풀이

STEP II 2001, 문제 13

확률과 통계 • 연속 분포와 추정

STEP II 2001, 13번 문제

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

step.maths.org/questions/01-s2-q13

핵심 아이디어

생존 함수(survival function) $\mathrm{P}(T > t)$가 직접 주어져 있으므로, 중앙값(median)은 이것을 $\frac{1}{2}$과 같게 놓아 구하고, 평균(mean)은 미분하여 확률밀도함수(probability density function)를 얻은 뒤 적분하여 구한다. 실험에서는 표본 중앙값(sample median)(크기순으로 늘어놓은 아홉 개의 수명 중 다섯 번째)과 표본평균(sample mean)이 주어지고, 이것을 공식과 맞추면 α와 k에 대한 두 방정식을 얻는다. 한 식을 다른 식으로 나누면 k가 사라지고 α만의 방정식이 남는다. 마지막 확률은 조건부 확률(conditional probability)이며, 그곳에서 반가운 놀라움이 기다리고 있다.

풀이

중앙값과 평균

$t \geq 0$에 대하여 $S(t) = \mathrm{P}(T > t) = (1+kt)^{-\alpha}$라 하자. 이 값이 1에서 0으로 감소하려면 $k > 0$이어야 한다.

중앙값. 중앙값 M은 $\mathrm{P}(T > M) = \frac{1}{2}$을 만족하므로

$$(1+kM)^{\alpha} = 2 \quad \Longrightarrow \quad 1 + kM = 2^{1/\alpha} \quad \Longrightarrow \quad M = \frac{2^{1/\alpha} - 1}{k}.$$

평균. 누적분포함수(cumulative distribution function)는 $F(t) = 1 - (1+kt)^{-\alpha}$이므로 확률밀도함수는

$$f(t) = F'(t) = \alpha k\,(1+kt)^{-\alpha-1}, \qquad t \geq 0.$$

이제 $u = t$, $\mathrm{d}v/\mathrm{d}t = \alpha k(1+kt)^{-\alpha-1}$로 놓고, 즉 $v = -(1+kt)^{-\alpha}$로 하여 부분적

분(integration by parts)하면

$$m = \int_0^\infty t\,\alpha k(1+kt)^{-\alpha-1}\,\mathrm{d}t = \Big[-t(1+kt)^{-\alpha}\Big]_0^\infty + \int_0^\infty (1+kt)^{-\alpha}\,\mathrm{d}t$$
$$= 0 + \left[-\frac{(1+kt)^{1-\alpha}}{k(\alpha-1)}\right]_0^\infty = \frac{1}{k(\alpha-1)}.$$

무한대에서의 두 극한은 모두 $\alpha > 1$을 사용한다. $t(1+kt)^{-\alpha}$는 $t^{1-\alpha}/k^\alpha \to 0$처럼 행동하고, $(1+kt)^{1-\alpha} \to 0$이다. ($\alpha \le 1$이면 평균이 무한대가 되며, 문제에서 $\alpha > 1$을 굳이 요구하는 것은 이 때문이다.) 따라서

$$M = \frac{2^{1/\alpha}-1}{k}, \qquad m = \frac{1}{k(\alpha-1)}.$$

α와 k의 추정

전구가 아홉 개이므로, 다섯 번째 고장은 크기순으로 늘어놓은 아홉 개의 수명 중 한가운데 값이다. 즉 이것이 **표본 중앙값**이다. 따라서 $M \approx 1000$으로 추정하고, 표본평균으로부터 $m \approx 2400$으로 추정한다.

> **생각해 보기** 방정식이 두 개, 미지수가 두 개이지만 k가 두 식 모두에 들어 있다. 어떻게 하면 α만 따로 얻을 수 있을까?

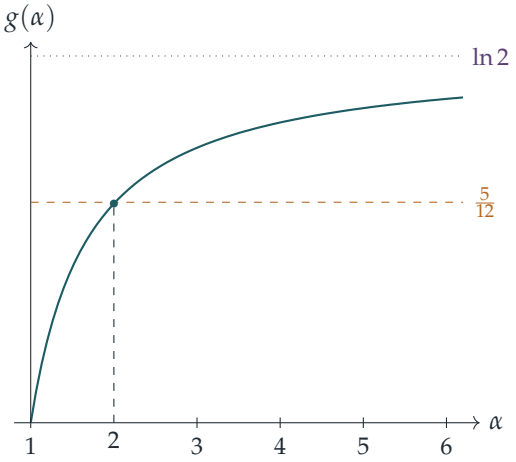
나누면 된다. M과 m이 모두 $1/k$에 비례하므로 k가 약분된다. 그러면

$$\frac{M}{m} = (2^{1/\alpha}-1)(\alpha-1) \approx \frac{1000}{2400} = \frac{5}{12} \approx 0.4167.$$

좌변을 $g(\alpha)$라 하자. $\alpha = 2$에서

$$g(2) = (\sqrt{2}-1)(2-1) = \sqrt{2}-1 \approx 0.4142,$$

이며, 이는 $5/12$과의 차이가 0.003 이내이다. 다른 값이 이보다 잘 맞지 않음을 보이기 위해 $g(1.9) \approx 0.396$, $g(2.1) \approx 0.430$임에 주목하면, 값 $5/12$은 이 둘 사이에 있다. 실제로 g는 $g(1) = 0$에서 시작하여 $\alpha \to \infty$일 때 $\ln 2 \approx 0.693$을 향해 꾸준히 증가하므로(그래프 참고), 값 $5/12$을 2보다 조금 큰 α에서 단 한 번만 가진다. 따라서 $\alpha \approx 2$이다.



