풀이

STEP III 2008, 문제 13

확률과 통계 • 지시확률변수의 합

STEP III 2008, 13번 문제

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

step.maths.org/questions/08-s3-q13

이 문제는 *Beyond STEP*의 12.10 절 **느슨한 끈 끝 묶기**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch12/loose-ends.html

핵심 아이디어

각 단계에서 무슨 일이 일어나든 끈 조각의 개수는 언제나 정확히 하나씩 줄어든다. 고리 하나가 판에서 빠지거나, 두 조각이 하나로 합쳐지기 때문이다. 따라서 이전에 무슨 일이 있었든, 매 단계가 끝나면 조각이 하나 적은 같은 상황으로 돌아온다. 그래서 전체 고리의 개수를 **독립**(*independent*)인 지시변수(indicator variable)들의 합으로 쓸 수 있고, 평균(mean)과 분산(variance)은 단순히 합으로 구해진다.

풀이

첫 번째 단계

조각이 n개이면 자유로운 끝이 $2n$개 있으므로, 두 끝을 고르는 방법은

$$\binom{2n}{2} = n(2n-1)$$

가지이고 모두 같은 확률(probability)로 일어난다. 두 끝이 같은 조각에 속할 때 정확히 고리가 만들어지며, 그런 쌍은 조각마다 하나씩 n개 있다. 따라서

$$\mathrm{P}(\text{첫 번째 단계에서 고리가 생김}) = \frac{n}{n(2n-1)} = \frac{1}{2n-1}.$$

같은 이야기를 다르게 하면, 먼저 어느 끝을 고르든 남은 $2n-1$개의 끝 가운데 그 짝은 정확히 하나이다. 첫 번째 단계에서 고리가 생기면 1, 그렇지 않으면 0인 값을 X_n이라 하면, 첫

번째 단계에서 생기는 고리 개수의 기댓값(expectation)은

$$\mathrm{E}(X_n) = \frac{1}{2n-1}.$$

고리 개수의 기댓값

첫 번째 단계가 끝나면 끈 조각은 $n-1$개가 되고, 각 조각에는 자유로운 끝이 두 개씩 있다. 이는 두 경우 모두 성립한다. 고리가 생기면 조각 하나가 판에서 빠지고, 서로 다른 두 조각을 이으면 두 조각이 하나가 된다. 조각의 길이는 상관없고 끝만이 중요하다.

> ? **생각해 보기** 이제 어떤 조각이 다른 조각보다 길다는 사실이나, 앞에서 고리가 만들어졌는지 아닌지가 문제가 될까?

그렇지 않다. 어떤 끝의 쌍이 선택되는지에 영향을 주는 것은 남아 있는 조각의 개수(따라서 끝의 개수)뿐이다. 따라서 조각이 k개 남아 있을 때, 그 단계에서 고리가 생길 확률은 앞선 모든 단계와 관계없이 $\frac{1}{2k-1}$이다.

각 단계를 그 직전에 있는 조각의 개수 $k = n, n-1, \ldots, 1$로 이름 붙이고, 그 단계에서 고리가 생기는 사건의 지시변수를 X_k라 하자. 그러면 $\mathrm{P}(X_k = 1) = \frac{1}{2k-1}$이고, 과정은 $k = 1$인 단계 뒤에 끝난다($X_1 = 1$이므로 이 단계에서는 언제나 고리가 생긴다). 전체 고리의 개수는

$$R = X_1 + X_2 + \cdots + X_n,$$

이므로 기댓값의 선형성(linearity of expectation)에 의해

$$\mathrm{E}(R) = \sum_{k=1}^{n} \frac{1}{2k-1} = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1}.$$

분산

분산을 구하려면 선형성만으로는 부족하고, X_k들이 독립이어야 한다. 실제로 독립이다. 앞선 단계들의 결과가 어떻게 주어지든, 조각이 k개인 단계에서 고리가 생길 확률은 여전히 $\frac{1}{2k-1}$이다. 따라서 앞선 모든 X들이 주어졌을 때 X_k의 조건부 분포(conditional distribution)는 조건이 없을 때의 분포와 같고, 그러므로 X_k들은 독립이다.

