풀이

STEP III 2017, 문제 13

확률과 통계 • 기댓값과 분산

> **STEP III 2017, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/17-s3-q13

핵심 아이디어

$X - x$를 $(X - \mu) + (\mu - x)$로 쓴다. 첫째 조각은 평균이 0이므로, 제곱하여 기댓값(expectation)을 취하면 교차항이 사라진다. 그러면 분산(variance)에 상수를 더한 것만 남는데, 이는 "$\mathrm{E}(X^2) = \mathrm{Var}(X) + \mu^2$"와 같은 항등식이다. 균등분포(uniform distribution)의 경우, Y는 X의 함수이지만 일대일(one-to-one)이 아니므로, Y의 확률밀도함수(density)를 구하는 가장 안전한 길은 누적분포함수(cumulative distribution function)를 거치는 것이다.

풀이

함수 V

임의의 고정된 수 x에 대하여

$$\begin{aligned}\mathrm{V}(x) = \mathrm{E}((X-x)^2) &= \mathrm{E}(((X-\mu)+(\mu-x))^2) \\ &= \mathrm{E}((X-\mu)^2) + 2(\mu-x)\,\mathrm{E}(X-\mu) + (\mu-x)^2 \\ &= \sigma^2 + (x-\mu)^2\end{aligned}$$

이다. 여기서 기댓값의 선형성(linearity of expectation)(여기서 $\mu - x$는 상수이다)과 $\mathrm{E}(X - \mu) = 0$을 이용하였다.

Y의 평균

$\mathrm{V}(x) = \sigma^2 + (x - \mu)^2$는 모든 실수 x에 대하여 성립하므로 $x = X$를 대입할 수 있다.

$$Y = \mathrm{V}(X) = \sigma^2 + (X - \mu)^2.$$

생각해 보기 V는 기댓값으로 정의되었으니 $\mathrm{V}(X)$는 그냥 하나의 수가 아닌가?

그렇지 않다. V는 실변수의 평범한 함수(x에 대한 이차식(quadratic))이다. 기댓값은 x를 고정한 채로 취한 것이다. 이 이차식에 확률변수(random variable) X를 넣으면 새로운 확률변수가 된다. 따라서

$$\mathrm{E}(Y) = \sigma^2 + \mathrm{E}((X-\mu)^2) = \sigma^2 + \sigma^2 = 2\sigma^2,$$

이고, 이것이 $(*)$이다.

균등분포의 경우

X가 $[0,1]$에서 균등분포를 따르면 $\mu = \frac{1}{2}$이고 $\sigma^2 = \int_0^1 x^2\,\mathrm{d}x - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$이다. 따라서

$$\mathrm{V}(x) = \tfrac{1}{12} + (x-\tfrac{1}{2})^2 = x^2 - x + \tfrac{1}{3}.$$

Y의 분포. 이제 $Y = \frac{1}{12} + (X-\frac{1}{2})^2$이다. X가 $[0,1]$ 전체를 움직이면 $|X-\frac{1}{2}|$는 $[0,\frac{1}{2}]$ 전체를 움직이므로, Y는 $[\frac{1}{12}, \frac{1}{12}+\frac{1}{4}] = [\frac{1}{12}, \frac{1}{3}]$의 값을 가진다.

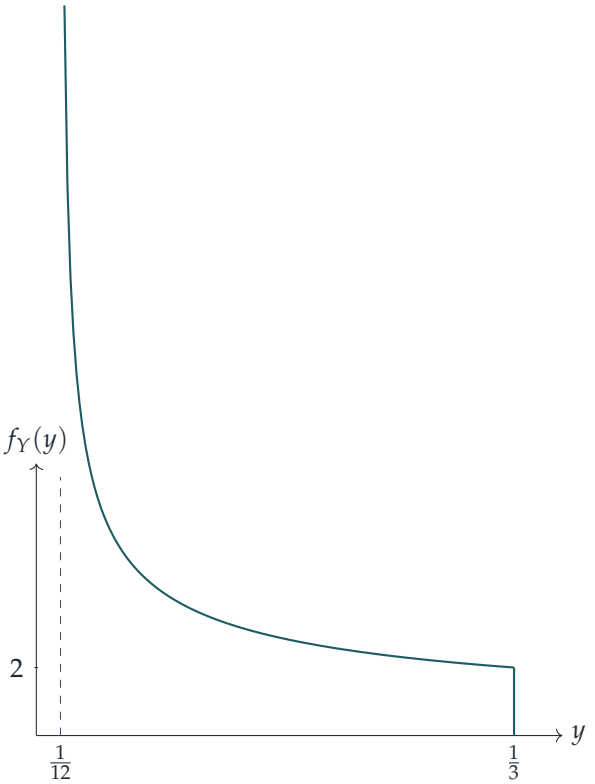
사상 $X \mapsto Y$는 이대일(두 점 $\frac{1}{2} \pm h$가 같은 Y를 준다)이므로 분포함수(distribution function)를 가지고 다룬다. $\frac{1}{12} \le y \le \frac{1}{3}$일 때

