解答例

STEP II 2001、第 13 問

確率・統計 ・ 連続分布と推定

> **STEP II 2001、問 13**
>
> 先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：
>
> step.maths.org/questions/01-s2-q13

鍵となる考え方

生存関数（survival function）$\mathrm{P}(T > t)$ が直接与えられているので、中央値（median）はこれを $\frac{1}{2}$ に等しいとおけば求まり、平均（mean）は微分して密度を求めてから積分すれば得られる。実験からは標本中央値（sample median）（小さい順に並べた 9 個の寿命のうち 5 番目）と標本平均（sample mean）が得られ、これらを公式に当てはめると α と k についての 2 つの方程式が得られる。一方を他方で割ると k が消え、α だけの方程式が残る。最後の確率は条件付き確率（conditional probability）であり、そこには嬉しい驚きが待っている。

解答

中央値と平均

$t \geq 0$ に対して $S(t) = \mathrm{P}(T > t) = (1+kt)^{-\alpha}$ とおく。これが 1 から 0 へ減少するためには $k > 0$ でなければならない。

中央値。 中央値 M は $\mathrm{P}(T > M) = \frac{1}{2}$ を満たすので、

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

平均。 累積分布関数（cumulative distribution function）は $F(t) = 1 - (1+kt)^{-\alpha}$ であるから、確率密度関数（probability density function）は

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

である。そこで、$u = t$、$\mathrm{d}v/\mathrm{d}t = \alpha k(1+kt)^{-\alpha-1}$ とし、したがって $v = -(1+kt)^{-\alpha}$

として部分積分（integration by parts）を行うと、

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

無限大での 2 つの極限はいずれも $\alpha > 1$ を用いている。すなわち、$t(1+kt)^{-\alpha}$ は $t^{1-\alpha}/k^\alpha \to 0$ のように振る舞い、また $(1+kt)^{1-\alpha} \to 0$ である。（$\alpha \le 1$ ならば平均は無限大になる。問題が $\alpha > 1$ を課しているのはこのためである。）よって

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

α と k の推定

電球が 9 個あるとき、5 番目の故障は、小さい順に並べた 9 個の寿命のちょうど真ん中にあたる。つまりこれは**標本中央値**である。そこで $M \approx 1000$ と推定し、標本平均から $m \approx 2400$ と推定する。

考えてみよう 方程式が 2 つ、未知数が 2 つあるが、k はその両方に現れる。どうすれば α だけを取り出せるだろうか。

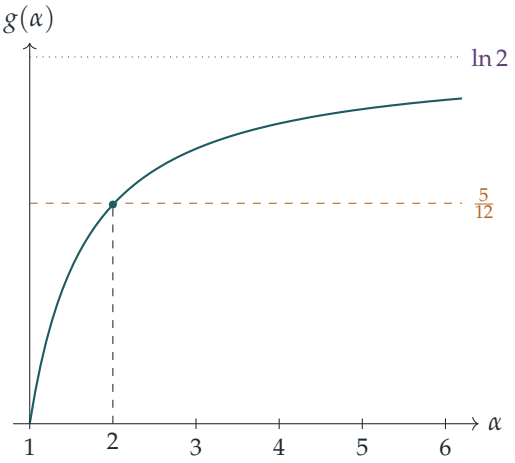
割り算をすればよい。M も m も $1/k$ に比例するので、k は約分されて消える。すると

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

を得る。左辺を $g(\alpha)$ とおく。$\alpha = 2$ のとき

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

であり、これは 5/12 との差が 0.003 以内である。これ以外の値がこれほどよく合うことはないことを見るために、$g(1.9) \approx 0.396$、$g(2.1) \approx 0.430$ であることに注意すると、値 5/12 はこの 2 つの間にある。実際、g は $g(1) = 0$ から、$\alpha \to \infty$ のとき $\ln 2 \approx 0.693$ に向かって単調に増加する（グラフを参照）。したがって g が値 5/12 をとるのは、2 をわずかに上回る α でのただ一度だけである。ゆえに $\alpha \approx 2$ である。



