풀이

STEP II 2005, 문제 14

확률과 통계 • 혼합분포

> **STEP II 2005, 14번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/05-s2-q14

핵심 아이디어

이 밀도함수(density)는 우리가 이미 잘 아는 두 밀도함수, 즉 표준정규분포(standard normal)와 $[0, \lambda]$ 위의 균등분포(uniform)를 가중 혼합(mixture)한 것이다. 따라서 f의 모든 적분은 정규 부분과 균등 부분으로 나뉘며, 각 부분은 잘 알려진 사실들(전체 확률 1, 평균, $\mathrm{E}(X^2)$)로부터 어려운 적분 없이 바로 읽어 낼 수 있다. 그러면 분산(variance)은 $\mathrm{E}(X^2) - \mu^2$로부터 얻는다.

풀이

상수 k, 평균, 분산

ϕ와 g는 모두 확률밀도함수(probability density function)이므로, 각각을 실수 전체에서 적분하면 1이다. 따라서

$$1 = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = k\,[1 + \lambda \cdot 1] \quad \Longrightarrow \quad k = \frac{1}{1+\lambda}.$$

평균(mean)을 구할 때는 두 성분의 평균을 이용한다. 표준정규분포의 평균은 0이고, $[0, \lambda]$ 위의 균등밀도함수 g의 평균은 $\frac{1}{2}\lambda$이다. 따라서

$$\mu = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = k\left[\int x\phi(x)\,\mathrm{d}x + \lambda \int x g(x)\,\mathrm{d}x\right] = \frac{1}{1+\lambda}\left[0 + \lambda \cdot \frac{\lambda}{2}\right] = \frac{\lambda^2}{2(1+\lambda)}.$$

같은 방법으로, 표준정규분포에서는 $\int x^2\phi(x)\,\mathrm{d}x = 1$이고(평균이 0이므로 이 값은 분산과 같다), 또

$$\int x^2 g(x)\,\mathrm{d}x = \frac{1}{\lambda}\int_0^{\lambda} x^2\,\mathrm{d}x = \frac{\lambda^2}{3}.$$

따라서

$$\mathrm{E}(X^2) = \frac{1}{1+\lambda}\left[1 + \lambda \cdot \frac{\lambda^2}{3}\right] = \frac{3+\lambda^3}{3(1+\lambda)}.$$

? **생각해 보기** $\mathrm{E}(X^2)$는 이렇게 나눌 수 있는데, 분산은 왜 그럴 수 없을까?

기댓값(expectation)은 적분이고 적분은 선형이므로, $\mathrm{E}(X^2)$는 f가 성분 밀도함수들의 가중 결합인 것과 똑같은 방식으로 성분들의 이차 적률(second moment)의 가중 결합이 된다. 분산은 선형이 아니므로(μ^2가 들어 있다), 이차 적률을 먼저 결합한 다음에야 μ^2를 빼야 한다. 그렇게 하면

$$\begin{aligned}\sigma^2 = \mathrm{E}(X^2) - \mu^2 &= \frac{3+\lambda^3}{3(1+\lambda)} - \frac{\lambda^4}{4(1+\lambda)^2} \\ &= \frac{4(3+\lambda^3)(1+\lambda) - 3\lambda^4}{12(1+\lambda)^2} \\ &= \frac{12 + 12\lambda + 4\lambda^3 + 4\lambda^4 - 3\lambda^4}{12(1+\lambda)^2} = \frac{\lambda^4 + 4\lambda^3 + 12\lambda + 12}{12(1+\lambda)^2},\end{aligned}$$

이고, 이것이 구하려던 결과이다.

$\lambda = 2$인 경우

이제 $k = \frac{1}{3}$이고 $[0,2]$에서 $\lambda g(x) = 1$이므로

$$f(x) = \begin{cases} \frac{1}{3}\phi(x) & x < 0 \text{ 또는} x > 2, \\ \frac{1}{3}(\phi(x) + 1) & 0 \le x \le 2. \end{cases}$$

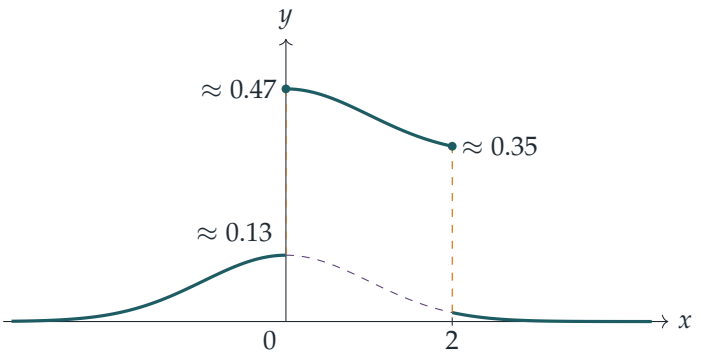
또한

$$\mu = \frac{4}{6} = \frac{2}{3}, \qquad \sigma^2 = \frac{16+32+24+12}{12 \cdot 9} = \frac{84}{108} = \frac{7}{9}, \qquad \sigma = \frac{\sqrt{7}}{3}.$$

소문항 (i)

이 곡선은 표준정규분포의 종 모양 곡선을 3분의 1 높이로 줄인 다음, 구간 $[0,2]$에서 $\frac{1}{3}$만큼 들어 올린 것이다. 따라서 $x = 0$에서는 위로 뛰어오르고 $x = 2$에서는 아래로 뚝 떨어지며, 각 도약의 높이는 $\frac{1}{3}$이다. $\phi(0) = 1/\sqrt{2\pi} \approx 0.399$와 $\phi(2) \approx 0.054$를 이용하면 다음과 같다.

