풀이

STEP III 1993, 문제 16

확률과 통계 • 절단정규분포

> **STEP III 1993, 16번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/93-s3-q16

핵심 아이디어

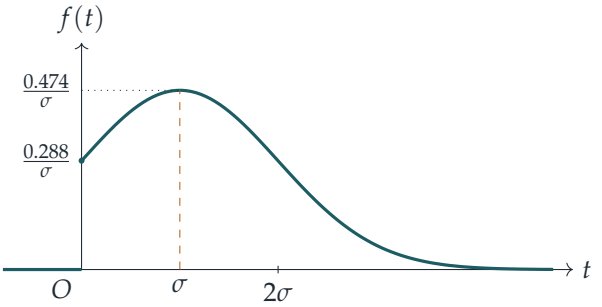
인수 $1/k$를 제외하면, 이 밀도(density)는 평균(mean)이 σ이고 표준편차(standard deviation)가 σ인 정규분포(normal distribution)의 밀도를 $t = 0$에서 잘라 낸 것과 정확히 같다. 따라서 T에 관한 모든 확률(probability)은 정규분포의 확률을, 잘라 내기 전의 확률변수가 음이 아닐 확률로 나눈 것이다. $z = (t - \sigma)/\sigma$로 표준화(standardising)하면 모든 것이 $Z \sim \mathrm{N}(0,1)$과 Φ의 표에 관한 명제로 바뀐다.

풀이

이 풀이 전체에서 $Z \sim \mathrm{N}(0,1)$이고, 그 밀도는 $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, 분포함수(distribution function)는 Φ이다.

그래프 그리기

$t \geq 0$에서 그래프는 $t = \sigma$를 중심으로 하는 정규분포의 종 모양 곡선을 $1/k$배만큼 확대한 것이다. 최댓값은 $t = \sigma$에서 갖고, 그 값은 $f(\sigma) = \frac{1}{k\sigma\sqrt{2\pi}} \approx \frac{0.474}{\sigma}$이며, 변곡점(point of inflexion)은 $t = 2\sigma$에 있다. 잘린 지점에서 곡선은 양의 높이 $f(0) = \frac{e^{-1/2}}{k\sigma\sqrt{2\pi}} \approx \frac{0.288}{\sigma}$에서 시작하고, $t < 0$에서는 0으로 뚝 떨어진다. $t \to \infty$일 때 $f(t) \to 0$이다.



k의 값

전체 확률은 1이어야 한다.

$$1 = \int_0^\infty \frac{1}{k\sigma\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(\tfrac{t-\sigma}{\sigma}\right)^2\right) \mathrm{d}t.$$

$z = (t-\sigma)/\sigma$로 치환(substitution)하면 $\mathrm{d}t = \sigma\,\mathrm{d}z$이고 $t = 0$은 $z = -1$이 되므로

$$1 = \frac{1}{k}\int_{-1}^\infty \phi(z)\,\mathrm{d}z = \frac{1-\Phi(-1)}{k} = \frac{\Phi(1)}{k}.$$

따라서 $k = \Phi(1) \approx 0.8413$이고, 요구한 대로 $k \approx 0.841$이다.

평균

같은 치환을 쓰면 $t = \sigma(1+z)$이고

$$\mathrm{E}(T) = \frac{1}{k}\int_{-1}^\infty \sigma(1+z)\,\phi(z)\,\mathrm{d}z = \sigma\left[\frac{1}{k}\int_{-1}^\infty \phi(z)\,\mathrm{d}z + \frac{1}{k}\int_{-1}^\infty z\,\phi(z)\,\mathrm{d}z\right].$$

첫 번째 적분은 방금 보였듯이 k이다.

? **생각해 보기** $\phi(z)$ 자체는 적분할 수 없는데, $z\,\phi(z)$는 어떻게 적분할까?

$\phi'(z) = -z\,\phi(z)$라는 사실을 이용한다. 즉 $z\,\phi(z)$의 원시함수(antiderivative)는 $-\phi(z)$이다.

$$\int_{-1}^\infty z\,\phi(z)\,\mathrm{d}z = \Big[-\phi(z)\Big]_{-1}^\infty = \phi(-1) = \frac{e^{-1/2}}{\sqrt{2\pi}} \approx 0.24197.$$

그러므로

$$\mathrm{E}(T) = \sigma\left(1 + \frac{e^{-1/2}}{k\sqrt{2\pi}}\right) \approx \sigma\left(1 + \frac{0.24197}{0.84134}\right) \approx 1.288\,\sigma.$$

이 값은 마땅히 σ보다 크다. 왼쪽 꼬리를 없애면 평균이 오른쪽으로 밀려나기 때문이다.

