풀이

STEP I 2012, 문제 13

확률과 통계 • 지시변수를 통한 기댓값

> **STEP I 2012, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/12-s1-q13

핵심 아이디어

먼저 조건부 표본공간(conditional sample space)을 정확히 파악한다. 어떤 숫자가 제외되는지 알고 나면, 그 수는 각 자리가 $\{1,2,3,4,5\}$에서 선택된 다섯 자리 숫자열일 뿐이고, 이러한 숫자열은 모두 같은 확률로 나타난다. 그다음 서로 다른 숫자의 개수의 분포 전체를 구하는 대신, 그 개수를 가능한 숫자마다 하나씩 대응하는 지시확률변수(indicator variable)들의 합으로 나타낸다. 기댓값의 선형성(linearity of expectation)을 이용하면 문제는 하나의 확률을 구하는 것으로 바뀌고, 이 확률은 여사건(complement)을 생각하면 쉽게 구할 수 있다.

풀이

조건부 표본공간

10000부터 99999까지의 모든 정수는 $d_1 \neq 0$인 다섯 자리 숫자열 $d_1d_2d_3d_4d_5$이고, 이러한 숫자열은 각각 정확히 한 번씩 나타난다. 조건은 모든 자리의 숫자가 $\{1,2,3,4,5\}$에 속한다는 것이다. 0은 어차피 제외되므로 $d_1 \neq 0$이라는 제한은 아무것도 더하지 않는다. 따라서 조건을 만족하는 정수는 $\{1,2,3,4,5\}$의 숫자로 이루어진

$$5^5 = 3125$$

개의 다섯 자리 숫자열과 정확히 일대일로 대응한다.

원래의 선택은 균등(uniform)했다. 어떤 사건으로 조건을 주어도 그 사건 안의 결과들은 여전히 같은 확률을 가지므로, 조건이 주어졌을 때 이 3125개의 숫자열은 각각 확률 $\frac{1}{3125}$를 가진다. 달리 말하면, 다섯 자리의 숫자는 서로 독립(independent)이고 각각 $\{1,2,3,4,5\}$에서 균등분포를 따른다. 즉 3125개의 결과 각각의 확률은 $\frac{1}{5}$ 다섯 개의 곱이다.

지시확률변수

선택된 수에 나타나는 서로 다른 숫자의 개수를 N이라 하자. $j = 1, 2, 3, 4, 5$에 대하여

$$I_j = \begin{cases} 1 & \text{숫자}j\text{가 적어도 한 번 나타나면,} \\ 0 & \text{그 외의 경우.} \end{cases}$$

로 정의한다. 나타나는 각 숫자는 합에 정확히 1씩 기여하므로

$$N = I_1 + I_2 + I_3 + I_4 + I_5.$$

이다.

생각해 보기 I_j들은 독립이 아니다. 그래도 기댓값을 더할 수 있을까?

그렇다. 기댓값의 선형성 $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$는 독립이든 아니든 모든 확률변수에 대하여 성립한다. 따라서

$$\mathrm{E}(N) = \sum_{j=1}^{5} \mathrm{E}(I_j) = \sum_{j=1}^{5} \mathrm{P}(\text{숫자}j\text{가 나타남}).$$

이다.

숫자 j가 나타나지 않는 것은 서로 독립인 다섯 자리의 숫자가 모두 나머지 네 값 중 하나일 때이고, 그 확률은 각 자리마다 $\frac{4}{5}$이다. 그러므로

$$\mathrm{P}(\text{숫자}j\text{가 나타남}) = 1 - \left(\tfrac{4}{5}\right)^5 = 1 - \frac{1024}{3125} = \frac{2101}{3125}.$$

이다. 대칭성에 의해 이 값은 다섯 개의 j 값에 대하여 모두 같으므로

$$\mathrm{E}(N) = 5 \cdot \frac{2101}{3125} = \frac{2101}{625} = 3.3616,$$

이고, 이것이 증명하려던 결과이다.

다른 방법: 분포 전체 구하기

직접 세는 것을 선호한다면, 정확히 k개의 서로 다른 숫자를 사용하는 숫자열의 개수는 $\binom{5}{k}$(숫자를 고르는 방법의 수)에, 선택한 k개의 숫자가 모두 나타나도록 다섯 자리를 채우는 방법의 수를 곱한 것이다. 두 번째 개수는 포함배제의 원리(inclusion–exclusion)로 구한다. 예를 들어 $k = 3$이면 그 값은 $3^5 - 3 \cdot 2^5 + 3 \cdot 1^5 = 150$이다.

k	1	2	3	4	5	합계
숫자열 수	$5 \cdot 1$	$10 \cdot 30$	$10 \cdot 150$	$5 \cdot 240$	$1 \cdot 120$	
	5	300	1500	1200	120	3125

그러면

$$\mathrm{E}(N) = \frac{1\cdot 5 + 2\cdot 300 + 3\cdot 1500 + 4\cdot 1200 + 5\cdot 120}{3125} = \frac{10505}{3125} = 3.3616.$$

이다. 이 방법도 통하지만 더 길고 틀릴 수 있는 곳이 더 많다. 지시확률변수를 이용하는 방법은 확률 하나만 구하면 된다.

검산

- 표의 개수를 모두 더하면 $3125 = 5^5$이므로, 모든 숫자열이 정확히 한 번씩 세어졌음을 확인할 수 있다.
- 답이 타당하다. N은 1과 5 사이에 있고, 다섯 개의 값에서 다섯 번 뽑으면 중복이 꽤 생길 것으로 예상되므로 3보다 약간 큰 값은 그럴듯하다. 일반적인 검산으로, m개의 값에서 n번 뽑으면 같은 논리로 $m\left(1-\left(1-\frac{1}{m}\right)^n\right)$을 얻는데, 이 값은 $n=1$일 때 마땅히 그래야 하듯 1이다.
- 컴퓨터로 10000부터 99999까지의 모든 정수를 나열하고, 숫자 1–5만 사용하는 3125개를 골라 서로 다른 숫자의 개수를 평균 내면 정확히 $\frac{2101}{625} = 3.3616$을 얻는다.

흔한 실수

- 표본공간을 90000개의 정수 전체로 보는 것. 조건에 의해 표본공간은 같은 확률을 가지는 5^5개의 숫자열로 줄어든다.
- 기댓값의 선형성에 독립성이 필요하다고 생각하는 것. 독립성은 필요하지 않으며, 바로 그 때문에 지시확률변수가 그토록 유용하다.
- P(숫자j가 나타남)을 다섯 자리 각각에 대해 $\frac{1}{5}$씩 더해서 구하는 것. 이렇게 하면 j가 두 번 이상 나타나는 숫자열을 여러 번 세게 되어 1이 나온다. 여사건을 이용해야 한다.

STEP 너머

이것은 고전적인 "점유 문제(occupancy problem)" 또는 쿠폰 수집가 문제(coupon-collector problem)이다. n개의 공을 m개의 상자에 무작위로 던지고, N은 결국 공이 들어 있는 상자의 개수이다. 지시확률변수를 이용한 기법 $\mathrm{E}(N) = m\left(1-\left(1-\frac{1}{m}\right)^n\right)$은 대학 1학년 확률론과 해싱(hashing) 같은 알고리즘 분석에서 가장 유용한 도구 중 하나이다. 정확한 분포에는 제2종 스털링 수(Stirling numbers of the second kind)가 등장하는데, 이는 위의 포함배제 계산에서 이미 나타났다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 책 전체에서 지시확률변수를 사용하며, 다음 단계로 읽기에 훌

륭한 책이다.

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.