풀이

STEP I 1996, 문제 12

확률과 통계 • 확률표본의 최댓값

> **STEP I 1996, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/96-s1-q12

핵심 아이디어

채점자의 선택은 결국 $\{1, 2, \ldots, m\}$에서 서로 다른 점수 n개로 이루어진 **집합**(set)을 고르는 것이고, 이런 집합은 모두 같은 확률로 나온다. 따라서 확률(probability)은 집합의 개수를 $\binom{m}{n}$으로 나눈 값이다. 가장 큰 점수가 k라고 하면 점수 하나가 정해지고, 나머지 $n-1$개는 그보다 작은 점수 중에서 고르게 된다. 확률의 합이 1이므로 항등식(identity)은 저절로 따라 나온다. 기댓값(expectation)을 구할 때의 요령은 $k\binom{k-1}{n-1}$을 다시 하나의 이항계수(binomial coefficient)로 바꾸어 같은 항등식을 한 번 더 쓰는 것이다.

풀이

K의 분포

채점자는 서로 다른 점수 n개를 주며, 대칭성에 의해 $\{1, \ldots, m\}$의 크기가 n인 부분집합은 모두 같은 확률로 사용된 점수의 집합이 된다. (어느 답안지가 어느 점수를 받는지는 K에 아무 영향을 주지 않는다.) 이런 부분집합은 $\binom{m}{n}$개 있다.

사건 $K = k$는 집합이 k를 포함하고 나머지 $n-1$개의 점수가 모두 $\{1, \ldots, k-1\}$에 속할 때, 그리고 그때에만 일어난다. 그 $n-1$개의 점수를 고르는 방법은 $\binom{k-1}{n-1}$가지이다. 이를 위해서는 $k - 1 \geq n - 1$, 즉 $k \geq n$이어야 한다. 서로 다른 양의 정수 n개 중 가장 큰 수는 적어도 n이기 때문이다. 물론 $k \leq m$이기도 하다.



(그림은 $m = 10, n = 4, k = 7$인 경우이다.) 따라서

$$\mathrm{P}(K = k) = \binom{k-1}{n-1} \Big/ \binom{m}{n}, \qquad n \le k \le m.$$

항등식

K는 $n, n+1, \ldots, m$ 중 정확히 하나의 값을 가져야 하므로, 이 확률들의 합은 1이다.

$$\sum_{k=n}^{m} \binom{k-1}{n-1} \Big/ \binom{m}{n} = 1 \quad \Longrightarrow \quad \sum_{k=n}^{m} \binom{k-1}{n-1} = \binom{m}{n}.$$

이 식은 $1 \le n \le m$인 **모든** 정수에 대하여 성립하며, 다음 단계에서 이 사실이 필요하다.

기댓값

정의에 의해

$$\mathrm{E}(K) = \frac{1}{\binom{m}{n}} \sum_{k=n}^{m} k \binom{k-1}{n-1}.$$

> ? **생각해 보기** 항등식을 다시 적용할 수 있도록 인수 k를 없애려면 어떻게 해야 할까?

k를 이항계수 안으로 흡수시킨다.

$$k\binom{k-1}{n-1} = \frac{k\,(k-1)!}{(n-1)!\,(k-n)!} = n \cdot \frac{k!}{n!\,(k-n)!} = n\binom{k}{n}.$$

이제 $\binom{k}{n} = \binom{(k+1)-1}{(n+1)-1}$이다. k가 n부터 m까지 변할 때 $j = k+1$은 $n+1$부터 $m+1$까지 변한다. 따라서 n과 m 대신 $n+1$과 $m+1$을 넣어 항등식을 쓰면($1 \le n+1 \le m+1$이므로 가능하다)

$$\sum_{k=n}^{m} \binom{k}{n} = \sum_{j=n+1}^{m+1} \binom{j-1}{(n+1)-1} = \binom{m+1}{n+1}.$$

그러므로

$$\begin{aligned} \mathrm{E}(K) &= \frac{n\binom{m+1}{n+1}}{\binom{m}{n}} = n \cdot \frac{(m+1)!}{(n+1)!\,(m-n)!} \cdot \frac{n!\,(m-n)!}{m!} \\ &= \frac{n(m+1)}{n+1}. \end{aligned}$$

검산

▶ $n = m$: 모든 점수가 사용되므로 확실히 $K = m$이고, 실제로 $\frac{m(m+1)}{m+1} = m$이다.

- $n = 1$: K는 점수 하나로서 $\{1, \ldots, m\}$ 위의 균등분포(uniform distribution)를 따르고 평균은 $\frac{m+1}{2}$이므로 일치한다.
- $m = 6, n = 2$: $\{1, \ldots, 6\}$에서 나오는 15개의 쌍을 모두 나열하면 $\mathrm{E}(K) = \frac{70}{15} = \frac{14}{3}$이고, 공식으로는 $\frac{2 \cdot 7}{3} = \frac{14}{3}$이다. 우리는 $1 \leq n \leq m \leq 9$인 모든 경우에 대해 부분집합을 모두 나열하여 분포, 항등식, 평균을 확인하였다.
- 답을 보는 멋진 방법이 있다. 선택된 n개의 점수는 나머지 $m - n$개의 점수를 $n + 1$개의 틈으로 나누며, 각 틈의 크기의 기댓값은 $\frac{m-n}{n+1}$로 모두 같다. 그러면 $\mathrm{E}(K) = m - \frac{m-n}{n+1} = \frac{n(m+1)}{n+1}$이다.

흔한 실수

- 분자에서는 답안지에 점수를 배정하는 순서 있는 경우를 세고, 분모에서는 순서 없는 집합을 세는 것(또는 그 반대). 일관되게만 세면 어느 쪽이든 괜찮은데, 인수 $n!$이 약분되기 때문이다.
- $\binom{k-1}{n-1}$ 대신 $\binom{k}{n}$을 쓰는 것. 이것은 가장 큰 점수가 k와 **같은** 집합이 아니라 모든 점수가 k **이하**인 집합의 개수를 센다.

STEP 너머

이 항등식은 파스칼(Pascal)의 삼각형에 나오는 "하키 스틱(hockey-stick)" 항등식이며, 여기서는 같은 대상을 두 가지 방법으로 세는 표준적인 조합론적 기법으로 증명하였다. K는 비복원추출(sampling without replacement)로 얻은 무작위 표본의 최댓값으로, **순서통계량**(order statistic)이다. "독일 전차 문제(German tank problem)"는 이 문제를 거꾸로 뒤집어, 관측된 K로부터 m을 추정한다. $\mathrm{E}(K) = \frac{n(m+1)}{n+1}$을 m에 대해 풀면 불편추정량(unbiased estimator) $\frac{n+1}{n}K - 1$을 얻는다. 순서통계량과 세기 논법에 대해 더 알고 싶다면 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)를 보라.

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.