풀이

STEP II 1989, 문제 16

확률과 통계 • 이항분포

> **STEP II 1989, 16번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/89-s2-q16

핵심 아이디어

한 묶음(batch)을 서로 다르게 행동하는 두 부분, 곧 표본(sample)에 들어간 n개의 부품과 한 번도 검사받지 않는 나머지 N개로 나눈다. 각 부분의 불량품 개수는 이항분포(binomial distribution)를 따르고, 두 개수는 서로 독립(independent)이다. 모든 것은 표본에 달려 있다. 표본에 불량품이 두 개 이상 있으면 묶음은 결국 완벽한 상태가 되고, 그렇지 않으면 아무것도 바뀌지 않는다. 따라서 마지막에 불량품이 $k \geq 1$개인 묶음은 표본에 불량품이 0개 또는 1개 있었고, 나머지는 다른 N개 가운데 있었어야 한다.

풀이

$q = 1 - p$라 하자. X를 표본 속 불량 부품의 개수, Y를 묶음의 나머지 N개 가운데 불량 부품의 개수라 하자. 각 부품은 독립적으로 확률 p로 불량이므로

$$X \sim \mathrm{B}(n, p), \qquad Y \sim \mathrm{B}(N, p),$$

이고, X와 Y는 서로소인 부품 집합에서 불량품을 세므로 독립이다.

K를 품질 관리 후 묶음 속 불량 부품의 개수라 하자. $X \geq 2$이면 모든 불량이 바로잡혀 $K = 0$이다. $X \leq 1$이면 아무것도 바로잡히지 않아 $K = X + Y$이다. 따라서

$$K = \begin{cases} X + Y, & X \leq 1, \\ 0, & X \geq 2. \end{cases}$$

K의 분포

$1 \le k \le N$이라 하자. $k \ge 1$이므로 사건 $K = k$는 $X \ge 2$인 경우에서 나올 수 없고, 따라서 정확히 $X = 0,\ Y = k$ 또는 $X = 1,\ Y = k-1$일 때 일어난다. 이 두 사건은 서로 배반이고, 독립성에 의해

$$\begin{aligned}\mathrm{P}(K=k) &= \mathrm{P}(X=0)\,\mathrm{P}(Y=k) + \mathrm{P}(X=1)\,\mathrm{P}(Y=k-1)\\ &= q^n\binom{N}{k}p^k q^{N-k} + npq^{n-1}\binom{N}{k-1}p^{k-1}q^{N-k+1}\\ &= \left[\binom{N}{k} + n\binom{N}{k-1}\right]p^k q^{N+n-k}.\end{aligned}$$

두 항의 p의 거듭제곱과 q의 거듭제곱이 같다는 점 덕분에 두 항을 하나로 묶을 수 있다. 이제 이항계수(binomial coefficient)를 공통분모 $k!\,(N-k+1)!$로 통분하면

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

더하면 분자는 $N!\,(N-k+1+nk) = N!\,[N+1+k(n-1)]$이므로

$$\mathrm{P}(K=k) = \frac{[N+1+k(n-1)]\,N!}{(N-k+1)!\,k!}\,p^k(1-p)^{N+n-k},$$

이고, 이것이 구하려던 식이다.

> **생각해 보기** $k = 0$이면 어떻게 될까?

이 공식은 $k \ge 1$일 때만 성립한다. $k = 0$을 넣으면 $\frac{(N+1)N!}{(N+1)!}q^{N+n} = q^{N+n}$이 되는데, 이는 불량품이 전혀 없을 확률이다. 그러나 $K = 0$은 $X \ge 2$일 때마다 일어나기도 하므로, 실제로는

$$\mathrm{P}(K=0) = q^{N+n} + \mathrm{P}(X \ge 2) = q^{N+n} + 1 - q^n - npq^{n-1}.$$

문제의 "$k \le N$"은 $1 \le k \le N$으로 읽어야 한다.