ここで $\alpha = 2$ をどちらかの公式に代入する。平均の式からは

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

となり、中央値の式からは $k = (\sqrt{2} - 1)/1000 \approx 4.14 \times 10^{-4}$ となる。両者はよく一致しており、

$$k \approx 4.2 \times 10^{-4} (1\text{時間あたり}).$$

である。

条件付き確率

M 時間もち、さらに $m - M$ 時間もった電球は、合計で m 時間もったことになる。$m > M$ であるから、事象 $\{T > m\}$ は $\{T > M\}$ に含まれ、したがって

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

である。（厳密には、電球は M 時間の時点でまだ点灯していることが確認されたのである。寿命が連続型であれば、$T \geq M$ と $T > M$ の確率は等しい。）ここで驚くべきことに、k が何であっても $km = 1/(\alpha - 1)$ がちょうど成り立つので、$1 + km = \alpha/(\alpha - 1)$ となり、

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

を得る。$\alpha = 2$ のとき、これは $2 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{2}$ である。よって求める確率の推定値は

$$\mathrm{P}(\text{さらに} m - M \text{時間もつ}) \approx \tfrac{1}{2}.$$

である。この最後の段階では、k の推定値はまったく必要ない。

検算

- 中央値において $S(M) = (2^{1/\alpha})^{-\alpha} = \frac{1}{2}$ であり、確かにそうなっている。ランダムに選んだ $\alpha > 1$ と $k > 0$ について $S(t)$ を $[0, \infty)$ 上で数値積分すると、$m = 1/(k(\alpha - 1))$ が再現される（非負の確率変数については $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$ が成り立つことを思い出そう。これはまさに上の計算に現れた 2 つ目の積分である）。
- $g(\alpha) = 5/12$ を正確に解くと $\alpha \approx 2.015$ となり、このとき条件付き確率は $2(1 - 1/\alpha)^\alpha \approx 0.502$ である。したがって $\alpha = 2$ に丸めても、ほとんど何も失われない。
- $\alpha = 2$、$k = 1/2400$ として 400 000 個の寿命をシミュレーションしたところ（U を $(0,1)$ 上の一様分布として $T = (U^{-1/2} - 1)/k$ とした）、$\mathrm{P}(T > m \mid T > M) \approx 0.501$ を得た。

よくある間違い

- $\mathrm{P}(T > t)$ を累積分布関数とみなし、マイナス符号をつけずに微分してしまい、負の「密度」を得てしまう。
- 電球が時刻 M で新品に戻ったかのように、$\mathrm{P}(T > m - M)$ を答えてしまう。この分布は無記憶性（memoryless）をもたない（無記憶性をもつのは指数分布（exponential distribution）だけである）ので、条件 $T > M$ が効いてくる。
- 5 番目の故障を、中央値ではなく平均の推定値として使ってしまう。

STEP のその先へ

これはパレート（Pareto）型分布（ロマックス分布（Lomax distribution））であり、寿命や所得の標準的なモデルである。この分布は裾が重く（heavy tail）、$\mathrm{E}(T^r)$ が有限になるのは $r < \alpha$ のときに限られる。ここで行ったのは**積率法**（method of moments）である（厳密には、分位点（quantile）1 つとモーメント（moment）1 つを一致させている）。すなわち、モデルの要約統計量が標本の要約統計量に等しくなるようにパラメータを選ぶのである。大学 1 年次の統計学では、これを最尤法（maximum likelihood）と比較し、わずか 9 個の電球からの推定がどれほど正確なのかを問う。比 $\mathrm{P}(T > s + t \mid T > s)$ は信頼性理論（reliability theory）で用いられる生存関数である。ここではこれが s とともに**増加**するので、古い電球のほうが新しい電球よりも点灯し続ける可能性が高い。次の一歩としては、Grimmett and Welsh, *Probability: An Introduction*（確率論入門）がよい。

Beyond STEP 解答集：https://beyondstep.keithmcnulty.org/ja/。解答本文は CC0 1.0 である。問題そのものは © Cambridge University Press & Assessment に帰属し、ここには掲載していない。英語版から機械翻訳したものであり、まだ校閲を経ていない。