풀이

STEP III 2014, 문제 12

확률과 통계 • 로그정규분포

> **STEP III 2014, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/14-s3-q12

핵심 아이디어

$y = \mathrm{e}^x$는 순증가(strictly increasing) 함수이므로, Y에 관한 명제는 로그를 취하기만 하면 X에 관한 명제로 바뀐다. 즉 $Y \leq y$일 필요충분조건은 $X \leq \ln y$이다. 이로부터 Y의 중앙값(median)과 밀도(density)를 곧바로 얻는다. 평균(mean)을 구할 때의 요령은 지수에서 완전제곱식을 만들어(complete the square) 피적분함수를 또 다른 정규분포의 밀도로 바꾸는 것이다. 그 적분값이 1이라는 것은 이미 알고 있다.

풀이

소문항 (i)

e^x는 순증가하므로 사건 $\{Y \leq \mathrm{e}^{x_m}\}$과 $\{X \leq x_m\}$은 같은 사건이다. 따라서

$$\mathrm{P}(Y \leq \mathrm{e}^{x_m}) = \mathrm{P}(X \leq x_m) = \tfrac{1}{2},$$

이므로 Y의 중앙값은

$$y_m = \mathrm{e}^{x_m}.$$

이다. (중앙값은 이처럼 임의의 증가함수를 그대로 통과하지만, 소문항 (iv)에서 보듯이 평균은 그렇지 않다.)

소문항 (ii)

Y는 양수 값만 가진다. $y > 0$에 대하여 G를 Y의 누적분포함수(cumulative distribution function)라 하자. 앞과 같은 동치 관계를 이용하면

$$G(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(X \leq \ln y) = F(\ln y).$$

연쇄법칙(chain rule)으로 미분하고 $F' = f$를 이용하면

$$g(y) = G'(y) = f(\ln y) \cdot \frac{1}{y} = \frac{f(\ln y)}{y}, \qquad y > 0,$$

이 되어 원하는 결과를 얻는다. ($y \le 0$이면 $g(y) = 0$이다.)

최빈값(mode) λ는 g가 최댓값을 갖는 점이다. 이 점은 g가 미분가능한 열린구간 $(0, \infty)$ 안에 있으므로 $g'(\lambda) = 0$이다. 몫의 미분법(quotient rule)에 의해

$$g'(y) = \frac{f'(\ln y) \cdot \frac{1}{y} \cdot y - f(\ln y)}{y^2} = \frac{f'(\ln y) - f(\ln y)}{y^2}.$$

$y^2 \neq 0$이므로 $g'(\lambda) = 0$으로 두면

$$f'(\ln \lambda) = f(\ln \lambda),$$

을 얻고, 이것이 구하는 조건이다.

소문항 (iii)

함수

$$\frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-(x-\mu-\sigma^2)^2/(2\sigma^2)}$$

는 정확히 $\mathrm{N}(\mu + \sigma^2, \sigma^2)$ 확률변수(random variable)의 확률밀도함수(probability density function)이다. 즉 주어진 정규분포(normal distribution)의 밀도에서 μ를 $\mu + \sigma^2$로 바꾼 것이다. 모든 확률밀도함수는 실수 전체에서 적분하면 1이므로, 주어진 적분값이 설명된다.

이제

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{e}^X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^x \, \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \, \mathrm{d}x.$$

합쳐진 지수에서 완전제곱식을 만든다. $(x-\mu-\sigma^2)^2 = (x-\mu)^2 - 2\sigma^2 x + 2\mu\sigma^2 + \sigma^4$로 전개하면

$$\begin{aligned} x - \frac{(x-\mu)^2}{2\sigma^2} &= -\frac{(x-\mu)^2 - 2\sigma^2 x}{2\sigma^2} \\ &= -\frac{(x-\mu-\sigma^2)^2 - 2\mu\sigma^2 - \sigma^4}{2\sigma^2} \\ &= -\frac{(x-\mu-\sigma^2)^2}{2\sigma^2} + \mu + \tfrac{1}{2}\sigma^2. \end{aligned}$$

을 얻는다. 상수 $\mathrm{e}^{\mu + \frac{1}{2}\sigma^2}$를 적분 밖으로 꺼내면

$$\mathrm{E}(Y) = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-(x-\mu-\sigma^2)^2/(2\sigma^2)} \, \mathrm{d}x = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2},$$

이다. 여기서 위에서 설명한 적분을 이용하였다. 이것이 구하는 결과이다.

소문항 (iv)

세 값을 모두 명시적으로 구한다.

중앙값. 정규분포의 밀도는 μ에 대하여 대칭이므로 $x_m = \mu$이고, 소문항 (i)에 의해 $y_m = \mathrm{e}^{\mu}$이다.

