详细解答

STEP II 2004，第 12 题

概率与统计 · 连续分布

STEP II 2004，第 12 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/04-s2-q12

关键思路

一切都源自一个原函数（antiderivative）：除了一个常数因子之外，$x\mathrm{e}^{-ax^2}$ 就是 e^{-ax^2} 的导数。由此可求出归一化常数（normalising constant），并得到累积分布函数（cumulative distribution function）的闭合形式，所以只需解一条指数方程便可求得中位数（median）。众数（mode）就是曲线达到顶点的位置，前提是该顶点位于 $[0,1]$ 之内；否则概率密度在 $x=1$ 处仍在上升。至于条件概率（conditional probability），只需用到累积分布函数在 m 和 h 的值。

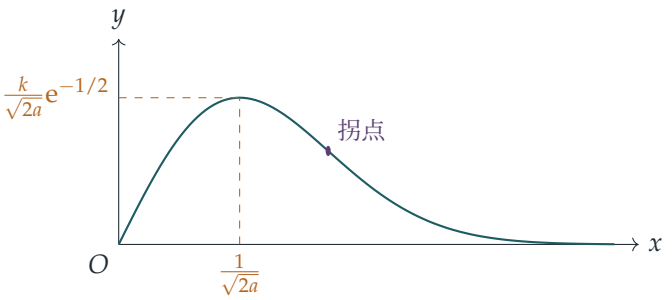
解答

草图

设 $g(x)=kx\mathrm{e}^{-ax^2}$，其中 $x\geq 0$。则 $g(0)=0$，当 $x>0$ 时 $g(x)>0$，而且由积法则（product rule）可得

$$g'(x)=k\mathrm{e}^{-ax^2}(1-2ax^2),$$

它在 $0\leq x<1/\sqrt{2a}$ 时为正，其后为负。所以 g 由原点开始上升（在该处的斜率为 k），在 $x=1/\sqrt{2a}$ 达到唯一的极大值，此时 $y=\dfrac{k}{\sqrt{2a}}\mathrm{e}^{-1/2}$，然后下降。当 $x\to\infty$ 时，指数因子压倒线性因子，所以 $y\to 0$：x 轴是渐近线（asymptote）。（再求导一次可知曲线在 $x=\sqrt{3/(2a)}$ 有一个拐点（point of inflection），曲线在该处由凹（concave）转为凸（convex）。）



常数 k

由于 $\dfrac{\mathrm{d}}{\mathrm{d}x}e^{-ax^2} = -2axe^{-ax^2}$，对于 $0 \le t \le 1$，

$$\int_0^t kxe^{-ax^2}\,\mathrm{d}x = \left[-\frac{k}{2a}e^{-ax^2}\right]_0^t = \frac{k}{2a}(1-e^{-at^2}).$$

总概率为 1，所以取 $t=1$ 可得 $\dfrac{k}{2a}(1-e^{-a}) = 1$，即

$$k = \frac{2a}{1-e^{-a}},$$

这正是所需的结果。（因为 $a>0$，所以分母为正。）把这个值代回去，在 $[0,1]$ 上的累积分布函数为

$$F(t) = P(X \le t) = \frac{1-e^{-at^2}}{1-e^{-a}}.$$

众数

概率密度函数（density）就是我们刚才所画的曲线，在 $x=1$ 处截断。从草图可见，在 $[0,1]$ 上，如果转向点（turning point）位于区间之内，密度便在该点最大，否则在区间的右端点最大。

- 若 $a > \frac{1}{2}$，则 $2a>1$，所以 $1/\sqrt{2a}<1$：顶点位于区间之内，故 $m = \dfrac{1}{\sqrt{2a}}$。
- 若 $a < \frac{1}{2}$，则 $1/\sqrt{2a}>1$，所以在整个 $[0,1)$ 上 $f'(x)>0$：密度在整个区间上递增，故 $m=1$。

（当 $a=\frac{1}{2}$ 时，两条公式都给出 $m=1$。）

中位数

中位数 h 满足 $F(h) = \frac{1}{2}$：

$$1-e^{-ah^2} = \tfrac{1}{2}\left(1-e^{-a}\right) \quad\Longleftrightarrow\quad e^{-ah^2} = \frac{1+e^{-a}}{2}.$$

取对数，并利用 $h > 0$，

$$h = \sqrt{\frac{1}{a}\ln\left(\frac{2}{1+\mathrm{e}^{-a}}\right)}.$$

由于 $1+\mathrm{e}^{-a} < 2$，对数为正，所以这是一个实数。

何时 $h > m$?