이제 $\alpha = 2$를 어느 한 공식에 다시 대입한다. 평균으로부터

$$k = \frac{1}{m(\alpha - 1)} \approx \frac{1}{2400},$$

이고, 중앙값으로부터 $k = (\sqrt{2} - 1)/1000 \approx 4.14 \times 10^{-4}$이다. 두 값은 거의 일치하며,

$$k \approx 4.2 \times 10^{-4} \text{ (시간당)}.$$

조건부 확률

M시간 동안 작동한 뒤 다시 $m - M$시간 더 작동한 전구는 모두 합쳐 m시간 동안 작동한 것이다. $m > M$이므로 사건 $\{T > m\}$은 $\{T > M\}$에 포함되고, 따라서

$$\mathrm{P}(T > m \mid T > M) = \frac{\mathrm{P}(T > m)}{\mathrm{P}(T > M)} = \frac{(1 + km)^{-\alpha}}{1/2} = \frac{2}{(1 + km)^{\alpha}}.$$

(엄밀히 말하면 전구가 M시간 시점에 아직 작동하고 있음이 확인된 것이다. 연속적인 수명에서는 $T \geq M$과 $T > M$의 확률이 같다.) 이제 놀라운 점이 나온다. k가 무엇이든 정확히 $km = 1/(\alpha - 1)$이므로 $1 + km = \alpha/(\alpha - 1)$이고

$$\mathrm{P}(T > m \mid T > M) = 2\left(\frac{\alpha - 1}{\alpha}\right)^{\alpha}.$$

$\alpha = 2$이면 이 값은 $2 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{2}$이다. 따라서 추정 확률은

$$\mathrm{P}(\text{앞으로} m - M\text{시간 더 작동함}) \approx \tfrac{1}{2}.$$

이 마지막 단계에서는 k의 추정값이 전혀 필요하지 않다.

검산

- 중앙값에서 $S(M) = (2^{1/\alpha})^{-\alpha} = \frac{1}{2}$이며, 당연히 그래야 한다. 임의의 $\alpha > 1$과 $k > 0$에 대하여 $S(t)$를 $[0, \infty)$에서 수치적분하면 $m = 1/(k(\alpha - 1))$이 재현된다(음이 아닌 확률변수에 대하여 $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$임을 떠올리자. 이것이 바로 풀이의 두 번째 적분이다).
- $g(\alpha) = 5/12$을 정확히 풀면 $\alpha \approx 2.015$이고, 이때 조건부 확률은 $2(1 - 1/\alpha)^\alpha \approx 0.502$이므로, $\alpha = 2$로 반올림해도 손해가 거의 없다.
- $\alpha = 2$, $k = 1/2400$으로 수명 400 000개를 시뮬레이션한 결과 ($(0,1)$에서 균등분포를 따르는 U에 대하여 $T = (U^{-1/2} - 1)/k$로 놓음) $\mathrm{P}(T > m \mid T > M) \approx 0.501$을 얻었다.

흔한 실수

- $\mathrm{P}(T > t)$를 누적분포함수로 착각하고 음의 부호 없이 미분하여 음수인 "밀도"를 얻는 것.
- 전구가 시각 M에서 새로 시작하는 것처럼 $\mathrm{P}(T > m - M)$을 답하는 것. 이 분포는 비기억성(memoryless)을 갖지 않으므로(그런 분포는 지수분포(exponential distribution)뿐이다), 조건 $T > M$이 중요하다.
- 다섯 번째 고장을 중앙값이 아니라 평균의 추정값으로 사용하는 것.

STEP 너머

이것은 파레토형(Pareto-type) 분포, 곧 로맥스(Lomax) 분포로, 수명과 소득에 대한 표준적인 모형이며 두꺼운 꼬리(heavy tail)를 가진다. $\mathrm{E}(T^r)$은 $r < \alpha$일 때에만 유한하다. 여기서 한 것은 **적률법**(method of moments)이다(엄밀히는 분위수(quantile) 하나와 적률(moment) 하나를 맞춘 것이다). 즉 모형의 요약 통계량이 표본의 요약 통계량과 같아지도록 모수를 고른다. 대학 1학년 통계학에서는 이것을 최대가능도(maximum likelihood)와 비교하고, 전구 아홉 개만으로 얻은 이런 추정값이 얼마나 정확한지를 묻는다. 비율 $\mathrm{P}(T > s + t \mid T > s)$는 신뢰성 이론(reliability theory)에서 쓰이는 생존 함수이다. 여기서는 이 값이 s에 따라 **증가**하므로, 오래된 전구가 새 전구보다 계속 작동할 가능성이 더 크다. Grimmett와 Welsh의 *Probability: An Introduction*(확률론 입문)이 다음 단계로 좋은 책이다.

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.