풀이

STEP III 2011, 문제 13

확률과 통계 • 이산분포의 최빈값

> **STEP III 2011, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/11-s3-q13

핵심 아이디어

이산 확률변수(discrete random variable)가 가장 가질 법한 값을 찾을 때는 미분하지 않는다. 대신 이웃한 확률끼리 비교한다. 비 $\mathrm{P}(X=r+1)/\mathrm{P}(X=r)$는 r에 대한 간단한 유리함수(rational function)이고, 이 비가 언제 1보다 커지는지를 묻는 것은 r에 대한 일차부등식(linear inequality)이 된다. 비가 1보다 큰 동안에는 확률이 증가하고, 1보다 작아지면 감소하므로, 최댓값은 바뀌기 직전의 마지막 r에서 나타난다. 같은 방법이 복원추출(with replacement)과 비복원추출(without replacement) 모두에 통하며, 달라지는 것은 비뿐이다.

풀이

소문항 (i)

복원추출에서는 각 시행이 독립(independent)이고 각각 확률 $p=b/n$으로 검은 공이 나오므로 $X\sim \mathrm{B}(k,p)$이다.

$$\mathrm{P}(X=r)=\binom{k}{r}p^r(1-p)^{k-r},\qquad r=0,1,\ldots,k.$$

먼저 $0<b<n$이라고 가정하자. 그러면 이 확률들은 모두 양수이다. $0\le r\le k-1$에 대하여

$$\frac{\mathrm{P}(X=r+1)}{\mathrm{P}(X=r)}=\frac{k-r}{r+1}\cdot\frac{p}{1-p}=\frac{(k-r)\,b}{(r+1)(n-b)}.$$

분모가 양수이므로 이 비가 1보다 클 필요충분조건은 다음과 같다.

$$\begin{aligned} & (k-r)b > (r+1)(n-b) \\ \iff \quad & kb - rb > rn - rb + n - b \\ \iff \quad & (k+1)b > (r+1)n \\ \iff \quad & r+1 < \frac{(k+1)b}{n}. \end{aligned}$$

같은 방식으로, 비가 1과 같을 필요충분조건은 $r+1 = m$이고, 1보다 작을 필요충분조건은 $r+1 > m$이다. 여기서

$$m = \frac{(k+1)b}{n}.$$

따라서 r가 증가할 때, $\mathrm{P}(X=r)$는 도착하는 값이 $r+1 < m$인 각 단계에서 순증가하고, 도착하는 값이 $r+1 > m$인 각 단계에서 순감소한다. $0 < b < n$이므로 $0 < m < k+1$이고, 따라서 증가에서 감소로 바뀌는 지점은 범위 $0, \ldots, k$ 안에 있다.

? 생각해 보기 m이 정수가 아니라면 가장 큰 확률은 정확히 어디에 있는가?

m이 정수가 아니면 $M = \lfloor m \rfloor$로 두자. 그러면 $M < m < M+1$이고 $0 \le M \le k$이다. 확률은 $\mathrm{P}(X=M)$까지 순증가하고(모든 단계가 $\le M < m$인 값에서 끝나므로), 그 뒤로는 순감소한다(이후의 모든 단계가 $\ge M+1 > m$인 값에서 끝나므로). 따라서 가장 가능성이 높은 검은 공의 개수는 유일하게

$$\left\lfloor \frac{(k+1)b}{n} \right\rfloor,$$

이고, 이것이 증명하려던 것이다.

유일하지 않은 경우. m이 정수이면 $r = m-1$에서 $r = m$으로 가는 단계의 비가 정확히 1이므로 $\mathrm{P}(X = m-1) = \mathrm{P}(X=m)$이고, 이 두 같은 값은 다른 모든 값보다 크다(확률은 $m-1$까지 순증가하고 m 이후로 순감소한다). $1 \le m \le k$이므로 두 값 모두 가능한 값이다. 그러므로 답이 유일하지 않을 필요충분조건은 ($0 < b < n$일 때) $(k+1)b/n$이 정수인 것이며, 이때 $m-1$과 m이 똑같이 가장 가능성이 높다. 예를 들어 $n=3, b=1, k=5$이면 $m=2$이고, 실제로 $\mathrm{P}(X=1) = \mathrm{P}(X=2) = \frac{80}{243}$이다.

자명한 경우는 여기서 제외된다. $b=0$이면 확실히 $X=0$이고, 이는 공식과 일치한다. $b=n$이면 확실히 $X=k$인데, 공식은 ($k+1$을 주게 되므로) 적용되지 않는다. 위의 비가 $n-b$로 나누는 것을 포함하기 때문이다.