$k = N+1$인 경우

$K = N+1$이려면 $X \le 1$이고 $X + Y = N+1$이어야 한다. $Y \le N$이므로 $X = 1,\ Y = N$일 수밖에 없고, 따라서

$$\mathrm{P}(K=N+1) = npq^{n-1}\cdot p^N = np^{N+1}q^{n-1}.$$

공식에 $k = N+1$을 넣으면 계수는

$$\frac{[N+1+(N+1)(n-1)]\,N!}{0!\,(N+1)!} = \frac{(N+1)\,n\,N!}{(N+1)!} = n,$$

이고 q의 거듭제곱은 $q^{N+n-(N+1)} = q^{n-1}$이다. 따라서 공식은 $np^{N+1}q^{n-1}$을 주며, 이는 올바르다. ($K \le N+1$이므로 $k \ge N+2$이면 확률은 0이다.)

불량품 개수의 기댓값

$k\,\mathrm{P}(K=k)$를 모두 더해도 되지만, X와 Y로 나타낸 K의 표현을 쓰는 편이 훨씬 빠르다. 사건 $X \le 1$의 지시함수(indicator)를 $\mathbf{1}_{\{X\le 1\}}$로 쓰면

$$K = (X+Y)\,\mathbf{1}_{\{X\le 1\}} = X\,\mathbf{1}_{\{X\le 1\}} + Y\,\mathbf{1}_{\{X\le 1\}}.$$

첫째 항 $X\,\mathbf{1}_{\{X\le 1\}}$은 $X=1$일 때 1이고 그 밖에는 0이므로, 그 기댓값(expectation)은 $\mathrm{P}(X=1) = npq^{n-1}$이다. 둘째 항에서 Y는 X와 독립이므로

$$\mathrm{E}[Y\,\mathbf{1}_{\{X\le 1\}}] = \mathrm{E}[Y]\,\mathrm{P}(X\le 1) = Np\left(q^n + npq^{n-1}\right).$$

두 항을 더하고 인수 pq^{n-1}을 묶어 내면

$$\begin{aligned}\mathrm{E}[K] &= npq^{n-1} + Np\,q^{n-1}(q+np)\\ &= pq^{n-1}\left[n + N(1-p) + Nnp\right]\\ &= \left[N + n + pN(n-1)\right]p(1-p)^{n-1},\end{aligned}$$

이고, 이것이 구하려던 식이다.

검산

- 확률의 합은 1이어야 한다. $n=2$, $N=1$일 때 $\mathrm{P}(K=1) = (2+1)pq^2 = 3pq^2$, $\mathrm{P}(K=2) = 2p^2q$, $\mathrm{P}(K=0) = q^3 + p^2$이고, ($p^2 = p^2(p+q)$를 이용하면) 이들의 합은 $q^3 + 3pq^2 + 3p^2q + p^3 = 1$이다.
- n이 매우 크면 표본에 불량품이 있는 묶음은 거의 모두 수리되므로 기댓값은 0에 가까워지며, 이는 인수 $(1-p)^{n-1}$과 들어맞는다.
- $2 \le n \le 6$, $0 \le N \le 6$과 여러 p 값에 대해 모든 공식을 정확한 분수로 확인했고, 절차를 모의실험(simulation)으로도 돌려 보았다. $n=4$, $N=6$, $p=0.3$일 때 공식은 $\mathrm{E}[K] \approx 1.585$를 주고, 200 000개의 모의 묶음에서는 1.581이 나왔다.

흔한 실수

- "두 개 이상" 규칙을 묶음 전체의 불량품 개수에 적용하는 것. 불량을 바로잡을지는 오직 표본의 개수 X만이 결정한다.
- $K=0$에는 수리된 묶음에서 오는 몫이 더 있다는 점을 잊어, 공식이 $k=0$에서 성립하지 않는다는 것을 놓치는 것.
- $\mathrm{E}[K]$를 무작정 합을 계산해서 구하는 것. 그래도 답은 나오지만, 지시함수로 K를 쪼개고 독립성을 쓰는 편이 훨씬 짧다.

STEP 너머

확률변수(random variable)를 지시함수를 곱한 조각들의 합으로 쓰고 기댓값의 선형성(linearity of expectation)을 이용하는 것은 대학 1학년 확률론에서 가장 쓸모 있는 기법 가운데 하나이다. 평균만 필요할 때 분포를 구하는 수고를 덜어 준다. 이 문제는 품질 관리의 합격 판정 샘플링(acceptance sampling)을 작게 본뜬 모형이기도 하다. 여기서는 검사 비용과 빠져나가는 불량품 사이의 균형을 맞추도록 n을 정한다. 더 알고 싶다면 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문), 또는 처음부터 끝까지 지시함수를 활용하는 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.