详细解答

STEP III 1993，第 16 题

概率与统计 · 截尾正态分布

> **STEP III 1993，第 16 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/93-s3-q16

关键思路

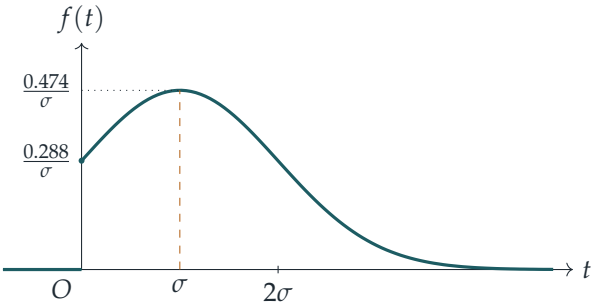
除了因子 $1/k$ 之外，这个密度函数正是平均值为 σ、标准差为 σ 的正态分布（normal distribution）的密度函数，只是在 $t = 0$ 处被截断。因此，凡是关于 T 的概率，都等于一个正态概率除以“未截断的变量为非负”的概率。以 $z = (t - \sigma)/\sigma$ 作标准化（standardising），一切便化为关于 $Z \sim \mathrm{N}(0,1)$ 的命题，可查 Φ 的数值表求得。

解答

以下一律设 $Z \sim \mathrm{N}(0,1)$，其密度函数（density）为 $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$，分布函数（distribution function）为 Φ。

略图

当 $t \geq 0$ 时，图像是以 $t = \sigma$ 为中心的正态钟形曲线，再按 $1/k$ 的比例放大。它在 $t = \sigma$ 处取得极大值，$f(\sigma) = \frac{1}{k\sigma\sqrt{2\pi}} \approx \frac{0.474}{\sigma}$，并在 $t = 2\sigma$ 处有一个拐点（point of inflexion）。在截断处，曲线由一个正的高度开始，$f(0) = \frac{e^{-1/2}}{k\sigma\sqrt{2\pi}} \approx \frac{0.288}{\sigma}$，而当 $t < 0$ 时则跳至 0。当 $t \to \infty$ 时，$f(t) \to 0$。



k 的值

总概率必须等于 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 $z = (t-\sigma)/\sigma$，则 $\mathrm{d}t = \sigma\,\mathrm{d}z$，而 $t = 0$ 变成 $z = -1$，于是

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

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

平均值

用同一个代换，$t = \sigma(1+z)$，于是

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

第一个积分等于 k，这是刚才已经证明的。

> ? **想一想** 既然 $\phi(z)$ 本身也积不出来，我们怎样对 $z\,\phi(z)$ 求积分?

我们利用 $\phi'(z) = -z\,\phi(z)$ 这个事实，所以 $z\,\phi(z)$ 的原函数（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

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

这个值比 σ 大，也理应如此：除去左尾之后，平均值会向右移。

求 σ

拟题时间的平均值是 3 小时，所以

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

对任意 $a \geq 0$，用同一个代换可得

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

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 $(*)$，

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

现在我们假设 16 道题目的拟题时间互相独立（independent），而且各自服从 T 的分布（这是唯一合理而可用的模型）。那么

$$\mathrm{P}(\text{没有题目需时超过 } 4 \text{ 小时}) = \mathrm{P}(T \leq 4)^{16} \approx 0.7186^{16} \approx 0.0051.$$

所以，每道题目都能在 4 小时内拟好的机会极微（大约二百分之一）。

第 (ii) 部分

> ? **想一想** 我们是以哪一个事件作为条件?

“2 小时后仍未令人满意”表示该题目总共需要多于 2 小时，所以这是事件 $T > 2$。“再过 2 小时后仍然不获接纳”则是 $T > 4$。由于 $\{T > 4\} \subseteq \{T > 2\}$，

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

取 $a = 2$：$\frac{2}{\sigma} - 1 \approx -0.142$，所以 $1 - \Phi(-0.142) = \Phi(0.142) \approx 0.5563$。把两个 $(*)$ 式相除时，因子 $\Phi(1)$ 互相抵消：

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

验算

- 直接对密度函数作数值积分（辛普森法则（Simpson's rule），取 $\sigma = 1$），在除以 k 之前得总面积为 0.84134，平均值为 1.2876σ。
- 模拟 4×10^5 个平均值和标准差均为 $\sigma = 2.330$ 的正态值，并舍去其中的负值，得样本平均值 3.004，$\mathrm{P}(T > 4) \approx 0.282$ 及 $\mathrm{P}(T > 4 \mid T > 2) \approx 0.426$。
- 第 (i) 部分对舍入非常敏感：$0.719^{16} \approx 0.0051$，但 $0.718^{16} \approx 0.0050$。答案写作约 0.005 即可。
- 即使没有截断，$\mathrm{P}(T > 4 \mid T > 2)$ 也会相同，因为因子 $1/k$ 会抵消。只有第 (i) 部分和平均值才受 k 影响。

常见错误

- 计算 $\mathrm{P}(T > 4)$ 时忘记除以 k：这样会得到 0.237 而不是 0.281，并在 (i) 中得到 $0.7633^{16} \approx 0.013$，是正确值的两倍有多。
- 把 T 的平均值当作 σ（即截断前的平均值），这样会得到 $\sigma = 3$。
- 在 (ii) 中求了 $\mathrm{P}(2 < T \leq 4)$ 或 $\mathrm{P}(T > 2)$，而不是条件概率（conditional probability）。

STEP 之外

平均值中出现的量 $\phi(1)/\Phi(1)$ 是*逆米尔斯比率*（inverse Mills ratio）的一个例子：一般而言，正态分布 $\mathrm{N}(\mu, \sigma^2)$ 在 a 处截去下方之后，其平均值为 $\mu + \sigma\phi(\alpha)/(1 - \Phi(\alpha))$，其中 $\alpha = (a - \mu)/\sigma$。凡是数据只能在某个门槛以上才观察得到的场合，都会出现截断分布（truncated distribution），例如存活分析（survival analysis）以及处理设限数据（censored data）的计量经济模型。第 (ii) 部分所讨论的是*存活函数*（survival function）$\mathrm{P}(T > t)$；指数分布（exponential distribution）是唯一能令 $\mathrm{P}(T > s + t \mid T > s)$ 与 s 无关的情况。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门教材）涵盖连续分布和条件概率。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。