- 0의 바로 왼쪽에서 $f \approx 0.133$이고, 0의 바로 오른쪽에서 $f \approx 0.466$(최댓값)이다.
- 2의 바로 왼쪽에서 $f \approx 0.351$이고, 2의 바로 오른쪽에서 $f \approx 0.018$이다.
- $[0,2]$에서 곡선은 정규분포 곡선의 모양을 따라 감소한다.
- $x \to \pm\infty$일 때 $f \to 0$이고, 모든 곳에서 $f > 0$이다.



$[0,2]$ 위의 자주색 점선 곡선은 비교를 위해 그린 $\frac{1}{3}\phi(x)$이고, 실선 곡선은 $y = f(x)$이다. (세로 축척은 늘려서 그렸다.)

소문항 (ii)

$F(x) = \mathrm{P}(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$라 하자. $\frac{1}{3}\phi$를 적분하면 $\frac{1}{3}\Phi$가 되고, $[0,2]$에서 추가된 $\frac{1}{3}$은 $[0,2]$ 가운데 x보다 왼쪽에 있는 부분의 길이에 $\frac{1}{3}$을 곱한 만큼 기여한다. 따라서

$$F(x) = \begin{cases} \frac{1}{3}\Phi(x) & x < 0, \\ \frac{1}{3}\big(\Phi(x) + x\big) & 0 \le x \le 2, \\ \frac{1}{3}\big(\Phi(x) + 2\big) & x > 2. \end{cases}$$

각 조각은 $x = 0$과 $x = 2$에서 값이 일치하는데, 당연히 그래야 한다(점 질량(point mass)이 없으면 누적분포함수(CDF)는 연속이다). 또 $x \to \infty$일 때 $F(x) \to \frac{1}{3}(1+2) = 1$이다.

소문항 (iii)

여기서 $\mu + 2\sigma = \frac{2}{3} + \frac{2}{3}\sqrt{7} \approx 2.43$이고 이는 2보다 크므로, 그 점에서는 F의 마지막 식을 사용한다. X는 연속확률변수이므로

$$\begin{aligned} \mathrm{P}(0 < X < \mu + 2\sigma) &= F\big(\tfrac{2}{3} + \tfrac{2}{3}\sqrt{7}\big) - F(0) \\ &= \tfrac{1}{3}\big(\Phi(\tfrac{2}{3} + \tfrac{2}{3}\sqrt{7}) + 2\big) - \tfrac{1}{3}\Phi(0) \\ &= \tfrac{1}{3}\big(0.9921 + 2 - 0.5\big) = \frac{2.4921}{3} = 0.8307. \end{aligned}$$

검산

- $\lambda \to 0$이면 균등 부분이 사라진다. 이때 공식은 $\mu \to 0$과 $\sigma^2 \to \frac{12}{12} = 1$을 주는데, 당연히 그래야 하듯 이는 표준정규분포의 값이다.

- $\lambda = 2$일 때 전체 분산의 법칙(law of total variance, 아래 참조)을 이용하면 다른 방법으로 같은 답을 얻는다. 두 성분의 가중치는 $\frac{1}{3}, \frac{2}{3}$, 분산은 $1, \frac{1}{3}$, 평균은 $0, 1$이다. 분산들의 평균은 $\frac{1}{3} + \frac{2}{9} = \frac{5}{9}$, 평균들의 분산은 $\frac{2}{3} - \frac{4}{9} = \frac{2}{9}$이고, 그 합은 $\frac{7}{9}$이다.
- $\lambda = 0.5, 1, 2, 3.7$에 대해 f를 수치적분(중점 공식, 0과 λ에서 구간을 나눔)한 결과, 전체 확률 1, 평균, 분산 공식이 10^{-6}보다 작은 오차로 재현되었다. $\lambda = 2$일 때는 $\mu = 0.6667$과 $\sigma^2 = 0.7778$을 얻었다.
- $\Phi(2.4305)$를 직접 계산하면 문제에 주어진 0.9921이 아니라 0.99246이 되고, 확률도 0.8307이 아니라 0.8308이 된다. 이 작은 차이는 문제에 주어진 값에서 비롯된 것이며, 우리는 그 주어진 값을 사용해야 한다.

흔한 실수

- 인수 $k = \frac{1}{3}$을 빠뜨려서 전체 확률이 3이 되는 것.
- 분산을 $k[1 + \lambda \cdot \frac{\lambda^2}{12}]$로, 즉 성분들의 분산을 결합하여 계산하는 것. 혼합분포의 분산은 선형으로 결합되지 않는다. 선형으로 결합되는 것은 이차 적률이다.
- f를 연속인 곡선으로 그리는 것. 실제로는 $x = 0$과 $x = \lambda$에서 값이 뛴다.

STEP 너머

가중치 $w_i \geq 0$의 합이 1일 때, $\sum_i w_i f_i(x)$ 꼴의 밀도함수를 **혼합분포**(mixture distribution)라 한다. 이는 확률 w_i로 성분 i를 고른 다음 f_i에서 표본을 뽑는 분포이다. 여기서 가중치는 $\frac{1}{1+\lambda}$와 $\frac{\lambda}{1+\lambda}$이다. 혼합분포의 분산은 모습만 바꾼 전체 분산의 법칙으로, 성분 분산들의 평균에 성분 평균들의 분산을 더한 것이다. 혼합분포는 통계학과 기계학습(machine learning)에서 군집화(clustering) 방법의 기초가 된다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 조건부 기댓값(conditional expectation)과 전체 분산의 법칙을 명료하게 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.