$$\begin{aligned}\mathrm{P}(Y \le y) &= \mathrm{P}((X-\tfrac{1}{2})^2 \le y - \tfrac{1}{12}) = \mathrm{P}(\tfrac{1}{2} - s \le X \le \tfrac{1}{2} + s), \quad \text{여기서} s = \sqrt{y - \tfrac{1}{12}},\\ &= 2s = 2\sqrt{y - \tfrac{1}{12}},\end{aligned}$$

이다. $0 \le s \le \frac{1}{2}$이므로 이 구간은 $[0,1]$ 안에 놓이고 길이가 $2s$이기 때문이다. ($y = \frac{1}{3}$에서 이 값은 $2 \cdot \frac{1}{2} = 1$로, 마땅히 그래야 하는 값이 나온다.) 미분하면

$$f_Y(y) = \frac{1}{\sqrt{y - \frac{1}{12}}} \quad \text{단, } \tfrac{1}{12} < y \le \tfrac{1}{3}, \qquad f_Y(y) = 0 \text{ 그 외의 경우.}$$

이 확률밀도함수는 $y = \frac{1}{12}$ 근처에서 유계가 아니지만 여전히 적분가능(integrable)하다. $(X-\frac{1}{2})^2$이 $X = \frac{1}{2}$ 근처에서 평평하므로, Y는 최솟값 가까이에 있을 가능성이 가장 크다.



$(*)$ **확인하기.** 확률밀도함수를 이용하고 $t = y - \frac{1}{12}$로 치환하면(t는 0부터 $\frac{1}{4}$까지 움직인다)

$$\mathrm{E}(Y) = \int_{1/12}^{1/3} \frac{y}{\sqrt{y - \frac{1}{12}}}\,\mathrm{d}y = \int_0^{1/4} (t^{1/2} + \tfrac{1}{12}t^{-1/2})\,\mathrm{d}t = \left[\tfrac{2}{3}t^{3/2} + \tfrac{1}{6}t^{1/2}\right]_0^{1/4}$$
$$= \tfrac{2}{3}\cdot\tfrac{1}{8} + \tfrac{1}{6}\cdot\tfrac{1}{2} = \tfrac{1}{12} + \tfrac{1}{12} = \tfrac{1}{6}.$$

$2\sigma^2 = 2\cdot\frac{1}{12} = \frac{1}{6}$이므로 이 결과는 $(*)$와 일치한다.

검산

- 확률밀도함수를 적분하면 1이다: $\int_0^{1/4} t^{-1/2}\,\mathrm{d}t = 2\cdot\frac{1}{2} = 1$.
- $\mathrm{V}(x)$는 $x = \mu$에서 가장 작고 최솟값은 σ^2이다. 즉 평균은 제곱 거리의 기댓값을 최소로 만드는 점이다. 균등분포의 경우 $\mathrm{V}(\frac{1}{2}) = \frac{1}{12}$로, Y의 치역의 맨 아래 값이다.
- 균등분포 값 400 000개로 컴퓨터 시뮬레이션을 하였더니 Y의 표본평균이 약 0.1667로 $\frac{1}{6}$과 일치하였다.

흔한 실수

- $\mathrm{V}(X)$를 상수 σ^2로 취급하여 $\mathrm{E}(Y) = \sigma^2$를 얻는 것.
- 일대일 변수변환(change of variable) 공식 $f_Y = f_X\,|\mathrm{d}x/\mathrm{d}y|$를 한쪽 가지에만 적용하여 올바른 확률밀도함수의 절반을 얻는 것. $X = \frac{1}{2} + s$와 $X = \frac{1}{2} - s$가 모두 기여한다.

STEP 너머

$\mathrm{V}(x) = \sigma^2 + (x-\mu)^2$는 가장 단순한 편향–분산 분해(bias–variance decomposition)이다. 어떤 양을 x로 추정한다면, 평균제곱오차(mean squared error)는 데이터의 퍼짐에 편향(bias)의 제곱을 더한 것으로 나뉜다. $\mathrm{E}(Y) = 2\sigma^2$라는 결과는 X의 서로 독립(independent)인 두 복사본 사이의 제곱 거리의 기댓값이 $2\sigma^2$라는 뜻이다. 통계학자들은 이 사실을 거듭 사용하는데, 예를 들어 표본분산(sample variance)이 왜 $n-1$로 나누는지 설명할 때 쓴다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 이 두 아이디어를 모두 다루며, 확률변수의 단조(monotonic)가 아닌 함수의 확률밀도함수를 구하는 방법도 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.