최빈값. 정규분포의 밀도를 미분하면

$$f'(x) = -\frac{x-\mu}{\sigma^2} f(x).$$

> **생각해 보기** f가 결코 0이 되지 않을 때 소문항 (ii)의 조건을 어떻게 이용할 수 있을까?

모든 x에 대하여 $f(x) > 0$이므로, 조건 $f'(\ln\lambda) = f(\ln\lambda)$의 양변을 $f(\ln\lambda)$로 나누면

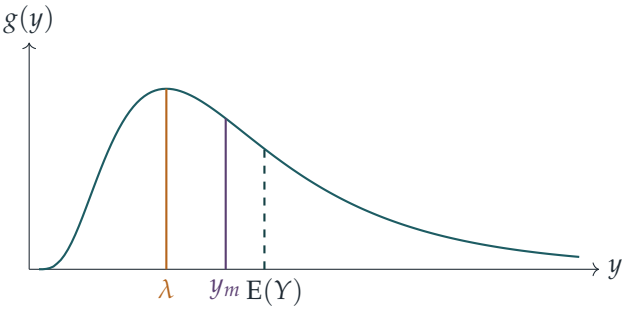
$$-\frac{\ln\lambda - \mu}{\sigma^2} = 1 \quad\Longrightarrow\quad \ln\lambda = \mu - \sigma^2 \quad\Longrightarrow\quad \lambda = \mathrm{e}^{\mu-\sigma^2}.$$

을 얻는다. 이것은 g의 유일한 정상점(stationary point)이고, Y의 최빈값이 유일하다고 주어졌으므로 이것이 최빈값이다.

비교. $\sigma > 0$이므로 $\mu - \sigma^2 < \mu < \mu + \frac{1}{2}\sigma^2$이고, e^t는 순증가하므로

$$\lambda = \mathrm{e}^{\mu-\sigma^2} \;<\; y_m = \mathrm{e}^{\mu} \;<\; \mathrm{E}(Y) = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2},$$

이 되어 원하는 결과를 얻는다. 엄격한 부등식은 $\sigma^2 > 0$에 의존하는데, 이는 진정한 정규분포라면 언제나 성립한다.



그림은 $\mu = 0, \sigma = 0.6$일 때 Y의 밀도를 나타낸다. 오른쪽으로 길게 뻗은 꼬리가 평균을 가장 오른쪽으로 끌어당기고, 중앙값은 그 가운데에 있으며, 봉우리는 둘보다 왼쪽에 있다.

검산

▶ $\sigma \to 0$이면 Y는 e^{μ}에 집중되고, 세 값 모두 e^{μ}로 수렴한다. 당연히 그래야 한다.

- $\mu = 0, \sigma = 1$이면 $\lambda = \mathrm{e}^{-1} \approx 0.368$, $y_m = 1$, $\mathrm{E}(Y) = \mathrm{e}^{1/2} \approx 1.649$이다. $\mathrm{e}^x f(x)$의 수치 적분과 g의 최댓값에 대한 수치 탐색 결과가 이 값들과 일치한다.
- 비 $\mathrm{E}(Y)/\lambda = \mathrm{e}^{3\sigma^2/2}$는 σ에만 의존한다. 이는 타당하다. μ를 바꾸는 것은 Y를 e^{μ}배로 다시 스케일링할 뿐이기 때문이다.

흔한 실수

- 연쇄법칙에서 나오는 인수 $1/y$를 빠뜨리고 Y의 밀도를 $f(\ln y)$로 쓰는 것. 이 인수가 없으면 최빈값이 중앙값과 같은 e^{μ}로 나오게 된다.
- $\mathrm{E}(\mathrm{e}^X) = \mathrm{e}^{\mathrm{E}(X)}$라고 가정하는 것. 이는 중앙값에 대해서는 참이지만 평균에 대해서는 그렇지 않다.
- 완전제곱식을 만들 때의 부호 실수. $(x - \mu - \sigma^2)^2$를 주의 깊게 전개하고 남는 상수를 확인하라.

STEP 너머

여기서 Y는 **로그정규분포**(*lognormal distribution*)를 따르며, 이 분포는 소득, 주가, 입자 크기를 모형화하는 데 쓰인다. 부등식 $\mathrm{E}(\mathrm{e}^X) \geq \mathrm{e}^{\mathrm{E}(X)}$는 볼록함수(convex function)에 대한 젠센 부등식(Jensen's inequality)의 한 경우이고, 소문항 (iii)의 계산은 정규분포의 적률생성함수(moment generating function) $\mathrm{E}(\mathrm{e}^{tX}) = \mathrm{e}^{\mu t + \frac{1}{2}\sigma^2 t^2}$에 $t = 1$을 대입한 것이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 확률변수의 변환, 적률생성함수, 로그정규분포를 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.