12.22 평균이 없는 분포

학습지

문제

STEP III 1995, 12번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/95-s3-q12

힌트

먼저 혼자 힘으로 문제를 풀어 보자. 막히면 힌트를 **하나씩** 읽고, 힌트를 하나 읽을 때마다 다시 문제로 돌아가자. **잠깐, 직접 해 보기**라고 표시된 힌트가 핵심 단계이다. 각 힌트 뒤에는 풀이를 적을 공간이 있다.

> ? **생각해 보기** (U, V)가 원점으로부터의 거리에만 의존하는 밀도를 가진 무작위 점이라면, 원점에서 그 점까지 이은 선분이 한 축과 이루는 각 Θ에 대해 무엇을 말할 수 있는가?

1단계 결합 밀도

다루는 내용: 결합 밀도와 회전 대칭

> ? **생각해 보기** X와 Y는 독립(independent)이므로 두 밀도를 곱한다. 그 결과는 무엇에만 의존하는가?

2단계 대칭성 이용하기

다루는 내용: $\mathrm{P}(U/V < k)$에 대한 항등식

> **잠깐, 직접 해 보기** $\frac{U}{V} < k$는 $V > 0$일 때에도, $V < 0$일 때에도 일어날 수 있다. 각 경우를 영역으로 나타내어 보자. 회전 $(u, v) \mapsto (-u, -v)$는 이 영역들을 어떻게 바꾸는가?

3단계 각

다루는 내용: U/V의 밀도

생각해 보기 영역 $\{V > 0, U < kV\}$를 그려 보자. 이것은 원점을 꼭짓점으로 하는 부채꼴 모양의 영역이다. 그 각의 크기는 얼마인가? 밀도가 회전 대칭(radially symmetric)이면 이러한 영역의 확률은 그 각의 크기를 2π로 나눈 값이다. 그다음 미분한다.

숨은 수학

평균이 없다

코시 분포의 평균은 $\int_{-\infty}^{\infty} \frac{k}{\pi(1+k^2)}\,\mathrm{d}k$여야 하겠지만, $\int_0^K \frac{k}{1+k^2}\,\mathrm{d}k = \frac{1}{2}\ln(1+K^2) \to \infty$이다. 즉 이 적분은 수렴(converge)하지 않고, 평균은 존재하지 않는다. (대칭성 때문에 답이 0일 것 같지만, 양의 부분과 음의 부분이 모두 무한대이고 $\infty - \infty$는 값을 갖지 않는다.) 분산(variance)도 무한대이다.

더 나쁜 것은, 독립인 코시 확률변수 n개의 평균이 다시 정확히 같은 분포의 코시 분포를 따른다는 점이다. (이를 확인하는 한 가지 방법은 특성함수(characteristic function) $\mathrm{E}(e^{itX}) = e^{-|t|}$를 이용하는 것이다.) 따라서 평균을 내는 것은 전혀 도움이 되지 않는다. 백만 번 측정한 값의 평균이 한 번 측정한 값보다 조금도 더 믿을 만하지 않다. 그림 12.22.1는 무슨 일이 일어나는지 보여 준다. 정규분포 표본의 누적 평균(running average)은 안정되지만, 코시 분포 표본의 누적 평균은 이따금 나타나는 거대한 값에 흔들리며 끝없이 요동친다.

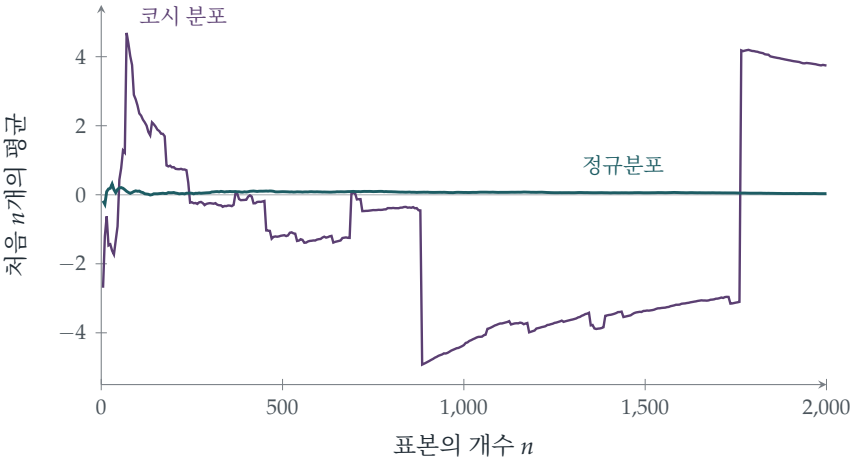

그림 12.22.1 독립인 표본 2000개의 누적 평균. 표준정규분포 표본(청록색)의 경우, 큰 수의 법칙이 예측하는 대로 평균이 0으로 안정된다. 코시 분포 표본(자주색)의 경우 평균은 결코 안정되지 않는다. 이따금 엄청나게 큰 값 하나가 평균을 새로운 곳으로 내던진다.

