풀이

STEP II 1991, 문제 16

확률과 통계 • 이항분포의 정규근사

STEP II 1991, 16번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/91-s2-q16

핵심 아이디어

부족량(deficit), 즉 수위가 댐 꼭대기보다 얼마나 아래에 있는지를 추적한다. 방류는 부족량을 일정하게 늘리고, 소나기는 한 번 올 때마다 부족량을 정확히 V만큼 줄인다. 부족량이 음수가 되면 물이 넘친다. 이틀 문제에서는 물이 넘칠 수 있는 강우 패턴이 단 하나뿐이고, 그 때에도 균등분포(uniform)를 따르는 시각 하나만 중요하다. 여름에는 방류가 멈추므로 모든 것이 소나기 횟수를 세는 문제로 귀결되며, 그 횟수는 이항분포(binomial)를 따른다.

풀이

소문항 (i): 이틀 휴가

휴가가 시작되는 자정부터 시간을 일 단위로 재고, 시각 t에서의 부족량을 $D(t)$라 하자. 그러면 $D(0) = V$이다. 소나기와 소나기 사이에 D는 비율 u로 증가하고, 소나기가 한 번 올 때마다 V만큼 감소한다. D가 한 번이라도 음수가 되면 댐이 넘친다. (문제의 의도대로, 저수지에는 이틀 내내 방류를 계속할 만큼 충분한 물이 있다고 가정한다.)

첫째 날과 둘째 날에 비가 오는 사건을 각각 R_1, R_2라 하고, 비가 오는 시각(하루에 대한 비율)을 T_1과 T_2라 하자. 이들은 각각 $[0,1]$에서 균등분포를 따른다. "어느 날이든 확률 p"라는 표현에 함축되어 있듯이, 각 날은 서로 독립(independent)이다.

- **비가 오지 않는 경우:** 부족량은 늘어나기만 하므로 넘치지 않는다.
- **하루만 비가 오는 경우**, 그 시각을 s라 하면: 소나기 직후 $D = V + us - V = us \geq 0$이므로 넘치지 않는다(확률이 0인 사건 $s = 0$은 제외한다).
- **이틀 모두 비가 오는 경우:** 첫 소나기 직후 $D = uT_1 \geq 0$이므로 안전하다. 두 번째 소나

기 직후, 즉 시각 $1+T_2$에서는

$$D = V + u(1+T_2) - 2V = u(1+T_2) - V.$$

T_1이 사라졌음에 주목하자. 첫 소나기가 언제 오든 방류는 같은 비율로 계속된다.

따라서 물이 넘치는 것은 이틀 모두 비가 오고 $u(1+T_2) < V$, 즉 $T_2 < \frac{V}{u} - 1$일 때이고, 그 때뿐이다. $\frac{1}{2}V \le u \le V$이면 수 $\frac{V}{u} - 1$은 $[0,1]$에 속하고, T_2가 균등분포를 따르므로

$$\mathrm{P}(\text{넘침}) = p^2\left(\frac{V}{u} - 1\right).$$

이것을 Q와 같다고 놓으면

$$\frac{V}{u} = 1 + \frac{Q}{p^2} \quad\Longrightarrow\quad u = \frac{p^2 V}{p^2 + Q}.$$

조건 $Q < p^2$ 덕분에 이 결과가 앞의 가정과 모순되지 않는다. $0 < Q/p^2 < 1$이므로 가정한 대로 $\frac{1}{2}V < u < V$가 된다. (이 조건이 없으면 해가 없다. 비 오는 날이 이틀 필요하므로 넘칠 확률은 결코 p^2를 넘을 수 없다.)

소문항 (i): 여름 휴가

이제 $u = 0$이므로 부족량은 결코 증가하지 않는다. 부족량은 kV에서 시작하여 소나기가 올 때마다 V씩 줄어든다. 따라서 18일 가운데 비 오는 날의 수를 N이라 하면

$$\text{넘침} \iff NV > kV \iff N > k.$$

각 날은 독립이고 각각 확률 $\frac{1}{3}$로 비가 오므로 $N \sim \mathrm{B}(18, \frac{1}{3})$이며, 평균(mean)은 6, 분산(variance)은 $18 \cdot \frac{1}{3} \cdot \frac{2}{3} = 4$이다.

생각해 보기 여기서 "적절한" 근사는 무엇이며, 왜 연속성 보정이 필요한가?

