詳細解答

STEP III 1993，第 16 題

概率與統計 • 截尾正態分佈

> **STEP III 1993，第 16 題**
>
> 繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：
>
> step.maths.org/questions/93-s3-q16

關鍵思路

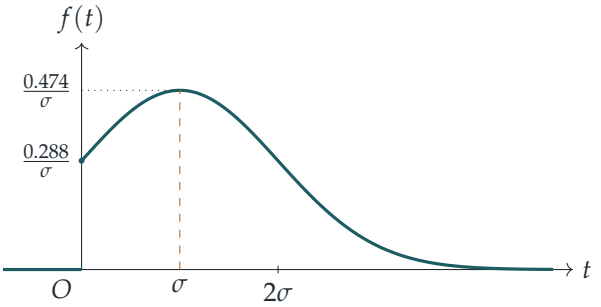
除了因子 $1/k$ 之外，這個密度函數正是平均值為 σ、標準差為 σ 的正態分佈（normal distribution）的密度函數，只是在 $t = 0$ 處被截斷。因此，凡是關於 T 的概率，都等於一個正態概率除以「未截斷的變量為非負」的概率。以 $z = (t - \sigma)/\sigma$ 作標準化（standardising），一切便化為關於 $Z \sim \mathrm{N}(0,1)$ 的命題，可查 Φ 的數值表求得。

解答

以下一律設 $Z \sim \mathrm{N}(0,1)$，其密度函數（density）為 $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$，分佈函數（distribution function）為 Φ。

略圖

當 $t \geq 0$ 時，圖像是以 $t = \sigma$ 為中心的正態鐘形曲線，再按 $1/k$ 的比例放大。它在 $t = \sigma$ 處取得極大值，$f(\sigma) = \frac{1}{k\sigma\sqrt{2\pi}} \approx \frac{0.474}{\sigma}$，並在 $t = 2\sigma$ 處有一個拐點（point of inflexion）。在截斷處，曲線由一個正的高度開始，$f(0) = \frac{e^{-1/2}}{k\sigma\sqrt{2\pi}} \approx \frac{0.288}{\sigma}$，而當 $t < 0$ 時則跳至 0。當 $t \to \infty$ 時，$f(t) \to 0$。



k 的值

總概率必須等於 1：

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

代入 $z = (t-\sigma)/\sigma$，則 $\mathrm{d}t = \sigma\,\mathrm{d}z$，而 $t = 0$ 變成 $z = -1$，於是

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

因此 $k = \Phi(1) \approx 0.8413$，即 $k \approx 0.841$，正如題目所求。

平均值

用同一個代換，$t = \sigma(1+z)$，於是

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

第一個積分等於 k，這是剛才已經證明的。

> **想一想** 既然 $\phi(z)$ 本身也積不出來，我們怎樣對 $z\,\phi(z)$ 求積分？

我們利用 $\phi'(z) = -z\,\phi(z)$ 這個事實，所以 $z\,\phi(z)$ 的原函數（antiderivative）是 $-\phi(z)$：

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

因此

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

這個值比 σ 大，也理應如此：除去左尾之後，平均值會向右移。

求 σ

擬題時間的平均值是 3 小時，所以

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

對任意 $a \geq 0$，用同一個代換可得

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

第 (i) 部分

取 $a = 4$：$\frac{4}{\sigma} - 1 \approx \frac{4}{2.330} - 1 \approx 0.717$，而 $\Phi(0.717) \approx 0.7633$。由 $(*)$，

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

現在我們假設 16 道題目的擬題時間互相獨立（independent），而且各自服從 T 的分佈（這是唯一合理而可用的模型）。那麼

$$\mathrm{P}(\text{沒有題目需時超過 } 4 \text{ 小時}) = \mathrm{P}(T \leq 4)^{16} \approx 0.7186^{16} \approx 0.0051.$$

所以，每道題目都能在 4 小時內擬好的機會極微（大約二百分之一）。

第 (ii) 部分

> ? **想一想** 我們是以哪一個事件作為條件？

「2 小時後仍未令人滿意」表示該題目總共需要多於 2 小時，所以這是事件 $T > 2$。「再過 2 小時後仍然不獲接納」則是 $T > 4$。由於 $\{T > 4\} \subseteq \{T > 2\}$，

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

取 $a = 2$：$\frac{2}{\sigma} - 1 \approx -0.142$，所以 $1 - \Phi(-0.142) = \Phi(0.142) \approx 0.5563$。把兩個 $(*)$ 式相除時，因子 $\Phi(1)$ 互相抵消：

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

驗算

- 直接對密度函數作數值積分（辛普森法則（Simpson's rule），取 $\sigma = 1$），在除以 k 之前得總面積為 0.84134，平均值為 1.2876σ。
- 模擬 4×10^5 個平均值和標準差均為 $\sigma = 2.330$ 的正態值，並捨去其中的負值，得樣本平均值 3.004，$\mathrm{P}(T > 4) \approx 0.282$ 及 $\mathrm{P}(T > 4 \mid T > 2) \approx 0.426$。
- 第 (i) 部分對捨入非常敏感：$0.719^{16} \approx 0.0051$，但 $0.718^{16} \approx 0.0050$。答案寫作約 0.005 即可。
- 即使沒有截斷，$\mathrm{P}(T > 4 \mid T > 2)$ 也會相同，因為因子 $1/k$ 會抵消。只有第 (i) 部分和平均值才受 k 影響。

常見錯誤

- 計算 $\mathrm{P}(T > 4)$ 時忘記除以 k：這樣會得到 0.237 而不是 0.281，並在 (i) 中得到 $0.7633^{16} \approx 0.013$，是正確值的兩倍有多。
- 把 T 的平均值當作 σ（即截斷前的平均值），這樣會得到 $\sigma = 3$。
- 在 (ii) 中求了 $\mathrm{P}(2 < T \leq 4)$ 或 $\mathrm{P}(T > 2)$，而不是條件概率（conditional probability）。

STEP 之外

平均值中出現的量 $\phi(1)/\Phi(1)$ 是*逆米爾斯比率*（inverse Mills ratio）的一個例子：一般而言，正態分佈 $\mathrm{N}(\mu, \sigma^2)$ 在 a 處截去下方之後，其平均值為 $\mu + \sigma\phi(\alpha)/(1 - \Phi(\alpha))$，其中 $\alpha = (a - \mu)/\sigma$。凡是數據只能在某個門檻以上才觀察得到的場合，都會出現截斷分佈（truncated distribution），例如存活分析（survival analysis）以及處理設限數據（censored data）的計量經濟模型。第 (ii) 部分所討論的是*存活函數*（survival function）$\mathrm{P}(T > t)$；指數分佈（exponential distribution）是唯一能令 $\mathrm{P}(T > s + t \mid T > s)$ 與 s 無關的情況。Grimmett 與 Welsh 的 *Probability: An Introduction*（概率論入門教材）涵蓋連續分佈和條件概率。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。