13.5 곤란한 질문 던지기

학습지

문제

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

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 Y는 확률 p로 X_1이고 확률 q로 X_2이다. $\mathrm{E}(Y)$는 무엇인가? $\mathrm{E}(Y^2)$는 무엇인가?

1단계 혼합 분포의 분산

다루는 내용: 소문항 (i)

생각해 보기 $\mathrm{E}(Y)$와 $\mathrm{E}(Y^2)$를 구하려면 어느 확률변수가 선택되는지를 조건으로 하여 계산한다(조건부 기댓값, conditioning). $\mathrm{E}(X_i^2) = \sigma_i^2 + \mu_i^2$로 쓴다.

잠깐, 직접 해 보기 $(\mathrm{E}(Y))^2$을 빼고, $p - p^2 = pq$와 $q - q^2 = pq$를 이용하여 μ가 들어 있는 항을 정리한다.

2단계 독립적인 무작위화

다루는 내용: 소문항 (ii)

생각해 보기 $\mathrm{P}(B = 1)$은 얼마인가? n개의 응답은 서로 독립(independent)인가? 그렇다면 Z_1의 분포는 무엇인가?

생각해 보기 "평균의 10% 이내"를 표준화된 단위로 나타내어 보자. $n \to \infty$일 때 어떻게 되는가?

3단계 모두를 위한 동전 하나

다루는 내용: 소문항 (iii)

잠깐, 직접 해 보기 이제 동전 하나가 n개의 주사위 모두에 대한 결정을 내린다. 앞면이 나왔을 때 Z_2의 분포는 무엇인가? 뒷면이 나왔을 때는? 평균과 분산은 소문항 (i)의 결과를 이용한다.

생각해 보기 Z_2의 확률은 대부분 어디에 몰려 있는가? $\frac{n}{2}$ 근처에도 많이 있는가? 각 봉우리가 $\frac{n}{2}$에서 표준편차의 몇 배만큼 떨어져 있는지 어림해 보자.

숨은 수학

확률화 응답

Warner의 방식에서 각 응답자는 남몰래 선택하여, 확률 p로 "당신은 집단 A에 속합니까?"에 답하고, 확률 $1-p$로 "당신은 집단 A에 속하지 **않습니까**?"에 답한다. 비율 π만큼의 사람이 A에 속한다면 "예"라고 답할 확률은

$$\lambda = p\pi + (1-p)(1-\pi),$$

이므로, 관측된 "예"의 비율 $\hat{\lambda}$로부터 $\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$로 추정할 수 있다(단, $p \neq \frac{1}{2}$). 문제의 소문항 (ii)가 바로 이것이며, 각자의 주사위가 결정을 내린다. 프라이버시의 대가는 정밀도이다. $2p-1$로 나누면 분산이 커지므로, 확률화된 조사는 직접 조사보다 더 많은 응답자가 필요하다. 하지만 각 사람이 독립적으로 확률화하므로 큰수의 법칙(law of large numbers)이 작동한다. 소문항 (iii)은 확률화를 공유하면 무엇이 잘못되는지 보여 준다. 단 하나의 동전에 대한 불확실성은 결코 평균화되어 사라지지 않는다.

차분 프라이버시

확률화 응답은 **차분 프라이버시**(*differential privacy*)의 조상이다. 차분 프라이버시는 2006년 Cynthia Dwork와 동료들이 도입한 프라이버시의 수학적 정의로, 어떤 한 사람의 데이터를 추가하거나 제거해도 모든 출력의 확률이 기껏해야 작은 배수만큼만 변한다면 그 절차는 프라이버시를 보장한다고 한다. 기술 기업들은 이를 이용해 사용 통계를 수집하며, 미국 인구조사국(United States Census Bureau)은 2020년 인구조사에서 이를 사용하여 공개되는 수치에 신중하게 보정한 무작위 잡음을 더함으로써 어떤 개인의 응답도 재구성할 수 없게 했다.

전분산 법칙

소문항 (i)의 항등식은 **전분산 법칙**(*law of total variance*)의 특수한 경우이다. 즉 $\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$이며, 여기서 C는 숨겨진 선택이다. 이 법칙은 **분산분석**(*analysis of variance*)(ANOVA)의 바탕이 된다. 분산분석은 1920년대에 피셔(Ronald Fisher)가 로섬스테드(Rothamsted)에서의 농업 실험을 위해 도입한 것으로, 작물 수확량의 변동을 비료나 품종으로 설명되는 부분과 설명되지 않고 남는 부분으로 나눈다.

더 도전하기

1 **쌍둥이.** STEP II 2005, 12번 문제 (step.maths.org/questions/05-s2-q12)를 풀어 보라. 이것은 겉모습만 바꾼 확률화 응답 퍼즐이다. 어떤 답이 동전에 대해 아무것도 알려 주지 않는 것은 언제인가?

2 **π 추정하기.** $p = \frac{3}{4}$인 Warner 조사에서 400명이 응답했고 160명이 "예"라고 답했다. 집단 A의 비율 π를 추정하고, 그 추정값의 근사적인 표준오차(standard error)를 구하라.

3 **전분산.** 여러 값을 갖는 이산 숨은 변수 C에 대해 전분산 법칙을 증명하라.

4 **공유된 동전.** 소문항 (iii)에서 $Z_2 > \frac{n}{2}$일 확률은 얼마인가? Z_2가 $\frac{n}{2}$에 가까운 경우는 거의 없는데도, n이 클 때 이 확률이 $\frac{1}{2}$에 가까운 이유를 설명하라.

더 읽을거리

- **Stanley Warner, "Randomized response: a survey technique for eliminating evasive answer bias"**, *Journal of the American Statistical Association* 60 (1965).
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*** (온라인 무료 공개). 현대 이론을 다루며, 앞부분의 장들은 쉽게 읽을 수 있다.
- **David Spiegelhalter, *The Art of Statistics*** (Pelican). 조사와 그 함정을 다룬다.

안내에 따른 전체 풀이는 제 13.5절에 있다:

https://beyondstep.keithmcnulty.org/ko/ch13/awkward-questions.html

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