풀이

STEP II 2023, 문제 11

확률과 통계 • 혼합분포

> **STEP II 2023, 11번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 OCR에서 내려받을 수 있는 2023년 STEP 시험지 묶음(zip 파일) 가운데 STEP II 시험지에 있다:
>
> www.ocr.org.uk/Images/707102-2023-past-papers.zip

이 문제는 *Beyond STEP*의 13.5 절 **곤란한 질문 던지기**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch13/awkward-questions.html

핵심 아이디어

Y는 **혼합**(*mixture*) 분포를 따른다. 먼저 분포 하나를 무작위로 고른 다음, 그 분포에서 값을 뽑는다. Y의 평균(mean)은 두 평균의 가중평균이지만, 분산(variance)에는 항이 하나 더 붙는다. 두 평균 사이의 차이가 그 자체로 변동을 더하기 때문이다. 소문항 (ii)와 (iii)은 모두 주사위를 쓰기 전에 동전을 던지지만, 그 결과는 완전히 다르다. (ii)에서는 주사위를 던질 때마다 동전을 새로 던지므로 혼합의 효과가 평균화되어 사라진다. (iii)에서는 동전 한 번이 모든 것을 결정하고, 그 추가 항이 결국 전체를 지배한다.

풀이

소문항 (i)

어느 확률변수를 쓸지 고르는 일은 X_1, X_2의 값과 독립적으로 이루어지므로

$$\mathrm{P}(Y = x_i) = p\,a_i + q\,b_i \qquad (i = 1, \ldots, n).$$

따라서

$$\mathrm{E}(Y) = \sum_i x_i(pa_i + qb_i) = p\sum_i x_i a_i + q\sum_i x_i b_i = p\mu_1 + q\mu_2.$$

같은 방법으로 $\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2+\mu_1^2) + q(\sigma_2^2+\mu_2^2)$이다. 그러므로

$$\begin{aligned}\mathrm{Var}(Y) &= p\sigma_1^2 + q\sigma_2^2 + p\mu_1^2 + q\mu_2^2 - (p\mu_1+q\mu_2)^2 \\ &= p\sigma_1^2 + q\sigma_2^2 + p(1-p)\mu_1^2 + q(1-q)\mu_2^2 - 2pq\mu_1\mu_2 \\ &= p\sigma_1^2 + q\sigma_2^2 + pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2 \\ &= p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1-\mu_2)^2,\end{aligned}$$

이고, 여기서 $1-p=q$와 $1-q=p$를 이용하였다. 이것이 구하려던 결과이다.

소문항 (ii)

동전이 앞면이 나오는 사건을 H, 주사위가 6이 나오는 사건을 S라 하자. 동전과 주사위는 독립이므로

$$\mathrm{P}(B=1) = \mathrm{P}(H)\mathrm{P}(S) + \mathrm{P}(H^c)\mathrm{P}(S^c) = \tfrac{1}{2}\cdot\tfrac{1}{6} + \tfrac{1}{2}\cdot\tfrac{5}{6} = \tfrac{1}{2}.$$

따라서 각 B는 0 또는 1을 각각 확률 $\frac{1}{2}$로 가진다. Z_1은 서로 독립인 n개의 B 가운데 1의 개수이며, 각각은 같은 확률 $\frac{1}{2}$로 성공한다. 이것은 정확히 이항분포(binomial distribution)의 상황이므로 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$이고, 평균은 $\frac{n}{2}$, 분산은 $\frac{n}{4}$이다.

n이 클 때 Z_1을 $\mathrm{N}(\frac{n}{2}, \frac{n}{4})$로 근사하며, 이 분포의 표준편차(standard deviation)는 $\frac{\sqrt{n}}{2}$이다. 평균의 10% 이내에 있다는 것은 $|Z_1 - \frac{n}{2}| \le \frac{n}{20}$을 뜻한다. $Z \sim \mathrm{N}(0,1)$로 두고 표준화(standardising)하면

$$\mathrm{P}\big(|Z_1 - \tfrac{n}{2}| \le \tfrac{n}{20}\big) \approx \mathrm{P}\Big(|Z| \le \frac{n/20}{\sqrt{n}/2}\Big) = \mathrm{P}\Big(|Z| \le \frac{\sqrt{n}}{10}\Big) = 2\Phi\Big(\frac{\sqrt{n}}{10}\Big) - 1.$$

$n \to \infty$일 때 $\sqrt{n}/10 \to \infty$이므로 $\Phi(\sqrt{n}/10) \to 1$이고, 확률은 $2-1=1$로 수렴한다. (연속성 보정(continuity correction)을 하더라도 경계가 표준편차의 $\frac{1}{\sqrt{n}}$배만큼만 바뀌므로 극한에는 영향이 없다.)

