풀이

STEP II 2007, 문제 13

확률과 통계 • 생일 문제

> **STEP II 2007, 13번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/07-s2-q13

이 문제는 *Beyond STEP*의 9.9 절 **생일 문제**에서 단계별 안내와 함께 자세히 다룬다: https://beyondstep.keithmcnulty.org/ko/ch09/birthday.html

핵심 아이디어

어떤 두 손님도 생일이 같지 **않을** 확률(probability)은 $\frac{365-r}{365}$ 꼴의 인수들의 곱이고, 곱은 다루기 불편하다. 첫 번째 부분은 각 인수를 지수함수(exponential)로 바꾸어 주는데, 지수함수의 곱은 합의 지수함수이므로 그 합을 정확히 계산할 수 있다. 주인의 경우에는 손님들에 관한 사건이 서로 독립(independent)이므로 여사건(complement)의 확률이 하나의 거듭제곱이 되고, 같은 근사(approximation)를 적용할 수 있다.

풀이

근사식

임의의 실수 x에 대하여

$$\mathrm{e}^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots .$$

$x = r/n$으로 놓자. r이 n보다 훨씬 작으므로 x는 작고, x^2 이상의 항은 x에 비해 무시할 수 있다. 따라서

$$\mathrm{e}^{-r/n} \approx 1 - \frac{r}{n} = \frac{n-r}{n},$$

이므로 원하는 결과를 얻는다.

적어도 두 손님의 생일이 같은 경우

k명의 손님 가운데 적어도 두 명의 생일이 같은 사건을 A라 하자. k명의 생일이 모두 다를 확률인 $\mathrm{P}(A')$를 구하는 편이 더 쉽다. 손님을 한 명씩 차례로 생각한다. 첫 번째 손님의 생일은 어느 날이어도 된다. 처음 r명의 손님의 생일이 서로 다른 r개의 날이라는 조건 아래에서, $(r+1)$번째 손님은 그 r개의 날을 피해야 하고, 생일은 서로 독립이며 365일에 걸쳐 균등(uniform)하게 분포하므로 그 확률은 $\frac{365-r}{365}$이다. 이들을 곱하면

$$\mathrm{P}(A') = \prod_{r=0}^{k-1} \frac{365-r}{365}.$$

파티 규모가 크지 않다면 여기서 모든 r은 365보다 훨씬 작으므로, 첫 번째 부분에서 $n = 365$로 놓으면

$$\begin{aligned}\mathrm{P}(A') &\approx \prod_{r=0}^{k-1} \mathrm{e}^{-r/365} = \exp\Big(-\frac{1}{365}\sum_{r=0}^{k-1} r\Big) \\ &= \exp\Big(-\frac{1}{365}\cdot\frac{k(k-1)}{2}\Big) = \mathrm{e}^{-k(k-1)/730}.\end{aligned}$$

($r = 0$인 인수는 정확히 $1 = \mathrm{e}^0$이므로 포함해도 아무 손해가 없다.) 그러므로

$$\mathrm{P}(A) = 1 - \mathrm{P}(A') \approx 1 - \mathrm{e}^{-k(k-1)/730},$$

이고, 이것이 구하려던 결과이다.

> **생각해 보기** 많은 인수를 한꺼번에 근사하는 것이 걱정스럽지 않은가? 오차가 쌓일 수도 있다.

각 인수의 오차는 $(r/365)^2$ 정도의 크기이고, 지수 자체는 $k^2/730$ 정도의 크기이다. 우리에게 필요한 k의 값(약 20)에서 가장 큰 $r/365$는 약 0.06이고, 지수의 전체 상대 오차는 몇 퍼센트에 불과하다. 실제로 $k = 23$일 때 근사값은 0.500이고 참값은 0.507이다.

확률이 절반 이상이 되는 가장 작은 파티

필요한 조건은

$$1 - \mathrm{e}^{-k(k-1)/730} \geq \tfrac{1}{2} \iff \mathrm{e}^{-k(k-1)/730} \leq \tfrac{1}{2} \iff \frac{k(k-1)}{730} \geq \ln 2,$$

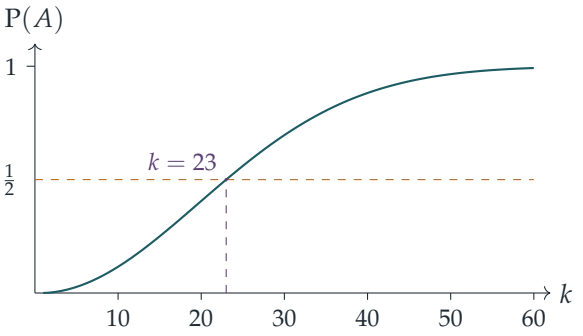
이다. 여기서 로그(logarithm)를 취하고(로그는 증가함수이므로 부등식의 방향이 유지된다) -1을 곱하였다(이때 부등식의 방향이 바뀐다). $\ln 2 \approx \frac{253}{365}$을 이용하면 조건은

$$k(k-1) \geq 730 \times \frac{253}{365} = 2 \times 253 = 506.$$

이 된다. $k \geq 1$일 때 $k(k-1)$은 k에 대하여 증가하고,

$$22 \times 21 = 462 < 506, \qquad 23 \times 22 = 506,$$

이므로 가장 작은 값은 $k = 23$이다.