σ 구하기

문항 출제에 걸리는 평균 시간이 3시간이므로

$$\sigma = \frac{3}{1.2876} \approx 2.330\text{시간}.$$

임의의 $a \geq 0$에 대하여 같은 치환을 쓰면

$$\mathrm{P}(T > a) = \frac{1}{k}\int_{(a-\sigma)/\sigma}^{\infty} \phi(z)\,\mathrm{d}z = \frac{1-\Phi\left(\frac{a}{\sigma}-1\right)}{\Phi(1)}. \qquad (*)$$

소문항 (i)

$a = 4$이면 $\frac{4}{\sigma} - 1 \approx \frac{4}{2.330} - 1 \approx 0.717$이고, $\Phi(0.717) \approx 0.7633$이다. $(*)$에 의해

$$\mathrm{P}(T > 4) \approx \frac{1-0.7633}{0.8413} \approx 0.2814, \qquad \mathrm{P}(T \leq 4) \approx 0.7186.$$

이제 16개 문항의 출제 시간이 서로 독립(independent)이고 각각 T의 분포를 따른다고 가정한다(이것이 사용할 수 있는 유일하게 합리적인 모형이다). 그러면

$$\mathrm{P}(4\text{시간보다 오래 걸리는 문항이 없음}) = \mathrm{P}(T \leq 4)^{16} \approx 0.7186^{16} \approx 0.0051.$$

따라서 모든 문항이 4시간 안에 출제될 가능성은 매우 낮다(약 200분의 1).

소문항 (ii)

> **생각해 보기** 어떤 사건을 조건으로 하는가?

"2시간 후에도 만족스럽지 않다"는 것은 그 문항에 모두 합쳐 2시간보다 많이 필요하다는 뜻이므로, 이는 사건 $T > 2$이다. "2시간을 더 들인 후에도 여전히 받아들일 수 없다"는 $T > 4$이다. $\{T > 4\} \subseteq \{T > 2\}$이므로

$$\mathrm{P}(T > 4 \mid T > 2) = \frac{\mathrm{P}(T > 4)}{\mathrm{P}(T > 2)}.$$

$a = 2$이면 $\frac{2}{\sigma} - 1 \approx -0.142$이므로 $1 - \Phi(-0.142) = \Phi(0.142) \approx 0.5563$이다. 두 $(*)$ 식의 비에서 인수 $\Phi(1)$은 약분된다.

$$\mathrm{P}(T > 4 \mid T > 2) \approx \frac{1-0.7633}{0.5563} = \frac{0.2367}{0.5563} \approx 0.426.$$

검산

- 밀도를 직접 수치적분하면(심프슨 공식(Simpson's rule), $\sigma = 1$) k로 나누기 전의 전체 넓이는 0.84134이고, 평균은 1.2876σ이다.

- 평균과 표준편차가 $\sigma = 2.330$인 정규분포 값 4×10^5개를 모의실험(simulation)으로 만들어 음수인 값을 버리면, 표본평균은 3.004, $\mathrm{P}(T > 4) \approx 0.282$, $\mathrm{P}(T > 4 \mid T > 2) \approx 0.426$이다.
- 소문항 (i)은 반올림에 매우 민감하다. $0.719^{16} \approx 0.0051$이지만 $0.718^{16} \approx 0.0050$이다. 따라서 답은 약 0.005로 제시한다.
- 잘라 내지(truncation) 않았더라도 $\mathrm{P}(T > 4 \mid T > 2)$는 같았을 것이다. 인수 $1/k$가 약분되기 때문이다. k의 영향을 받는 것은 소문항 (i)과 평균뿐이다.

흔한 실수

- $\mathrm{P}(T > 4)$에서 k로 나누는 것을 잊는 실수. 그러면 0.281 대신 0.237을 얻고, (i)에서는 $0.7633^{16} \approx 0.013$을 얻어 올바른 값의 두 배가 넘는다.
- T의 평균을 σ(잘라 내기 전의 평균)로 보아 $\sigma = 3$을 얻는 실수.
- (ii)에서 조건부 확률(conditional probability) 대신 $\mathrm{P}(2 < T \leq 4)$나 $\mathrm{P}(T > 2)$를 구하는 실수.

STEP 너머

평균에 나타난 양 $\phi(1)/\Phi(1)$은 **역 밀스 비**(*inverse Mills ratio*)의 한 예이다. 일반적으로 정규분포 $\mathrm{N}(\mu, \sigma^2)$를 a 아래에서 잘라 낸 분포의 평균은 $\mu + \sigma\,\phi(\alpha)/(1 - \Phi(\alpha))$이며, 여기서 $\alpha = (a - \mu)/\sigma$이다. 절단 분포(truncated distribution)는 자료가 어떤 문턱값보다 클 때만 관측될 수 있는 곳이라면 어디에서나 나타난다. 예를 들어 생존 분석(survival analysis)이나 중도절단 자료(censored data)를 다루는 계량경제학 모형에서 볼 수 있다. 소문항 (ii)는 **생존 함수**(*survival function*) $\mathrm{P}(T > t)$에 관한 것이다. 지수분포(exponential distribution)는 $\mathrm{P}(T > s + t \mid T > s)$가 s에 의존하지 않는 유일한 경우이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 연속분포와 조건부 확률을 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.