若 $a \le \frac{1}{2}$，则 $m = 1$，而 $h < 1$（一半的概率在 h 以下，而 1 以上则没有概率），所以 $h < m$。因此只有当 $a > \frac{1}{2}$ 时才可能有 $h > m$，而此时 $m = 1/\sqrt{2a}$。比较两者的平方（两者皆为正），

$$\begin{aligned} h > m &\iff \frac{1}{a}\ln\left(\frac{2}{1+\mathrm{e}^{-a}}\right) > \frac{1}{2a} \iff \ln\left(\frac{2}{1+\mathrm{e}^{-a}}\right) > \frac{1}{2} \\ &\iff \frac{2}{1+\mathrm{e}^{-a}} > \mathrm{e}^{1/2} \iff 1+\mathrm{e}^{-a} < 2\mathrm{e}^{-1/2} \\ &\iff \mathrm{e}^{-a} < 2\mathrm{e}^{-1/2} - 1 \iff a > -\ln\left(2\mathrm{e}^{-1/2}-1\right). \end{aligned}$$

第一步我们乘以 $a > 0$，第三步用到 exp 是递增函数，最后一步用到 ln 是递增函数。最后一步还需要 $2\mathrm{e}^{-1/2} - 1 > 0$，这是成立的，因为 $\mathrm{e}^{1/2} < 2$（因为 $\mathrm{e} < 4$）。

> **想一想** 以上一连串的推导假设了 $a > \frac{1}{2}$。结论不是需要把它列为额外条件吗?

不需要，而这正是下一部分要求证明它的原因。我们在下面会证明 $c = -\ln\left(2\mathrm{e}^{-1/2}-1\right)$ 大于 $\frac{1}{2}$。所以若 $a > c$，则自动有 $a > \frac{1}{2}$，公式 $m = 1/\sqrt{2a}$ 适用，而上述推导说明 $h > m$。连同 $a \le \frac{1}{2}$ 的情况，这就证明了 $h > m$ 当且仅当 $a > c$。

该常数大于 $\frac{1}{2}$

由于 ln 是递增函数，

$$-\ln\left(2\mathrm{e}^{-1/2}-1\right) > \tfrac{1}{2} \iff 2\mathrm{e}^{-1/2} - 1 < \mathrm{e}^{-1/2} \iff \mathrm{e}^{-1/2} < 1,$$

而最后一个不等式成立。（数值上，$c \approx 1.546$。）

条件概率

假设 $a > c$。则 $m < h$，所以事件 $\{X > m\} \cap \{X < h\}$ 就是事件 $\{m < X < h\}$。由于 $P(X < h) = \frac{1}{2}$，

$$P(X > m \mid X < h) = \frac{P(m < X < h)}{P(X < h)} = \frac{F(h) - F(m)}{\frac{1}{2}} = 1 - 2F(m).$$

当 $m = 1/\sqrt{2a}$ 时，$am^2 = \frac{1}{2}$，所以 $F(m) = \dfrac{1 - \mathrm{e}^{-1/2}}{1 - \mathrm{e}^{-a}}$。因此

$$P(X > m \mid X < h) = \frac{(1 - \mathrm{e}^{-a}) - 2(1 - \mathrm{e}^{-1/2})}{1 - \mathrm{e}^{-a}} = \frac{2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1}{1 - \mathrm{e}^{-a}},$$

这正是所需的结果。

验算

- 当 $a = c$ 时，分子 $2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1$ 恰好为 0，与该处 $m = h$ 吻合；当 $a > c$ 时它为正，符合概率应有的性质。
- 当 $a \to \infty$ 时，答案趋向 $2\mathrm{e}^{-1/2} - 1 \approx 0.213$，小于 1。
- 取 $a = 2.5$：$k \approx 5.447$，$m \approx 0.447$，$h \approx 0.496$，而对密度作数值积分可得 $P(X > m \mid X < h) \approx 0.1427$，与公式一致。取 $a = 1$ 则得 $m \approx 0.707 > h \approx 0.616$，与 $1 < c$ 相符。

常见错误

- 在所有情况下都答 $m = 1/\sqrt{2a}$：当 $a < \frac{1}{2}$ 时，转向点落在 $x = 1$ 之外，而该处的密度为零。
- 忘记 F 里面的 k 必须是归一化后的值，才能使 $F(1) = 1$。
- 写成 $P(X > m \mid X < h) = P(X > m)/P(X < h)$；分子必须是两个事件同时发生的概率。

STEP 之外

这个密度是截断瑞利分布（truncated Rayleigh distribution）：若不在 $x = 1$ 处截断，取 $k = 2a$ 的 $kx\mathrm{e}^{-ax^2}$ 就是一个点与原点距离的分布，而该点的两个坐标是互相独立的正态变量（normal variables）。比较平均值（mean）、中位数和众数，以及截断如何使它们移动，是大学第一门统计课程中反复出现的主题。Grimmett and Welsh 的 *Probability: An Introduction*（概率入门教材）是进一步了解连续随机变量（continuous random variables）及其密度函数和分布函数的好去处。

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。