$np = 6, n(1-p) = 12$이므로 정규근사(normal approximation) $N \approx \mathrm{N}(6,4)$는 타당하다. ($p$가 작지 않으므로 푸아송 근사(Poisson approximation)는 적절하지 않다.) N은 정수 값을 가지므로 정수 k에 대하여 사건 $N > k$는 사건 $N \ge k+1$과 같고, 이를 $X \sim \mathrm{N}(6,4)$에 대한 $X > k + \frac{1}{2}$로 근사한다. 이것이 연속성 보정(continuity correction)이다. 필요한 조건은

$$\mathrm{P}\left(Z > \frac{k + \frac{1}{2} - 6}{2}\right) < \frac{1}{10} \iff \frac{k - 5.5}{2} > 1.2816 \iff k > 8.06.$$

구체적으로, $k = 8$이면 $\mathrm{P}(Z > 1.25) \approx 0.106$으로 너무 크고, $k = 9$이면 $\mathrm{P}(Z > 1.75) \approx 0.040$으로 충분하다. 넘침 여부는 $N > k$인지에만 달려 있으므로 8과 9 사이의 정수가 아닌 k는 $k = 9$보다 나을 것이 없다. 따라서 기술자는 저수지의 물을

$$k = 9,$$

즉 꼭대기에서 $9V$ 아래까지 빼야 한다.

소문항 (ii)

이제 반나절이 36개 있고, 이들은 독립이며 각각 확률 $\frac{1}{6}$로 비가 온다. 방류는 여전히 0이므로 넘침은 다시 $N > k$를 뜻하는데, 이번에는 $N \sim \mathrm{B}(36, \frac{1}{6})$이다. 평균은 여전히 6이지만 분산은 $36 \cdot \frac{1}{6} \cdot \frac{5}{6} = 5$로 4보다 크다.

평균이 같을 때 분산이 크면 위쪽 꼬리(upper tail)가 더 두껍다. 즉 평균보다 큰 각 k에 대하여 $\mathrm{P}(N > k)$가 이전보다 크다. 따라서 필요한 k는 **더 커진다**. 연속성 보정을 적용하여 $\mathrm{N}(6, 5)$로 근사하면 필요한 조건은

$$\frac{k - 5.5}{\sqrt{5}} > 1.2816 \iff k > 8.37,$$

이고, 이전에는 $k > 8.06$이었다. (여기서 정규근사는 $\mathrm{P}(N \geq 9) \approx 0.132$를 주며, 소문항 (i)에서는 0.106이었다.) 따라서 기준값이 올라간다. k를 정수로 올림하면 두 경우 모두 여전히 $k = 9$이다. 그러나 $k = 9$에서의 꼬리 확률이 0.04가 아니라 약 0.06이므로 안전 여유가 더 줄어든다.

검산

- $p = 0.5$, $Q = 0.1$, $u = \frac{0.25}{0.35}V \approx 0.714V$로 이틀 휴가를 몬테카를로 시뮬레이션(Monte Carlo simulation)한 결과(400,000회), 넘침의 상대도수는 0.0996이었다. 다른 p와 Q 값에서도 마찬가지로 잘 맞았다.
- 극한의 경우: $Q \to p^2$이면 $u \to \frac{1}{2}V$인데, 이는 비 오는 날 이틀로 여전히 넘칠 수 있는 가장 느린 방류 속도이다. $Q \to 0$이면 $u \to V$인데, 이 속도에서는 두 번째 소나기가 결코 넘침을 일으킬 수 없다.
- 정확한 이항분포 꼬리 확률이 정규근사를 뒷받침한다. $\mathrm{B}(18, \frac{1}{3})$에서는 $\mathrm{P}(N \geq 9) = 0.1076$, $\mathrm{P}(N \geq 10) = 0.0433$이다. $\mathrm{B}(36, \frac{1}{6})$에서는 $\mathrm{P}(N \geq 9) = 0.1332$, $\mathrm{P}(N \geq 10) = 0.0652$이다.

흔한 실수

- 이틀 문제에서 첫 비가 오는 시각 T_1이 중요하다고 생각하는 것. 방류가 일정하므로 이 값은 소거된다.
- 연속성 보정을 잊는 것. 보정을 하지 않으면 부등식이 $k > 8.56$이 되어 N이 정수라는 사실이 가려진다.
- 평균이 같다는 이유로 (ii)의 답이 "같다"고 말하는 것. 꼬리를 좌우하는 것은 퍼짐의 정도이다.

STEP 너머

무작위한 순간에 비가 내리고 물이 일정한 비율로 빠져나가는 상황은 응용확률론의 고전적인 모형으로, Moran이 댐을 위해 발전시킨 저장 이론(storage theory)이다. 이는 대기행렬(queue) 이론, 보험 파산(insurance ruin) 이론과 밀접하게 관련되어 있다. 소문항 (ii)에서 평균을 고정한 채 하루를 더 많고 더 작은 조각으로 나누면 $\mathrm{B}(n, 6/n)$은 평균 6, 분산 6인 푸아송 분포(Poisson distribution)로 수렴하고, 비가 오는 사건들은 푸아송 과정(Poisson process)이 된다. 더 알고 싶다면 Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문서)을 참고하라.

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.