解答例

STEP III 1993、第 16 問

確率・統計 ・ 切断正規分布

STEP III 1993、問 16

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/93-s3-q16

鍵となる考え方

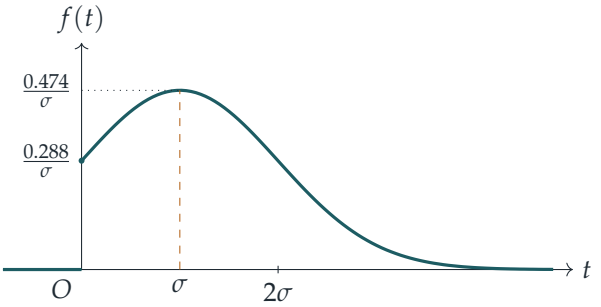
因子 $1/k$ を除けば、この確率密度関数（probability density function）は、平均（mean）σ、標準偏差（standard deviation）σ の正規分布（normal distribution）の密度を $t = 0$ で切断したものにほかならない。したがって T に関するあらゆる確率は、正規分布の確率を、切断する前の確率変数が非負となる確率で割ったものになる。$z = (t - \sigma)/\sigma$ と標準化（standardise）すれば、すべてが $Z \sim N(0,1)$ に関する主張と Φ の表の問題に帰着する。

解答

以下を通して、$Z \sim N(0,1)$ とし、その密度を $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$、分布関数（distribution function）を Φ とする。

概形

$t \geq 0$ では、グラフは $t = \sigma$ を中心とする正規分布の釣鐘型の曲線を $1/k$ 倍に拡大したものである。最大値は $t = \sigma$ でとり、そこで $f(\sigma) = \frac{1}{k\sigma\sqrt{2\pi}} \approx \frac{0.474}{\sigma}$ であり、$t = 2\sigma$ に変曲点（point of inflexion）をもつ。切断点では曲線は正の高さ $f(0) = \frac{e^{-1/2}}{k\sigma\sqrt{2\pi}} \approx \frac{0.288}{\sigma}$ から始まり、$t < 0$ では 0 に跳ぶ。$t \to \infty$ のとき $f(t) \to 0$ である。



k の値

全確率は 1 でなければならない：

$$1 = \int_0^\infty \frac{1}{k\sigma\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(\tfrac{t-\sigma}{\sigma}\right)^2\right) \mathrm{d}t.$$

$z = (t-\sigma)/\sigma$ と置換すると、$\mathrm{d}t = \sigma\,\mathrm{d}z$ であり、$t = 0$ は $z = -1$ に対応するから、

$$1 = \frac{1}{k}\int_{-1}^\infty \phi(z)\,\mathrm{d}z = \frac{1-\Phi(-1)}{k} = \frac{\Phi(1)}{k}.$$

よって $k = \Phi(1) \approx 0.8413$ であり、求めるとおり $k \approx 0.841$ である。

平均

同じ置換により $t = \sigma(1+z)$ であるから、

$$\mathrm{E}(T) = \frac{1}{k}\int_{-1}^\infty \sigma(1+z)\,\phi(z)\,\mathrm{d}z = \sigma\left[\frac{1}{k}\int_{-1}^\infty \phi(z)\,\mathrm{d}z + \frac{1}{k}\int_{-1}^\infty z\,\phi(z)\,\mathrm{d}z\right].$$

最初の積分は、いま示したとおり k である。

> **考えてみよう** $\phi(z)$ 自体は積分できないのに、$z\,\phi(z)$ はどうすれば積分できるだろうか。

$\phi'(z) = -z\,\phi(z)$ という事実を用いる。すなわち $z\,\phi(z)$ の原始関数（antiderivative）は $-\phi(z)$ である：

$$\int_{-1}^\infty z\,\phi(z)\,\mathrm{d}z = \Big[-\phi(z)\Big]_{-1}^\infty = \phi(-1) = \frac{e^{-1/2}}{\sqrt{2\pi}} \approx 0.24197.$$

したがって

$$\mathrm{E}(T) = \sigma\left(1 + \frac{e^{-1/2}}{k\sqrt{2\pi}}\right) \approx \sigma\left(1 + \frac{0.24197}{0.84134}\right) \approx 1.288\,\sigma.$$

これは当然 σ より大きい。左側の裾を取り除くと、平均は右へ押しやられるからである。

σ を求める

問題作成時間の平均は 3 時間であるから、

$$\sigma = \frac{3}{1.2876} \approx 2.330 \text{ 時間}.$$

任意の $a \geq 0$ に対して、同じ置換により

$$\mathrm{P}(T > a) = \frac{1}{k}\int_{(a-\sigma)/\sigma}^{\infty} \phi(z)\,\mathrm{d}z = \frac{1-\Phi\left(\frac{a}{\sigma}-1\right)}{\Phi(1)}. \qquad (*)$$