주인과 생일이 같은 경우

이제 손님의 수를 m이라 하자(주인은 손님 수에 포함하지 않는다). 적어도 한 손님이 주인과 생일이 같은 사건을 B라 하자. 주인의 생일이 언제이든, 각 손님은 독립적으로 확률 $\frac{364}{365}$로 그 날을 비껴가므로

$$\mathrm{P}(B') = \left(\frac{364}{365}\right)^m \approx \left(\mathrm{e}^{-1/365}\right)^m = \mathrm{e}^{-m/365},$$

이다. 여기서 첫 번째 부분을 $n = 365$, $r = 1$로 놓고 사용하였다. 이번에는 m의 크기와 상관없이 $r/n = 1/365$이 정말로 작다.

$\mathrm{P}(B) \geq \frac{1}{2}$, 즉 $\mathrm{e}^{-m/365} \leq \frac{1}{2}$이어야 한다. 앞에서와 똑같이 하면 이는

$$\frac{m}{365} \geq \ln 2 \approx \frac{253}{365} \iff m \geq 253.$$

이다. 따라서 손님이 적어도 253명 있어야 한다.

? **생각해 보기** 왜 답이 23보다 훨씬 클까?

첫 번째 문제에서는 어떤 두 손님의 쌍이든 생일이 일치할 수 있고, 23명의 손님은 $\binom{23}{2} = 253$개의 쌍을 만든다. 두 번째 문제에서는 손님과 주인으로 이루어진 m개의 쌍만 고려된다. 일치할 기회의 수를 같게 하려면 $m = 253$이어야 하며, 바로 이 때문에 두 답 모두에 253이 나타난다.

검산

- 정확한 곱을 계산하면 두 답이 모두 확인된다. 생일이 같은 사람이 있을 참 확률은 손님이 22명일 때 0.4757, 23명일 때 0.5073이다. $1 - (364/365)^m$의 참값은 $m = 252$일 때 $\frac{1}{2}$보다 작고 $m = 253$일 때 0.5005이다.

- $\frac{253}{365} = 0.69315\ldots$이고 $\ln 2 = 0.69315\ldots$로 소수점 아래 다섯째 자리까지 일치하므로, 주어진 근사는 어느 답에도 영향을 주지 않는다.
- 23명이 모인 파티 200 000개를 시뮬레이션(simulation)한 결과, 50.7%의 경우에 생일이 같은 사람이 있었다.

흔한 실수

- **특정한** 두 손님의 생일이 같을 확률을 계산하거나, 각 쌍마다 $\frac{1}{365}$를 한 번씩 더하는 것. 서로 겹치는 사건의 확률은 단순히 더할 수 없으며, 여사건을 이용하면 이 문제를 피할 수 있다.
- 첫 번째 곱의 모든 인수에 $\frac{364}{365}$를 사용하는 것. $(r+1)$번째 손님은 하루가 아니라 r개의 날을 피해야 한다.
- 마지막 부분에서 주인을 손님 수에 포함하는 것. 주인은 별개의 사람이므로 손님이 253명 필요하다.

STEP 너머

여기서 얻는 경험적 법칙(heuristic)은, N개의 상자에 넣은 k개의 대상 가운데 일치가 일어날 가능성은 k가 대략 $\sqrt{2N\ln 2} \approx 1.18\sqrt{N}$일 때부터 높아진다는 것이다. 이는 암호학의 "생일 공격(birthday attack)"의 바탕이 되는데, 이 공격은 N비트 해시(hash)에서 2^N번이 아니라 약 $2^{N/2}$번의 시도만으로 충돌(collision)을 찾아낸다. 이 근사의 배후에는 푸아송 극한(Poisson limit)이 있다. 일치하는 쌍의 개수는 대략 평균이 $\binom{k}{2}/365$인 푸아송 분포를 따르고, $\mathrm{P}(\text{일치 없음}) \approx \mathrm{e}^{-k(k-1)/730}$이다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)에서 생일 문제와 그 푸아송 근사를 자세히 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.