풀이

STEP II 2014, 문제 13

확률과 통계 • 이산확률변수

> **STEP II 2014, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/14-s2-q13

핵심 아이디어

이 문제는 모습을 바꾼 생일 문제(birthday problem)이다. 사건 $X \geq k$는 처음 $k-1$개의 정수가 모두 다르다는 뜻일 뿐이고, $X = r$는 처음 $r-1$개가 서로 다르고 r번째가 그중 하나와 같다는 뜻이다. 두 확률(probability) 모두 간단한 인수들의 곱이다. 두 신기한 항등식은 같은 양을 두 가지 방법으로 계산하는 데서 나온다. 소문항 (i)에서는 전체 확률이 1이라는 사실을, 소문항 (iv)에서는 $\mathrm{E}(X)$에 대한 두 공식이 일치한다는 사실을 이용한다.

풀이

이 풀이 전체에서

$$Q_m = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right) \qquad (m \geq 1),$$

로 쓰고, $Q_1 = 1$(빈 곱(empty product))로 둔다. 각 정수는 앞의 정수들과 독립(independent)이고 $\{1,\ldots,n\}$ 위에서 균등(uniform)하게 분포하므로, Q_m은 $I_1,\ldots,I_m$이 모두 다를 확률이다. 처음 j개가 서로 다를 때, 다음 정수가 그 모두를 피할 확률은 $1-\frac{j}{n}$이기 때문이다.

$\mathrm{P}(X=4)$의 값

I_1, I_2, I_3이 서로 다르고, 그다음 I_4가 그 세 값 중 하나와 같아야 한다.

$$\mathrm{P}(X=4) = 1\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdot\frac{3}{n} = \frac{3(n-1)(n-2)}{n^3}.$$

소문항 (i)

같은 방법으로, $X = r$는 $I_1, \ldots, I_{r-1}$이 서로 다르고(확률 Q_{r-1}) I_r이 이 $r-1$개의 값 중 하나라는(독립적으로 확률 $\frac{r-1}{n}$) 뜻이다. 따라서

$$\mathrm{P}(X = r) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{r-2}{n}\right)\frac{r-1}{n} = Q_{r-1}\,\frac{r-1}{n}, \qquad 2 \le r \le n+1.$$

? **생각해 보기** 이 확률들을 모두 더하면 왜 1이 되는가?

X는 반드시 존재하고 2와 $n+1$ 사이에 있다. 첫 번째 정수는 반복일 수 없으므로 $X \ge 2$이다. 또 $I_1, \ldots, I_{n+1}$은 $n+1$개의 정수인데 가능한 값은 n개뿐이므로, 비둘기집 원리(pigeonhole principle)에 의해 $n+1$번째까지 어떤 값은 두 번 나타나고, 따라서 $X \le n+1$이다. (수열이 끝내 반복되지 않을 확률은 0이므로 빠진 경우는 없다.) 그러므로

$$\sum_{r=2}^{n+1} \mathrm{P}(X = r) = 1.$$

$r = 2, 3, 4, \ldots$ 인 항들을 풀어 쓰면

$$\frac{1}{n} + \left(1 - \frac{1}{n}\right)\frac{2}{n} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\frac{3}{n} + \cdots = 1,$$

이고, 이것이 바로 구하는 항등식이다. 이 급수(series)는 $r = n+1$에서 끝나며, 그 이후의 항들은 인수 $1 - \frac{n}{n}$을 포함하므로 모두 0이다. 증명 끝. $m = r - 1$로 놓으면 이 식은

$$\sum_{m=1}^{n} \frac{m}{n}\,Q_m = 1. \qquad (*)$$

이 된다.

소문항 (ii)

기댓값(expectation)의 정의에 의해

$$\mathrm{E}(X) = \sum_{r=2}^{n+1} r\,\frac{r-1}{n}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{r-2}{n}\right) = \sum_{m=1}^{n} \frac{m(m+1)}{n}\,Q_m.$$

소문항 (iii)

$X \ge k$는 $I_1, \ldots, I_{k-1}$ 사이에 반복이 없을 때, 즉 이 $k-1$개의 정수가 모두 다를 때 정확히 일어난다. 따라서

$$\mathrm{P}(X \ge k) = Q_{k-1} = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right), \qquad 2 \le k \le n+2,$$

이고, $\mathrm{P}(X \ge 1) = \mathrm{P}(X \ge 2) = 1$이다. ($k = n+2$일 때 이 곱은 $1 - \frac{n}{n} = 0$을 포함하므로, 마땅히 그래야 하듯 0이 된다.)

