수 있는가?

LEP가 n개의 수를 더한다고 하자. 전체 오차는 근사적으로 $\mathrm{N}\left(0, \frac{n}{12}\right)$를 따르고, 다음이 필요하다.

$$\mathrm{P}(|E_L| < 10) = 2\Phi\left(\frac{10}{\sqrt{n/12}}\right) - 1 \;\geq\; 0.9 \quad\Longleftrightarrow\quad \Phi\left(\frac{10}{\sqrt{n/12}}\right) \geq 0.95.$$

Φ는 증가함수이고 $\Phi(1.6449) = 0.95$이므로, 이는

$$\frac{10}{\sqrt{n/12}} \geq 1.6449 \iff n \leq 12\left(\frac{10}{1.6449}\right)^2 \approx 443.5.$$

와 같다. 이 확률은 n이 커질수록 감소하므로, $n \leq 443$일 때 0.9 이상을 유지하고 $n = 444$에서 처음으로 0.9 아래로 떨어진다. 따라서 LEP는 **443**개의 수를 더할 수 있다. ("덧셈"을 $+$ 기호를 한 번 쓰는 것으로 센다면, 443개의 수를 더하는 데는 442번의 덧셈이 필요하다. 어느 쪽으로 해석하든 어떤 뜻인지 밝히기만 하면 된다.)

검산

- 1.342는 그럴듯한가? 전체 오차 15는 표준편차(11.18)의 한 배를 조금 넘고, 정규분포에서 평균으로부터 1.34 표준편차보다 멀리 떨어진 부분은 약 18%이다. 대략 맞다.
- **수치적으로.** 반올림한 수 1500개씩으로 이루어진 묶음 10^5개를 시뮬레이션하면 $\mathrm{P}(|E_L| > 15) \approx 0.179$를 얻는다. 균등분포 합의 정확한 분포로 계산하면 0.1797이고, 정규근사로는 0.1797이다. 두 번째 부분에서 정확한 확률은 $n = 443$일 때 0.90019, $n = 444$일 때 0.89981이므로 경계값이 확인된다.

흔한 실수

- LEP의 오차가 $[0, \frac{1}{2}]$에서 균등분포를 따른다고 보는 것. 반올림은 올라갈 수도 내려갈 수도 있으므로 오차는 $[-\frac{1}{2}, \frac{1}{2}]$에서 균등분포를 따른다.
- VOZ에서 수 하나당 평균 $-\frac{1}{2}$를 잊어버려서 LEP와 같은 답을 내는 것.
- 마지막 부분에서 0.95 대신 $\Phi(z) = 0.9$를 쓰는 것. 사건 $|E| < 10$은 양쪽에 걸쳐 있으므로 각 꼬리에 0.05씩 돌아간다.

STEP 너머

이것은 부동소수점 연산(floating-point arithmetic)에서 반올림 오차를 분석하는 고전적인 방법이다. 편향 없는 반올림을 쓰면 누적 오차가 $\sqrt{n}$ 정도로 커지지만, 편향된 규칙을 쓰면 n 정도로 커진다. IEEE 754 표준이 절반을 항상 올림하는 대신 "짝수로 반올림(round half to even)"을 쓰는 것도 이 때문이다. n개의 균등분포 확률변수의 합은 정확한 분포(어윈–홀 분포(Irwin–Hall distribution))를 가지며, 이 분포가 얼마나 빨리 정규분포에 가까워지는지 살펴보는 것은 중심극한정리에 입문하는 좋은 방법이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)에서 이 정리를 증명하고 논의한다.

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.