소문항 (ii)

비복원추출에서 X는 초기하분포(hypergeometric distribution)를 따른다. 표본을 k개의 공으로 이루어진 순서 없는 집합으로 뽑는다고 보면, $\binom{n}{k}$개의 집합이 모두 같은 확률로 나오

므로

$$\mathrm{P}(X=r)=\frac{\binom{b}{r}\binom{n-b}{k-r}}{\binom{n}{k}}.$$

$k \le b$이므로 가능한 가장 큰 값은 $r=k$이다. 검지 않은 공은 $n-b$개보다 많이 뽑을 수 없으므로 가장 작은 값은 $r_0=\max(0,k-(n-b))$이다. $r_0 \le r \le k$에서 확률은 양수이고, $r_0 \le r \le k-1$에 대하여 $\binom{a}{r+1}=\binom{a}{r}\frac{a-r}{r+1}$과 $\binom{a}{s-1}=\binom{a}{s}\frac{s}{a-s+1}$을 이용하면

$$\frac{\mathrm{P}(X=r+1)}{\mathrm{P}(X=r)}=\frac{b-r}{r+1}\cdot\frac{k-r}{n-b-k+r+1}.$$

이 범위에서 두 분모는 모두 양수이므로, 비가 1보다 클 필요충분조건은

$$(b-r)(k-r)>(r+1)(n-b-k+r+1).$$

전개하면 좌변은 $bk-r(b+k)+r^2$이고 우변은 $(r+1)(n-b-k+1)+r^2+r$이다. r^2 항이 소거되어 부등식은 다음과 같이 된다.

$$\begin{aligned} & bk-r(b+k)>r(n-b-k+1)+(n-b-k+1)+r \\ \iff\quad & bk+b+k+1>r(n+2)+(n+2) \\ \iff\quad & r+1<\frac{(b+1)(k+1)}{n+2}. \end{aligned}$$

? 생각해 보기 바뀌는 지점이 실제로 가능한 값의 범위 $r_0,\dots,k$ 안에 있는가?

$m=\frac{(b+1)(k+1)}{n+2}$로 두자. $b+1<n+2$이므로 $m<k+1$이다. 아래쪽 끝에 대해서는, $m>0$은 자명하고, $r_0=k+b-n>0$이면 $(b+1)(k+1)>(n+2)(k+b-n)$임을 보여야 한다. $x=n-b$, $y=n-k$로 쓰면 좌변에서 우변을 뺀 값은

$$(n+1-x)(n+1-y)-[(n+1)^2-1-(n+2)(x+y)]=1+x+y+xy>0.$$

따라서 모든 경우에 $r_0<m<k+1$이고, 소문항 (i)의 논증이 그대로 적용된다. 즉 확률은 $r+1<m$인 동안 순증가하고 $r+1>m$이 되면 순감소한다. 그러므로 유일할 때, 가장 가능성이 높은 검은 공의 개수는

$$\left\lfloor\frac{(b+1)(k+1)}{n+2}\right\rfloor.$$

이것이 유일하지 않을 필요충분조건은 $(b+1)(k+1)/(n+2)$가 정수인 것이다. 이때 $m-1$과 m은 ($r_0 \le m-1$이고 $m \le k$이므로 둘 다 가능한 값이며) 같은 확률을 가지고 다른 어떤 값보다도 확률이 크다. 예를 들어 $n=10$, $b=7$, $k=5$이면 $m=4$이고, $X=3$과 $X=4$가 동률이며 각각의 확률은 $\frac{105}{252}=\frac{5}{12}$이다.

검산

- 컴퓨터로 모든 $n \leq 24, 0 \leq b \leq n, 1 \leq k \leq 19$(소문항 (ii)에서는 $k \leq b$)에 대하여 두 분포를 분수로 정확히 계산하고, 최빈값(mode)을 전수 조사로 찾아 모든 경우에 두 공식과 두 동률 조건을 확인하였다(소문항 (ii)는 2565가지 경우).
- $n = 10, b = 6, k = 4$일 때: $m = \frac{35}{12}$이므로 공식은 2를 주고, 실제로 $X = 2$가 유일한 최빈값이다.
- n이 k에 비해 매우 크면, 공 몇 개를 꺼내도 검은 공의 비율은 거의 변하지 않으므로 소문항 (ii)는 소문항 (i)과 비슷해야 한다. 실제로 그렇다. $(b+1)(k+1)/(n+2) \approx (k+1)b/n$.

흔한 실수

- 최빈값이 평균의 정수 부분인 $\lfloor kb/n \rfloor$라고 결론짓는 것. 최빈값은 평균에 가깝지만 항상 평균의 정수 부분과 같지는 않다.
- 비가 정확히 1일 때 최빈값이 하나가 아니라 두 개라는 것을 잊는 것.
- 소문항 (ii)에서 $(r+1)(n-b-k+r+1)$을 전개할 때 부호를 틀리는 것. r^2 항이 소거되는지 보는 것이 유용한 확인 방법이다.

STEP 너머

연속한 확률을 비교하는 것은 어떤 이산분포에서든 최빈값을 찾는 표준적인 방법이며, 같은 계산으로 소문항 (i)의 극한의 경우로서 포아송(Poisson) 분포의 최빈값 $\lfloor \lambda \rfloor$를 얻는다. 초기하분포는 표본추출 이론과 통계학의 피셔(Fisher)의 정확 검정(Fisher's exact test)에서 핵심적인 역할을 한다. 이처럼 확률이 증가하다가 감소하는 분포를 단봉분포(unimodal distribution)라고 한다. 이항분포와 초기하분포의 경우 이 성질은 더 강한 성질인 로그오목성(log-concavity)에서 나온다. 더 알고 싶다면 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 또는 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.