9.11 한 세트 모두 모으기

학습지

문제

STEP II 1994, 13번 문제

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

step.maths.org/questions/94-s2-q13

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** 각 봉지에서 독립적으로 확률 p로 새 카드가 나온다면, 첫 번째 새 카드가 나올 때까지(그 봉지 포함) 뜯는 봉지 수의 기댓값은 얼마인가?

1단계 새 카드를 기다리기

다루는 내용: 새 카드를 기다리기

? **생각해 보기** 이미 서로 다른 카드 r장을 가지고 있을 때, 다음 봉지에서 새 카드가 나올 확률(probability)은 얼마인가? 기하분포(geometric distribution)의 기대 대기 시간(expected wait)을 이용한다.

2단계 합과 적분 비교하기

다루는 내용: 적분 부등식

생각해 보기 $r \le x \le r+1$에서 $\frac{1}{x}$는 어떤 두 값 사이에 있는가? 길이가 1인 이 구간에서 적분(integral)한다.

잠깐, 직접 해 보기 $r = 1$부터 $r = n-1$까지의 부등식(inequality)을 모두 더한다. 적분들이 이어져 하나의 적분이 된다. 어떤 적분인가?

3단계 전체 세트 모으기

다루는 내용: 전체 세트 모으기

생각해 보기 전체 봉지의 개수는 새 카드 각각을 얻기까지의 대기 시간을 모두 더한 것이다. 그 기댓값(expectation)들을 더하고, 두 번째 단계를 이용하여 그 합을 $\ln n$과 비교한다.

숨은 수학

이 추정은 얼마나 정확한가?

비교를 더 정밀하게 할 수 있다. 그림 9.11.1 (온라인 전체 절에 수록)에서 곡선과 위쪽 직사각형 사이의 색칠된 틈들은 넓이의 합이 유한하므로, $H_n - \ln n$은 어떤 극한으로 수렴(converge)한다.

$$\gamma = \lim_{n\to\infty} (H_n - \ln n) = 0.5772156649\ldots,$$

이것이 **오일러 상수**(*Euler's constant*)이다. (γ가 유리수(rational)인지는 아무도 모른다.) 따라서 봉지 수의 기댓값은 $n\ln n + \gamma n + \frac{1}{2} + \cdots$이다. 스티커 600장짜리 앨범의 경우 $600 \times (\ln 600 + 0.577) \approx 4185$로, 도입에서 말한 수치가 바로 이것이다.

시간은 어디에 쓰이는가? 각 단계는 매우 불균등하다. 세트의 처음 절반을 모으는 데는 약 $n\ln 2 \approx 0.69n$봉지밖에 들지 않지만, 마지막 카드 하나에만 평균 n봉지가 든다. 노력의 대부분은 마지막 몇 장에 들어가며, 그래서 수집가들은 중복된 카드를 서로 교환한다.

집중

기댓값은 시작일 뿐이다. 전체는 얼마나 변동하는가? 전체의 분산(variance)은 $\sum nr/(n-r)^2$이고, 이는 약 $\frac{\pi^2}{6}n^2$이다(Section 2.4에서 나온 합 $\sum 1/k^2 = \pi^2/6$이 등장한다). 따라서 표준편차(standard deviation)는 약 $1.28n$으로, n이 클 때 평균 $n\ln n$보다 훨씬 작다. 즉 전체는 평균 주위에 **집중**(*concentrated*)되어 있다. 폴 에르되시(Paul Erdős)와 알프레드 레니(Alfréd Rényi)는 1961년에 더 정확하게, 전체에서 $n\ln n$을 빼고 n으로 나눈 값이 극한분

포(limiting distribution)를 가지며, 그것이 **굼벨 분포**(*Gumbel distribution*)임을 보였다. 굼벨 분포는 많은 확률적인 양 가운데 가장 큰 값을 기술하는 분포이기도 하다(Section 12.25).

어디에나 있는 쿠폰

이 문제는 아브라암 드무아브르(Abraham de Moivre, 1712)와 라플라스(Laplace)까지 거슬러 올라간다. 오늘날 이 문제는 여러 모습으로 나타난다. 시스템의 모든 부품을 점검하려면 무작위 검사가 몇 번 필요한가? 네트워크 위의 랜덤 워크(random walk)가 모든 노드를 방문할 때까지 얼마나 걸리는가? 컴퓨터 해시 테이블(hash table)을 몇 번 탐색해야 하는가? 어느 경우든 답에는 $n \ln n$이 들어 있다. 생물학에서는 같은 계산으로 유전체 전체를 덮는 데 필요한 염기서열 분석 "리드(reads)"의 수를 추정한다.

더 도전하기

1 **세트의 절반.** (n이 짝수일 때) 서로 다른 카드 $\frac{n}{2}$장을 모으는 데 필요한 봉지 수의 기댓값이 $n(H_n - H_{n/2})$이고, 이것이 약 $n \ln 2$임을 보여라.

2 **서로 다른 카드.** k봉지를 뜯은 뒤 가지고 있는 서로 다른 카드 수의 기댓값이 $n(1 - (1 - \frac{1}{n})^k)$임을 보여라. (**힌트:** 각 카드에 대해 그 카드를 가지고 있을 확률을 구하고, 모두 더하라.) 대략 몇 봉지를 뜯으면 세트의 95%를 얻는가?

3 **두 수집가.** 두 아이가 같은 세트를 모으면서 봉지를 함께 나누어, 한 아이에게 중복인 카드가 다른 아이에게는 새 카드가 될 수 있다. 두 세트를 모두 완성하는 데 필요한 봉지 수의 기댓값이 약 $n(\ln n + \ln \ln n)$임을 보여라. (이 문제는 더 어렵다. 먼저 이것이 $2n \ln n$보다 훨씬 작다는 것을 납득해 보라.)

4 **더 정밀한 한계.** 직사각형 대신 사다리꼴 공식(trapezium rule)의 그림을 이용하여 $H_n - \ln n$이 감소하고 $H_n - \ln n - \frac{1}{2n}$이 증가함을 보여라. 작은 n 값들로부터 $0.5 < \gamma < 0.7$을 이끌어 내라.

더 읽을거리

- **Geoffrey Grimmett & David Stirzaker, *One Thousand Exercises in Probability*** (Oxford University Press). 쿠폰 수집가 문제의 다양한 변형이 많이 실려 있다.
- **Julian Havil, *Gamma: Exploring Euler's Constant*** (Princeton University Press). γ와 조화급수에 관한 모든 것을 쉽게 읽을 수 있는 수준으로 다룬다.
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing*** (Cambridge University Press). 쿠폰 수집가 문제와 컴퓨터 과학에서의 활용을 다룬다.

안내에 따른 전체 풀이는 제 9.11절에 있다:

https://beyondstep.keithmcnulty.org/ko/ch09/coupon-collector.html

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