소문항 (iii)

n개의 주사위 가운데 6의 개수를 X_1, 6이 아닌 것의 개수를 X_2라 하자. 그러면 $X_1 \sim \mathrm{B}(n, \frac{1}{6})$, $X_2 \sim \mathrm{B}(n, \frac{5}{6})$이다. 이제 Z_2는 확률 $p=\frac{1}{2}$로 X_1의 값을, 확률 $q=\frac{1}{2}$로 X_2의 값을 가지며, 동전은 주사위와 독립이다. 이것은 정확히 소문항 (i)의 상황이고,

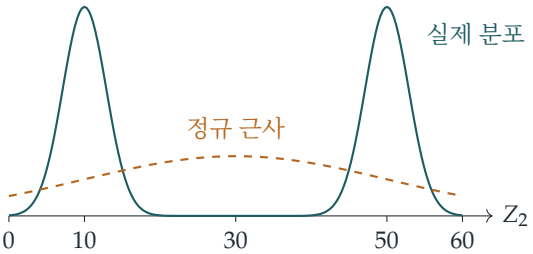
$$\mu_1 = \tfrac{n}{6}, \quad \mu_2 = \tfrac{5n}{6}, \quad \sigma_1^2 = \sigma_2^2 = n\cdot\tfrac{1}{6}\cdot\tfrac{5}{6} = \tfrac{5n}{36}.$$

이다. 그러므로

$$\begin{aligned}\mathrm{E}(Z_2) &= \tfrac{1}{2}\cdot\tfrac{n}{6} + \tfrac{1}{2}\cdot\tfrac{5n}{6} = \tfrac{n}{2}, \\ \mathrm{Var}(Z_2) &= \tfrac{1}{2}\cdot\tfrac{5n}{36} + \tfrac{1}{2}\cdot\tfrac{5n}{36} + \tfrac{1}{4}\Big(\tfrac{4n}{6}\Big)^2 = \frac{5n}{36} + \frac{n^2}{9}.\end{aligned}$$

? 생각해 보기 n이 클 때 Z_2의 분포는 실제로 어떤 모양일까?

확률 $\frac{1}{2}$로 Z_2는 $\frac{n}{6}$ 근처에 있고, 확률 $\frac{1}{2}$로 $\frac{5n}{6}$ 근처에 있다. 두 봉우리의 표준편차는 각각 $\frac{\sqrt{5n}}{6}$으로, 두 봉우리 사이의 거리 $\frac{2n}{3}$에 비하면 아주 작다. 따라서 이 분포는 **쌍봉**(*bimodal*) 분포이다. 좁은 봉우리가 두 개 있고, 평균 $\frac{n}{2}$ 근처에는 확률이 거의 없다. 정규분포는 평균에서 봉우리가 하나뿐이므로 여기서는 정반대로 틀린다. Z_2가 거의 떨어지지 않는 곳에 가장 많은 확률을 주기 때문이다. 문제가 하나 더 있다. 표준편차가 약 $\frac{n}{3}$이므로 $\mathrm{N}(\frac{n}{2}, \frac{n^2}{9})$는 음수 값에 약 $\mathrm{P}(Z < -1.5) \approx 0.067$의 확률을 주고, n보다 큰 값에도 같은 확률을 준다. 하지만 항상 $0 \le Z_2 \le n$이다. 아래 그림은 $n = 60$일 때의 두 그래프를 보여 준다.



