풀이

STEP III 2018, 문제 12

확률과 통계 • 순서통계량

> **STEP III 2018, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/18-s3-q12

핵심 아이디어

사건 "k번째로 작은 수가 y 이하이다"는 "n개의 수 중 적어도 k개가 y 이하이다"와 같은 사건이고, y 이하인 값의 개수는 이항분포(binomial distribution)를 따른다. 이로부터 분포함수(distribution function)를 곧바로 얻는다. 이를 미분하면 두 개의 합이 나오는데, (ii)의 두 이항계수 항등식이 바로 이 합들을 연쇄적으로 소거(telescope)되게 만든다. 확률밀도함수를 알고 나면, 그것을 적분한 값이 1이라는 사실로부터 어떤 적분값을 공짜로 얻을 수 있고, 같은 적분(n과 k를 옮긴 것)으로 평균도 구할 수 있다.

풀이

n개의 수를 $X_1, \ldots, X_n$이라 하자. 이들은 서로 독립이고, 각각 $[0,1]$에서 균등분포(uniform distribution)를 따른다.

소문항 (i)

$0 \le y \le 1$인 y를 고정하고, $X_i \le y$인 X_i의 개수를 N이라 하자. 각 X_i는 다른 것들과 독립적으로 확률 y로 $X_i \le y$를 만족하므로, $N \sim \mathrm{B}(n, y)$이다.

이제 $Y_k \le y$일 필요충분조건은 수들 중 적어도 k개가 y 이하인 것이다. $Y_k \le y$이면 Y_k와 그보다 작은 $k-1$개의 수가 모두 y 이하이므로 $N \ge k$이다. 역으로 $N \ge k$이면 가장 작은 k개의 수가 모두 y 이하이고, 특히 $Y_k \le y$이다. 따라서

$$\mathrm{P}(Y_k \le y) = \mathrm{P}(N \ge k) = \sum_{m=k}^{n} \binom{n}{m} y^m (1-y)^{n-m},$$

이고, 이것이 $(*)$이다.

소문항 (ii)

계승을 풀어 쓰면, $1 \le m \le n$일 때

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

이고, 마찬가지로 $0 \le m \le n-1$일 때

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

이다. ($m = n$일 때 좌변은 0이고, $\binom{n-1}{n}$을 0으로 읽으면 우변도 0이다.)

확률밀도함수. Y_k의 확률밀도함수(probability density function) $f(y)$는 $(*)$의 도함수이다. 각 항을 곱의 미분법으로 미분하면

$$f(y) = \sum_{m=k}^{n} \binom{n}{m} \left[m\,y^{m-1}(1-y)^{n-m} - (n-m)\,y^{m}(1-y)^{n-m-1} \right].$$

이다. 두 번째 부분에서 $m = n$인 항은 인수 $n-m$ 때문에 사라지므로, 그 합은 실제로는 $m = n-1$까지이다. 두 항등식을 쓰면

$$f(y) = n \sum_{m=k}^{n} \binom{n-1}{m-1} y^{m-1}(1-y)^{n-m} \;-\; n \sum_{m=k}^{n-1} \binom{n-1}{m} y^{m}(1-y)^{n-1-m}.$$

이다.

> ? **생각해 보기** 이 두 합은 거의 똑같아 보인다. 어떻게 하면 둘을 나란히 맞출 수 있을까?

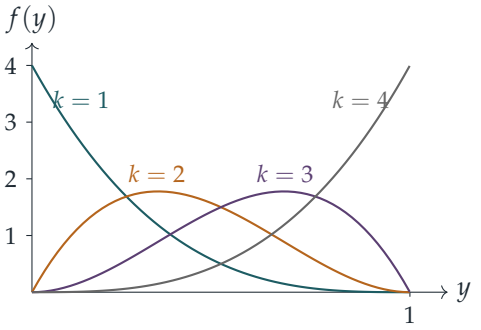
두 번째 합에서 $j = m+1$로 놓자. 그러면 합은

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

이 되는데, 이는 k가 아니라 $k+1$에서 시작한다는 점만 빼면 첫 번째 합과 항별로 똑같다. 첫 번째 합의 $m = k$인 항을 제외한 모든 것이 소거되어

$$f(y) = n\binom{n-1}{k-1} y^{k-1}(1-y)^{n-k},$$

를 얻으며, 이것이 구하는 결과이다. ($k = n$이면 두 번째 합은 비어 있고 첫 번째 합에는 $m = n$인 항만 있으므로, 결과는 여전히 성립한다.)