소문항 (iv)

$p_r = \mathrm{P}(Y = r)$로 쓰자. 그러면 $r = \sum_{k=1}^{r} 1$이므로

$$\mathrm{E}(Y) = \sum_{r=1}^{N} r\,p_r = \sum_{r=1}^{N}\sum_{k=1}^{r} p_r = \sum_{k=1}^{N}\sum_{r=k}^{N} p_r = \sum_{k=1}^{N} \mathrm{P}(Y \geq k).$$

합이 유한하므로 합의 순서를 바꾸어도 문제없다. 두 이중합 모두 $1 \leq k \leq r \leq N$인 모든 순서쌍 (k, r)에 대한 합이기 때문이다. 증명 끝.

> ? **생각해 보기** $\mathrm{E}(X)$에 대한 어떤 두 식을 비교해야 하는가?

이 결과를 $Y = X$에 적용하자. X는 $N = n+1$일 때 $\{1, \ldots, N\}$에 속하는 값을 가진다(값 1은 확률이 0일 뿐이다). 소문항 (iii)을 이용하면

$$\mathrm{E}(X) = \sum_{k=1}^{n+1} \mathrm{P}(X \geq k) = 1 + \sum_{k=2}^{n+1} Q_{k-1} = 1 + \sum_{m=1}^{n} Q_m.$$

이제 이것을 소문항 (ii)의 결과와 같다고 놓고, $\frac{m(m+1)}{n} = \frac{m^2}{n} + \frac{m}{n}$으로 나눈 뒤 $(*)$를 이용한다.

$$\sum_{m=1}^{n} \frac{m(m+1)}{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + \underbrace{\sum_{m=1}^{n} \frac{m}{n} Q_m}_{=1 \text{ 근거: } (*)}.$$

따라서 $1 + \sum_{m=1}^{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + 1$, 즉

$$\sum_{m=1}^{n} Q_m \left(1 - \frac{m^2}{n}\right) = 0.$$

$m = 1, 2, 3$인 항은 $\left(1 - \frac{1^2}{n}\right)$, $\left(1 - \frac{1}{n}\right)\left(1 - \frac{2^2}{n}\right)$, $\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3^2}{n}\right)$이므로, 이것이 정확히 구하는 항등식이다. 소문항 (i)에서와 마찬가지로 $m = n$ 이후의 모든 항은 인수 $1 - \frac{n}{n} = 0$을 가진다. 증명 끝.

검산

- $n = 1$이면 반드시 $X = 2$이다. 소문항 (i)은 $\frac{1}{1} = 1$을 주고, 마지막 합은 단 하나의 항 $1 - 1 = 0$이다.
- $n = 4$이면 마지막 합은 $\frac{3}{4} + \frac{3}{4}\cdot 0 + \frac{3}{4}\cdot\frac{1}{2}\cdot\left(-\frac{5}{4}\right) + \frac{3}{4}\cdot\frac{1}{2}\cdot\frac{1}{4}\cdot(-3) = \frac{24-15-9}{32} = 0$이고, 두 공식 모두 $\mathrm{E}(X) = \frac{103}{32}$를 준다.
- $n = 365$(생일)이면 $\mathrm{E}(X) = 1 + \sum Q_m \approx 24.62$이다. 10^5개의 수열에 대한 컴퓨터 시뮬레이션에서 평균은 24.61이었고, $n \leq 5$에 대한 정확한 전수 조사로 위의 모든 공식을 확인하였다.

흔한 실수

- $\mathrm{P}(X = r)$를 쓸 때 $1 - \frac{j}{n}$ 꼴의 인수를 $r - 2$개가 아니라 $r - 1$개로 쓰는 실수. 앞의 값들을 피해야 하는 것은 두 번째부터 $(r - 1)$번째 정수까지뿐이다.
- $\sum \mathrm{P}(X \geq k)$에서 $k = 1$인 항을 빠뜨리는 실수. 이 항은 1을 기여하며, 바로 이것이 $(*)$와 상쇄된다.

STEP 너머

꼬리합 공식(tail-sum formula) $\mathrm{E}(Y) = \sum_k \mathrm{P}(Y \geq k)$는 음이 아닌 정수 값을 가지는 모든 확률변수(random variable)에 대해 성립하며, 그 연속 버전 $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$는 대학 1학년 확률론의 표준적인 도구이다. n이 클 때 $\mathrm{E}(X) \approx \sqrt{\pi n/2}$이며, 이 때문에 해시 함수(hash function)의 충돌은 고작 $\sqrt{n}$번 정도의 시도 만에 나타난다(이른바 "생일 공격(birthday attack)"). Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문서)에서 두 아이디어를 모두 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.