풀이

STEP III 2016, 문제 13

확률과 통계 • 합의 첨도

> **STEP III 2016, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/16-s3-q13

핵심 아이디어

첨도(kurtosis)는 평균(mean)으로부터의 편차(deviation)만으로 만들어지므로, 확률변수를 평행이동해도 변하지 않는다. 덕분에 모든 확률변수의 중심을 0에 두고 단순한 적률(moment)로 계산할 수 있다. 정규분포(Normal distribution)의 경우 4차 적률은 부분적분(integration by parts)으로 구하는데, x 하나를 떼어 내어 지수함수와 짝지으면 된다. 평균이 0인 독립(independent) 확률변수들의 합에서는 T^4를 전개하여 기댓값을 취하면 어떤 변수가 1제곱으로 들어 있는 항이 모두 사라진다. 이것이 소문항 (ii)의 전부이며, 그 결과는 곧바로 소문항 (iii)으로 이어진다.

풀이

평행이동. $W = X - a$라 하자. 기댓값의 선형성(linearity of expectation)에 의해 $\mathrm{E}(W) = \mu - a$이므로

$$W - \mathrm{E}(W) = (X - a) - (\mu - a) = X - \mu.$$

W의 평균으로부터의 편차는 X의 평균으로부터의 편차와 정확히 같다. 따라서 W는 같은 분산(variance) $\mathrm{E}((X-\mu)^2) = \sigma^2$과 같은 4차 중심적률(fourth central moment) $\mathrm{E}((X-\mu)^4)$을 가지며, 따라서 첨도도 같다. 특히 $a = \mu$로 두면, 모든 확률변수는 평균이 0인 중심화된 변수 $X - \mu$와 같은 첨도를 가진다.

소문항 (i)

$X \sim \mathrm{N}(0, \sigma^2)$이고 확률밀도함수가

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-x^2/(2\sigma^2)}, \qquad \text{이므로} \qquad f'(x) = -\frac{x}{\sigma^2}\,f(x).$$

라 하자. $\mu = 0$이므로 구해야 할 것은 $\mathrm{E}(X^4) = \int_{-\infty}^{\infty} x^4 f(x)\,\mathrm{d}x$이다.

> **생각해 보기** $\mathrm{e}^{-x^2/(2\sigma^2)}$이 초등함수로 된 역도함수(antiderivative)를 갖지 않는데, $x^4\mathrm{e}^{-x^2/(2\sigma^2)}$은 어떻게 적분할까?

함수 $xf(x)$는 역도함수를 **가지며**, 그것은 $-\sigma^2 f(x)$이다. 그러므로 $x^4 f(x) = x^3 \cdot xf(x)$로 쓰고 $u = x^3$, $\mathrm{d}v = xf(x)\,\mathrm{d}x$, $v = -\sigma^2 f(x)$로 놓아 부분적분하면

$$\int_{-\infty}^{\infty} x^4 f(x)\,\mathrm{d}x = \Big[-\sigma^2 x^3 f(x)\Big]_{-\infty}^{\infty} + 3\sigma^2 \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = 3\sigma^2 \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x.$$

경계항이 사라지는 것은 $x \to \pm\infty$일 때 $x^3\mathrm{e}^{-x^2/(2\sigma^2)} \to 0$이기 때문이다(지수함수는 어떤 거듭제곱이 커지는 것보다도 빨리 감소한다). $u = x$로 같은 과정을 거치면

$$\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = \Big[-\sigma^2 x f(x)\Big]_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \sigma^2,$$

을 얻는다. f는 확률밀도함수(probability density)이므로 적분값이 1이기 때문이다. (이로써 분산이 σ^2임도 확인된다.) 따라서 $\mathrm{E}(X^4) = 3\sigma^2 \cdot \sigma^2 = 3\sigma^4$이고

$$\kappa = \frac{3\sigma^4}{\sigma^4} - 3 = 0.$$

평행이동에 관한 결과에 의해, 평균이 무엇이든 모든 정규분포의 첨도는 0이다.

소문항 (ii)

$T^4 = (Y_1 + \cdots + Y_n)^4$을 다항정리(multinomial theorem)로 전개한다. 각 항은 네 개의 인수 Y_r의 곱이며, 네 인수가 서로 다른 첨자들에 어떻게 나뉘는지에 따라 항은 다섯 가지 유형으로 나뉜다.

$$Y_r^4, \qquad 4\,Y_r^3 Y_s, \qquad 6\,Y_r^2 Y_s^2, \qquad 12\,Y_r^2 Y_s Y_t, \qquad 24\,Y_r Y_s Y_t Y_u,$$

여기서 서로 다른 문자는 서로 다른 첨자를 나타낸다. 계수는 배열의 가짓수를 센 것이다. 예를 들어 $Y_r^2 Y_s^2$은 4개의 괄호 가운데 어느 2개에서 Y_r를 택할지 정하는 데서 나오므로 $\binom{4}{2} = 6$가지이다. 순서 없는 쌍 $\{r, s\}$ 하나는 $Y_r^2 Y_s^2$ 유형에서 한 번씩 나타난다.