$n=4$일 때의 확률밀도함수들: 네 수 중 k번째로 작은 수는 $k/5$ 근처에 몰려 있다.

적분. 확률밀도함수를 그 범위(여기서는 $[0,1]$)에서 적분하면 1이다. 그러므로

$$n\binom{n-1}{k-1}\int_0^1 y^{k-1}(1-y)^{n-k}\,\mathrm{d}y = 1,$$

이고, 따라서

$$\int_0^1 y^{k-1}(1-y)^{n-k}\,\mathrm{d}y = \frac{1}{n\binom{n-1}{k-1}} = \frac{(k-1)!\,(n-k)!}{n!}.$$

이다. 이는 $1 \le k \le n$인 모든 정수에 대해 성립한다.

소문항 (iii)

정의에 의해

$$\mathrm{E}(Y_k) = \int_0^1 y\,f(y)\,\mathrm{d}y = n\binom{n-1}{k-1}\int_0^1 y^{k}(1-y)^{n-k}\,\mathrm{d}y.$$

이다.

> ? **생각해 보기** 이 적분은 방금 구한 적분과 꼭 같지는 않다. n과 k를 어떤 값으로 바꾸면 그 꼴이 될까?

피적분함수는 $K=k+1$, $N=n+1$로 둘 때 $y^{K-1}(1-y)^{N-K}$이고, $1 \le K \le N$이므로 (ii)의 공식을 적용할 수 있다:

$$\int_0^1 y^{k}(1-y)^{n-k}\,\mathrm{d}y = \frac{k!\,(n-k)!}{(n+1)!}.$$

$n\binom{n-1}{k-1} = \dfrac{n!}{(k-1)!\,(n-k)!}$이므로

$$\mathrm{E}(Y_k) = \frac{n!}{(k-1)!\,(n-k)!}\cdot\frac{k!\,(n-k)!}{(n+1)!} = \frac{k}{n+1}.$$

이다.

검산

- $n = 1, k = 1$이면 균등분포를 따르는 수 하나뿐이고, 마땅히 그래야 하듯 $\mathrm{E}(Y_1) = \frac{1}{2}$이다.
- 대칭성: 모든 X_i를 $1 - X_i$로 바꾸면 k번째로 작은 수가 k번째로 큰 수가 되므로 $\mathrm{E}(Y_k) + \mathrm{E}(Y_{n+1-k}) = 1$이다. 실제로 $\frac{k}{n+1} + \frac{n+1-k}{n+1} = 1$이다.
- n개의 수는 $[0,1]$을 $n+1$개의 간격으로 나누는데, 답에 따르면 각 간격의 평균 길이는 $\frac{1}{n+1}$이다. 이는 대칭성으로 보아 매우 그럴듯하다.
- $n = 5, k = 2$일 때: $y = 0.3$에서 $(*)$의 값은 0.47178이고, 200 000개의 표본으로 컴퓨터 시뮬레이션을 한 결과 0.4726을 얻었으며, Y_2의 표본평균은 $\frac{1}{3}$에 대해 0.33333이었다.

흔한 실수

- **적어도** k개가 아니라 y 아래에 **정확히** k개의 값이 있을 확률인 $\mathrm{P}(Y_k \leq y) = \binom{n}{k} y^k (1-y)^{n-k}$로 쓰는 것.
- 두 번째 합의 $m = n$인 항이 0이라는 것을 잊어서, 연쇄 소거가 한 항 차이로 실패하는 것처럼 보이게 되는 것.
- (iii)에서 적분을 다시 쓸 때 k만 옮기고 n은 옮기지 않는 것.

STEP 너머

Y_k들을 표본의 **순서통계량**(*order statistics*)이라 하며, Y_k는 모수가 k와 $n-k+1$인 베타분포(Beta distribution)를 따른다. (ii)의 적분은 베타함수 $B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}\,\mathrm{d}y = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$이며, 이 함수는 감마함수(Gamma function)와 함께 다시 만나게 될 것이다. 순서통계량은 표본중앙값과 비모수 통계학(non-parametric statistics)의 바탕이 된다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 순서통계량과 베타분포를 아름답게 다루며, 여기서의 논증과 매우 비슷한 베타 적분의 "이야기" 증명도 담고 있다.

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.