풀이

STEP III 2009, 문제 13

확률과 통계 • 중심극한정리

> **STEP III 2009, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/09-s3-q13

핵심 아이디어

모든 것은 하나의 각 θ의 함수이다: $X = \cos\theta$이고 $Y = \sin\theta$이다. 따라서 X의 분포는 누적분포함수(cumulative distribution function)로부터 구할 수 있고(원의 어느 호에서 $\cos\theta \le x$가 되는지 따져 보면 된다), 평균, 분산, 공분산은 삼각함수를 한 바퀴 전체에 걸쳐 평균한 값에 불과하다. 소문항 (ii)에서 $\overline{X}$는 서로 독립이고 같은 분포를 따르는 많은 확률변수의 평균인데, 이는 바로 중심극한정리(central limit theorem)가 적용되는 상황이다.

풀이

소문항 (i)

P가 단위원 위에 있으므로 $X = \cos\theta$, $Y = \sin\theta$이고, 여기서 θ는 $[0, 2\pi)$에서 확률밀도 $\frac{1}{2\pi}$를 가진다.

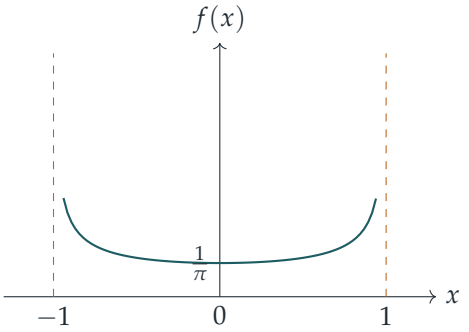
X의 확률밀도함수. 분명히 $-1 \le X \le 1$이다. $-1 < x < 1$인 x를 잡고 $\alpha = \arccos x \in (0, \pi)$라 하자. $[0, 2\pi)$에서 $\cos\theta > x$인 것은 정확히 $\theta < \alpha$ 또는 $\theta > 2\pi - \alpha$일 때이며, 이는 전체 길이가 2α인 호이다. 따라서

$$F(x) = \mathrm{P}(X \le x) = 1 - \frac{2\alpha}{2\pi} = 1 - \frac{\arccos x}{\pi}.$$

$\dfrac{\mathrm{d}}{\mathrm{d}x}\arccos x = -\dfrac{1}{\sqrt{1-x^2}}$을 이용하여 미분하면

$$f(x) = \frac{1}{\pi\sqrt{1-x^2}} \qquad (-1 < x < 1),$$

이고, 그 밖의 경우에는 $f(x) = 0$이다. (간단한 검산: $\int_{-1}^{1} f(x)\,\mathrm{d}x = \frac{1}{\pi}[\arcsin x]_{-1}^{1} = 1$.)



확률밀도는 $x = 0$에서 가장 작고 $x = \pm 1$에서 무한히 커진다. 점 P는 수평 지름의 양 끝 근처에서 수평 방향으로 가장 느리게 움직이므로, X는 대부분의 시간을 그곳에서 보낸다.

분산. 대칭성에 의해(f는 우함수이다) $\mathrm{E}(X) = 0$이다. θ를 직접 사용하여 계산하면

$$\mathrm{E}(X^2) = \frac{1}{2\pi}\int_0^{2\pi} \cos^2\theta \,\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} \tfrac{1}{2}(1+\cos 2\theta)\,\mathrm{d}\theta = \frac{1}{2},$$

이므로

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \tfrac{1}{2}.$$

같은 논법으로(또는 $\theta - \frac{\pi}{2}$도 한 바퀴 전체에서 균등분포를 따르므로) $\mathrm{E}(Y) = 0$이고 $\mathrm{Var}(Y) = \frac{1}{2}$이다.

상관관계가 없음.

$$\mathrm{E}(XY) = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta \sin\theta \,\mathrm{d}\theta = \frac{1}{4\pi}\int_0^{2\pi} \sin 2\theta \,\mathrm{d}\theta = 0,$$

이므로 $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 0 - 0 = 0$이다. 따라서 X와 Y는 서로 상관관계가 없다(uncorrelated).

독립이 아님. $X^2 + Y^2 = 1$이므로 두 확률변수는 결코 독립(independent)이 아니다. X가 정해지면 Y는 $\pm\sqrt{1-X^2}$ 중 하나로 고정된다. 구체적인 사건으로 보면, $\mathrm{P}(X > 0.8) > 0$이고 $\mathrm{P}(Y > 0.8) > 0$이지만 $\mathrm{P}(X > 0.8$이고$Y > 0.8) = 0$이다. 그런 경우에는 $X^2 + Y^2 > 1.28$이 되기 때문이다. 독립이라면 이 결합확률이 (양수인) 두 확률의 곱과 같아야 한다. 상관계수가 0이라는 것은 **선형** 관계가 없다는 것만을 말해 준다. 여기서 두 변수의 관계는 원이며, 원은 충분히 대칭적이어서 선형 경향이 전혀 없다.

소문항 (ii)

각 쌍 (X_i, Y_i)는 소문항 (i)의 분포를 따르고, 점들이 독립적으로 선택되므로 첨자가 다른 쌍끼리는 서로 독립이다.