小問 (i)

$a = 4$ とすると、$\frac{4}{\sigma} - 1 \approx \frac{4}{2.330} - 1 \approx 0.717$ であり、$\Phi(0.717) \approx 0.7633$ である。$(*)$ より

$$\mathrm{P}(T > 4) \approx \frac{1 - 0.7633}{0.8413} \approx 0.2814, \qquad \mathrm{P}(T \leq 4) \approx 0.7186.$$

ここで、16 問の作成時間は互いに独立（independent）で、それぞれ T と同じ分布に従うと仮定する（利用できる唯一の妥当なモデルである）。すると

$$\mathrm{P}(4 \text{ 時間を超える問題が } 1 \text{ 問もない}) = \mathrm{P}(T \leq 4)^{16} \approx 0.7186^{16} \approx 0.0051.$$

したがって、すべての問題が 4 時間以内に作成される可能性は非常に低い（およそ 200 回に 1 回）。

小問 (ii)

考えてみよう どの事象を条件としているのだろうか。

「2 時間後に満足のいくものでない」とは、その問題の作成に合計で 2 時間を超える時間が必要だということであり、事象 $T > 2$ である。「さらに 2 時間後もまだ受け入れられない」は $T > 4$ である。$\{T > 4\} \subseteq \{T > 2\}$ であるから、条件付き確率（conditional probability）は

$$\mathrm{P}(T > 4 \mid T > 2) = \frac{\mathrm{P}(T > 4)}{\mathrm{P}(T > 2)}.$$

$a = 2$ のとき $\frac{2}{\sigma} - 1 \approx -0.142$ であるから、$1 - \Phi(-0.142) = \Phi(0.142) \approx 0.5563$ である。$(*)$ の二つの式の比をとると、因子 $\Phi(1)$ は約分される：

$$\mathrm{P}(T > 4 \mid T > 2) \approx \frac{1 - 0.7633}{0.5563} = \frac{0.2367}{0.5563} \approx 0.426.$$

検算

- 密度を直接数値積分すると（シンプソンの公式（Simpson's rule）、$\sigma = 1$）、k で割る前の全面積は 0.84134、平均は 1.2876σ となる。

- 平均と標準偏差がともに $\sigma = 2.330$ の正規乱数を 4×10^5 個生成し、負のものを捨てるシミュレーションを行うと、標本平均は 3.004、$\mathrm{P}(T > 4) \approx 0.282$、$\mathrm{P}(T > 4 \mid T > 2) \approx 0.426$ となる。
- 小問 (i) は丸めの影響を非常に受けやすい。$0.719^{16} \approx 0.0051$ であるが、$0.718^{16} \approx 0.0050$ である。答えは約 0.005 と述べておくのがよい。
- 切断がなくても、$\mathrm{P}(T > 4 \mid T > 2)$ は同じ値になる。因子 $1/k$ が約分されるからである。k の影響を受けるのは小問 (i) と平均だけである。

よくある間違い

- $\mathrm{P}(T > 4)$ で k で割るのを忘れること。この場合 0.281 ではなく 0.237 となり、(i) では $0.7633^{16} \approx 0.013$ と、正しい値の 2 倍以上になってしまう。
- T の平均を σ（切断前の平均）としてしまい、$\sigma = 3$ とすること。
- (ii) で、条件付き確率ではなく $\mathrm{P}(2 < T \leq 4)$ や $\mathrm{P}(T > 2)$ を求めてしまうこと。

STEP のその先へ

平均に現れる量 $\phi(1)/\Phi(1)$ は、**逆ミルズ比**（inverse Mills ratio）の一例である。一般に、正規分布 $\mathrm{N}(\mu, \sigma^2)$ を a より下で切断したものの平均は、$\alpha = (a - \mu)/\sigma$ として $\mu + \sigma\,\phi(\alpha)/(1 - \Phi(\alpha))$ である。切断分布（truncated distribution）は、データがある閾値より上でしか観測できない場面ならどこにでも現れる。たとえば生存時間解析（survival analysis）や、打ち切りデータ（censored data）の計量経済モデルがそうである。小問 (ii) は**生存関数**（survival function）$\mathrm{P}(T > t)$ に関するものである。$\mathrm{P}(T > s + t \mid T > s)$ が s によらないのは、指数分布（exponential distribution）の場合に限られる。Grimmett and Welsh, *Probability: An Introduction*（確率論入門）は、連続分布と条件付けを扱っている。

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。