코시 확률변수는 어디에서 나오는가

균등하게 무작위인 방향을 향하도록 손전등을 벽에 비추면, 벽에 생기는 빛점의 위치는 코시 분포를 따른다. 물리학에서는 같은 밀도를 **로렌츠 분포**(*Lorentzian*)라고 부르며, 스펙트럼선과 공명 봉우리의 모양을 기술한다(Section 11.1와 비교해 보라). 금융에서는 정규분포 모형이 크게 과소평가하는 이따금의 폭락을 모형화하기 위해 코시 분포처럼 꼬리가 두꺼운(heavy-tailed) 분포를 사용한다. 통계학에 주는 교훈은 심각하다. 평균과 분산에 기반한 방법은 꼬리가 두꺼운 자료에서는 실패하며, 중앙값(median)을 사용하는 것과 같은 로버스트한(robust) 방법이 필요하다(코시 분포에서도 중앙값은 완벽하게 잘 정의되고 잘 작동한다).

푸아송, 코시, 그리고 논쟁

이 분포는 1824년 시메옹 드니 푸아송(Siméon Denis Poisson)이 연구했는데, 그는 이 분포가 큰 수의 법칙을 위반한다는 것을 알아차렸다. 오귀스탱루이 코시(Augustin-Louis Cauchy)는 1853년 최소제곱법(method of least squares)을 둘러싸고 Irénée-Jules Bienaymé와 공개 논쟁을 벌이던 중 이 분포를 만났다. 코시는 오차의 꼬리가 두꺼울 때 최소제곱법이 형편없는 방법일 수 있음을 보이는 데 이 분포를 사용했다. 이름은 코시에게 붙었는데, 이는 Stephen Stigler가 "스티글러의 명명 법칙(Stigler's law of eponymy)"이라고 부른 것의 한 예이다. 어떤 과학적 발견도 최초 발견자의 이름을 따서 불리지 않는다는 법칙이다.

더 도전하기

1 **손전등.** 길고 곧은 벽에서 거리 1만큼 떨어진 손전등이 벽의 수선과 각 Θ를 이루는 방향을 가리킨다. 여기서 Θ는 $(-\frac{\pi}{2}, \frac{\pi}{2})$에서 균등분포를 따른다. 빛점의 위치 $\tan\Theta$가 코시 분포를 따름을 보여라.

2 **역수.** C가 코시 분포를 따르면 $\frac{1}{C}$도 코시 분포를 따름을 보여라. (**힌트:** $\frac{U}{V}$와 $\frac{V}{U}$.)

3 **중앙값은 잘 작동한다.** 코시 분포의 중앙값이 0이고 사분위수(quartile)가 ± 1임을 보여라.

4 **반지름과 각.** X, Y가 독립인 표준정규 확률변수일 때, 극좌표(polar coordinates)로 바꾸어 $R^2 = X^2 + Y^2$가 평균이 2인 지수분포(exponential distribution)를 따르고 각과 독립임을 보여라. (이것이 정규분포 난수를 생성하는 Box–Muller 방법의 기초이다.)

더 읽을거리

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press). 확률변수의 변환을 다룬다.
- **Nassim Nicholas Taleb, *The Black Swan*** (Penguin). 일반 독자를 위해 두꺼운 꼬리와 그것이 왜 중요한지를 설명한다.
- **Stephen Stigler, *The History of Statistics*** (Harvard University Press). 코시와 Bienaymé의 논쟁을 담고 있다.

안내에 따른 전체 풀이는 제 12.22절에 있다:

https://beyondstep.keithmcnulty.org/ko/ch12/cauchy.html

매주 STEP 문제 하나를 끝까지 풀어 본다. The STEP Up은 이메일로 무료로 받아 볼 수 있다: keithmcnulty.substack.com