각 X_k는 $p_k = \frac{1}{2k-1}$인 베르누이(Bernoulli) 확률변수이므로 $\mathrm{Var}(X_k) = p_k(1-p_k)$이고, 독립인 확률변수들의 분산은 더해진다.

$$\begin{aligned}\mathrm{Var}(R) &= \sum_{k=1}^{n} \frac{1}{2k-1}\left(1 - \frac{1}{2k-1}\right) = \sum_{k=1}^{n} \frac{2k-2}{(2k-1)^2} \\ &= \sum_{k=1}^{n} \frac{1}{2k-1} - \sum_{k=1}^{n} \frac{1}{(2k-1)^2}.\end{aligned}$$

$n = 40\,000$인 경우

$H_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}$이라 쓰자. $2n-1$까지의 홀수의 역수의 합은 $2n$까지의 모든 역수의 합에서 짝수의 역수를 뺀 것이다.

$$\sum_{k=1}^{n} \frac{1}{2k-1} = H_{2n} - \left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2n}\right) = H_{2n} - \tfrac{1}{2}H_n.$$

m이 클 때 $H_m \approx \ln m$을 이용하면

$$\mathrm{E}(R) \approx \ln(2n) - \tfrac{1}{2}\ln n = \tfrac{1}{2}\left(2\ln 2n - \ln n\right) = \tfrac{1}{2}\ln(4n).$$

$n = 40\,000$이면 $4n = 160\,000 = 20^4$이므로

$$\mathrm{E}(R) \approx \tfrac{1}{2}\ln\left(20^4\right) = 2\ln 20 \approx 6.$$

따라서 끈 조각 40 000개로부터 기대할 수 있는 고리는 약 여섯 개에 불과하다.

검산

- $n = 1$: 조각이 하나이면 고리는 반드시 하나 생긴다. 공식에서도 $\mathrm{E}(R) = 1$, $\mathrm{Var}(R) = 0$이 나오므로 맞다.
- $n = 2$: 첫 번째 단계에서 고리가 생길 확률은 $\frac{1}{3}$이고 마지막 단계에서는 언제나 고리가 생기므로, R는 확률 $\frac{1}{3}$로 2이고 그 밖의 경우 1이다. 그러면 평균은 $\frac{4}{3}$, 분산은 $\frac{2}{9}$이며, 공식에서 얻은 $1 + \frac{1}{3}$, $\frac{2}{9}$와 일치한다.
- $n \le 5$에 대해 모든 선택의 순서를 정확히 나열해 보고 $n = 6$을 시뮬레이션해 보면 두 공식과 모두 일치한다. $n = 40\,000$일 때 정확한 합은 $\mathrm{E}(R) \approx 6.28$이다. 우리의 추정값보다 0.29만큼 더 큰 것은 $H_m \approx \ln m$이라는 거친 근사 때문이다(아래 참고).

흔한 실수

- 고리가 생길 확률을 $\frac{1}{2n-1}$(두 **끝**을 고름) 대신 $\frac{1}{n}$(두 **끈**을 고름)으로 잡는 것.
- X_k들이 독립임을 먼저 정당화하지 않고 분산을 더하는 것. 평균에는 선형성으로 충분하지만 분산에는 그렇지 않다.
- 짝수 항을 떼어 내지 않고 $\sum \frac{1}{2k-1}$을 $\ln n$이나 $\frac{1}{2}\ln n$으로 근사하는 것.

STEP 너머

더 정밀한 추정은 $H_m = \ln m + \gamma + o(1)$이며, 여기서 $\gamma \approx 0.5772$는 오일러-마스케로니 상수(Euler–Mascheroni constant)이다. 이로부터 $\mathrm{E}(R) \approx \frac{1}{2}\ln(4n) + \frac{1}{2}\gamma$를 얻고,

$\sum 1/(2k-1)^2 = \pi^2/8$이므로 $\mathrm{Var}(R) \approx \frac{1}{2}\ln(4n) + \frac{1}{2}\gamma - \frac{\pi^2}{8}$도 얻는다. 같은 "확률이 $1/k$인 독립 지시변수들의 합" 구조는 무작위 순열(random permutation)의 순환(cycle)의 개수와 무작위 수열의 기록값(record)의 개수를 센다. 두 경우 모두 그 개수는 $\ln n$처럼 증가한다. 지시변수와 이러한 무작위 구조를 더 깊이 공부하려면 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 좋은 출발점이다.

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.