이제 기댓값을 취한다. Y들은 독립이므로, 서로 다른 Y들의 거듭제곱의 곱의 기댓값은 각 기댓값의 곱이다. $Y_r^3 Y_s$, $Y_r^2 Y_s Y_t$, $Y_r Y_s Y_t Y_u$ 유형에서는 어떤 변수가 1제곱으로만 나타나고, 그 인수 $\mathrm{E}(Y_s) = 0$ 때문에 기댓값 전체가 0이 된다. 예를 들어

$$\mathrm{E}(Y_r^3 Y_s) = \mathrm{E}(Y_r^3)\,\mathrm{E}(Y_s) = 0.$$

첫째와 셋째 유형만 남으며, $r \neq s$일 때 역시 독립성에 의해 $\mathrm{E}(Y_r^2 Y_s^2) = \mathrm{E}(Y_r^2)\,\mathrm{E}(Y_s^2)$이다. 쌍을 $r < s$로 나열하면

$$\mathrm{E}(T^4) = \sum_{r=1}^{n} \mathrm{E}(Y_r^4) + 6 \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} \mathrm{E}(Y_r^2)\,\mathrm{E}(Y_s^2),$$

이 되어 증명이 끝난다. (이 소문항에서는 독립성과 평균이 0이라는 사실만 썼고, Y들이 동일한 분포(identically distributed)를 따른다는 사실은 쓰지 않았다.)

소문항 (iii)

각 X_r의 평균이 μ, 표준편차(standard deviation)가 σ라 하고 $S = X_1 + \cdots + X_n$이라 하자. $Y_r = X_r - \mu$로 두면 Y_r들은 평균이 0인 독립이고 동일한 분포를 따르는 확률변수이며

$$\mathrm{E}(Y_r^2) = \sigma^2, \qquad \mathrm{E}(Y_r^4) = (\kappa + 3)\sigma^4,$$

이다. 둘째 식은 κ의 정의를 정리하여 얻는다. 그러면 $T = \sum Y_r = S - n\mu$이다. 평행이동에 관한 결과에 의해 S는 T와 같은 첨도를 가지므로, 평균이 0인 T로 계산한다.

독립성에 의해 $\mathrm{Var}(T) = \sum \mathrm{Var}(Y_r) = n\sigma^2$이다. 소문항 (ii)의 이중합은 $\binom{n}{2} = \frac{1}{2}n(n-1)$개의 항으로 이루어지고 각 항은 σ^4이므로

$$\begin{aligned}\mathrm{E}(T^4) &= n(\kappa+3)\sigma^4 + 6 \cdot \tfrac{1}{2}n(n-1)\sigma^4 \\ &= n\sigma^4(\kappa + 3 + 3(n-1)) = n\sigma^4(\kappa + 3n).\end{aligned}$$

따라서 T의 첨도, 곧 S의 첨도는

$$\frac{\mathrm{E}(T^4)}{(\mathrm{Var}\, T)^2} - 3 = \frac{n\sigma^4(\kappa+3n)}{n^2\sigma^4} - 3 = \frac{\kappa}{n} + 3 - 3 = \frac{\kappa}{n},$$

이 되어 증명이 끝난다.

검산

- $Y = \pm 1$이 각각 확률 $\frac{1}{2}$로 나온다고 하자. 그러면 $\mathrm{E}(Y^4) = \sigma^4 = 1$이므로 $\kappa = 1 - 3 = -2$이다. $n = 2$일 때 합은 $-2, 0, 2$를 각각 확률 $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$로 가지므로 $\mathrm{E}(T^4) = 8$이고 $\mathrm{Var}\, T = 2$이다. 따라서 첨도는 $8/4 - 3 = -1 = \kappa/2$로, 예측과 일치한다.
- $n = 1$이면 소문항 (iii)은 아무것도 바뀌지 않는다고 말하며, 당연히 그래야 한다.
- 독립인 정규분포 확률변수들의 합은 정규분포를 따르므로, $\kappa = 0$일 때 소문항 (iii)은 다시 0을 주어야 하며 실제로 그렇다.
- 정확한 분수 계산을 이용하여, 세 값을 갖는 확률변수의 독립인 복사본을 최대 여섯 개까지 더한 합의 첨도가 정확히 κ/n임을 컴퓨터로 확인하였다. 또한 여러 σ 값에 대해 $\int x^4 f(x)\,\mathrm{d}x = 3\sigma^4$임을 수치적으로 확인하였다.

흔한 실수

- $Y_r^2 Y_s^2$의 계수를 잘못 구하는 것. 계수는 순서 없는 쌍마다 $\binom{4}{2} = 6$(순서 있는 쌍마다 3)이지, 4나 12가 아니다.
- 평균이 0이어야 하는 소문항 (ii)를 쓰기 전에 X_r를 중심화하는 것을 잊는 것.
- $(n\sigma^2)^2$ 대신 σ^4로 나누는 것. 합의 분산은 $n\sigma^2$이다.

STEP 너머

여기서 정의한 첨도는 4차 **누적률**(cumulant)을 σ^4으로 나눈 것이다. 누적률은 적률생성함수(moment generating function)의 로그를 전개했을 때의 계수이며, 독립인 확률변수들의 합에서는 단순히 더해진다. 그래서 합의 표준화된 4차 누적률은 $1/n$처럼 줄어들고, 3차 누적률(왜도(skewness))은 $1/\sqrt{n}$처럼 줄어든다. 이것은 중심극한정리(central limit theorem)를 정량적으로 엿보게 해 준다. 합은 점점 더 정규분포처럼 보이게 되는데, 정규분포는 2차보다 높은 누적률이 모두 0이다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)은 이 수준에서 생성함수와 중심극한정리를 전개한다.

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.