$\overline{X}$와 $\overline{Y}$는 상관관계가 없다. 공분산은 쌍선형(bilinear)이므로

$$\mathrm{Cov}(\overline{X},\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i,Y_j).$$

$i \neq j$이면 X_i는 P_i에만, Y_j는 P_j에만 의존하고 이 둘은 서로 독립이므로 $\mathrm{Cov}(X_i,Y_j)=0$이다. $i=j$이면 소문항 (i)에 의해 $\mathrm{Cov}(X_i,Y_i)=0$이다. 모든 항이 0이므로 $\mathrm{Cov}(\overline{X},\overline{Y})=0$이다. 즉 $\overline{X}$와 $\overline{Y}$는 상관관계가 없다.

근사 확률. X_i들은 서로 독립이고 평균이 0, 분산이 $\frac{1}{2}$인 같은 분포를 따른다. 따라서

$$\mathrm{E}(\overline{X})=0, \qquad \mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot\frac{n}{2} = \frac{1}{2n},$$

이다. 여기서 분산을 더할 수 있는 것은 X_i들이 서로 독립이기 때문이다.

> **생각해 보기** X는 정규분포와 전혀 닮지 않았다(확률밀도함수를 보라). 그런데 왜 정규근사를 쓸 수 있을까?

중심극한정리는 각 항이 서로 독립이고 같은 분포를 따르며 분산이 유한하기만 하면 되고, 그 분포의 모양은 상관없다. 따라서 n이 크면 $\overline{X}$는 근사적으로 $\mathrm{N}(0,\frac{1}{2n})$을 따르고, 표준편차는 $\sigma = \dfrac{1}{\sqrt{2n}}$이다. 이제

$$\sqrt{\frac{2}{n}} = \frac{2}{\sqrt{2n}} = 2\sigma,$$

이므로 $Z\sim\mathrm{N}(0,1)$이라 하면

$$\begin{aligned}\mathrm{P}\left(|\overline{X}| \le \sqrt{\tfrac{2}{n}}\right) &= \mathrm{P}\left(\frac{|\overline{X}|}{\sigma} \le 2\right) \approx \mathrm{P}(|Z|\le 2)\\ &= 2\Phi(2)-1 = 2(0.9772)-1 \approx 0.954 \approx 0.95,\end{aligned}$$

이고, 이것이 구하려던 결과이다.

검산

- 분산이 $\frac{1}{2}$인 것은 타당하다. $X^2+Y^2=1$이고 대칭성에 의해 $\mathrm{E}(X^2)=\mathrm{E}(Y^2)$이므로 각각은 $\frac{1}{2}$이어야 한다. 적분할 필요조차 없다!
- 4×10^5개의 무작위 각으로 모의실험한 결과, 여러 x값에서 $F(x)=1-\frac{1}{\pi}\arccos x$와 소수점 아래 셋째 자리 정도까지 일치했고, $\mathrm{Var}(X)\approx 0.5007$, $\mathrm{Cov}(X,Y)\approx 0.001$을 얻었다.
- $n=50$으로 40 000번 반복했을 때 모의실험으로 얻은 $\mathrm{Var}(\overline{X})$는 0.00997(이론값 0.01)이었고, $|\overline{X}|\le\sqrt{2/50}=0.2$인 비율은 0.954로 $\mathrm{P}(|Z|\le 2)=0.9545$와 비교된다.

흔한 실수

- $1 - \frac{1}{\pi}\arccos x$를 미분할 때 부호를 잘못 처리하여 음수인 "확률밀도"를 얻는 것. 확률밀도는 결코 음수가 될 수 없으니 부호를 확인하라.
- 공분산이 0이라는 것으로부터 독립이라고 결론 내리는 것. 독립이면 공분산이 0이지만, 그 역은 성립하지 않는다.
- $\mathrm{Var}(\overline{X}) = \frac{1}{2}$ 또는 $\frac{1}{2} \cdot \frac{1}{\sqrt{n}}$을 쓰는 것. 독립인 n개의 복사본의 평균의 분산은 하나의 분산을 n으로 나눈 것이다. $1/\sqrt{n}$에 비례하는 것은 **표준편차**이다.

STEP 너머

확률밀도 $1/(\pi\sqrt{1-x^2})$은 **아크사인 분포**(arcsine distribution)이며, 놀라울 만큼 자주 등장한다. 예를 들어 무작위 행보(random walk)가 양수 쪽에 머무는 시간의 비율의 극한 분포가 바로 이것이다. 이 문제의 나머지 부분은 상관관계가 없음과 독립 사이의 차이, 그리고 중심극한정리에 관한 것인데, 중심극한정리는 대학 1학년 확률론 강의에서 적률생성함수(moment generating function)나 특성함수(characteristic function)를 이용하여 증명하게 된다. 사실 $(\overline{X}, \overline{Y})$는 근사적으로 이변량 정규분포(bivariate normal distribution)를 따르며, 결합 정규분포를 따르는 확률변수들에 대해서는 상관계수가 0이면 독립이 **성립한다**. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 이 모든 내용을 잘 다루고 있다.

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.