이제 마지막 부분이다. Z_2가 평균의 10% 이내에 있는 것은 $0.45n \le Z_2 \le 0.55n$일 때이다. 동전의 결과에 대해 조건부(conditioning)로 나누면

$$\mathrm{P}(0.45n \le Z_2 \le 0.55n) = \tfrac{1}{2}\,\mathrm{P}(0.45n \le X_1 \le 0.55n) + \tfrac{1}{2}\,\mathrm{P}(0.45n \le X_2 \le 0.55n).$$

첫째 항에 대하여 $\mathrm{P}(0.45n \le X_1 \le 0.55n) \le \mathrm{P}(X_1 \ge 0.45n)$이고, $0.45n - \frac{n}{6} = \frac{17n}{60}$이다. 표준편차가 $\frac{\sqrt{5n}}{6}$인 정규근사 $X_1 \approx \mathrm{N}(\frac{n}{6}, \frac{5n}{36})$를 쓰면

$$\mathrm{P}(X_1 \ge 0.45n) \approx \mathrm{P}\Big(Z \ge \frac{17n/60}{\sqrt{5n}/6}\Big) = 1 - \Phi\Big(\frac{17\sqrt{n}}{10\sqrt{5}}\Big) \longrightarrow 1 - 1 = 0.$$

둘째 항은 첫째 항의 거울상이다. $X_2 = n - X_1$이므로 $0.45n \le X_2 \le 0.55n$은 $0.45n \le X_1 \le 0.55n$과 정확히 같은 조건이고, 둘째 항은 첫째 항과 같다. 따라서 Z_2가 평균의 10% 이내에 있을 확률은 많아야 약 $1 - \Phi(\frac{17\sqrt{n}}{10\sqrt{5}})$이며, 이는 $n \to \infty$일 때 0으로 수렴한다.

이 단계는 사실 근사에 의존하지 않는다. 체비쇼프 부등식(Chebyshev's inequality)을 쓰면 정확하게

$$\mathrm{P}\big(|X_1 - \tfrac{n}{6}| \ge \tfrac{17n}{60}\big) \le \frac{5n/36}{(17n/60)^2} = \frac{500}{289\,n} \longrightarrow 0.$$

을 얻는다.

검산

- (i)에서 $\mu_1 = \mu_2$이면 추가 항이 사라진다. 또 $\sigma_1 = \sigma_2 = 0$이면 Y는 μ_1 또는 μ_2의 값만 가지며, 분산은 크기를 조정한 베르누이(Bernoulli) 확률변수처럼 $pq(\mu_1 - \mu_2)^2$이다.

- (ii)에서 $n = 400$일 때 정확한 이항확률은 0.9598이고, 정규근사 값은 $2\Phi(2) - 1 = 0.9545$이다. $n = 1000$일 때 두 값은 0.9986과 0.9984이다.
- (iii)에서 $n = 30$으로 Z_2를 컴퓨터 시뮬레이션한 결과 평균 15.00, 분산 104.2를 얻었으며, 공식의 값은 15와 104.17이다. $n = 100$일 때 평균의 10% 이내에 있을 정확한 확률은 약 3×10^{-11}이다.

흔한 실수

- 두 평균 가운데 하나를 고르는 일 자체가 변동의 원천이라는 점을 잊고 $\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2$라고 답하는 것.
- (iii)의 Z_2를 (ii)의 Z_1처럼 다루는 것. (ii)에서는 주사위마다 동전을 다시 던지지만, (iii)에서는 동전 한 번이 n개의 주사위 전체를 좌우하므로 결과들이 독립이 아니다.
- 근사가 잘 맞지 않는 이유를 "분산이 크다"로만 설명하는 것. 진짜 문제는 모양이다. 봉우리가 두 개로 떨어져 있고, 평균 근처에는 확률이 아주 적다.

STEP 너머

소문항 (i)은 **전체 분산의 법칙**(*law of total variance*) $\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid C)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid C)\big)$의 가장 간단한 경우이며, 여기서 C는 동전이다. 소문항 (ii)는 큰 수의 약한 법칙(weak law of large numbers)으로, 대학의 첫 확률론 강의에서 체비쇼프 부등식을 이용해 엄밀하게 증명한다. 소문항 (iii)은 이 법칙에 왜 독립성이 필요한지 보여 준다. 공유된 동전 하나만으로도 법칙이 무너진다. Z_2와 같은 혼합분포는 통계적 모형화에서 늘 쓰이는 도구이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 세 가지 아이디어를 모두 많은 예제와 함께 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.