解答例

STEP II 2004、第 12 問

確率・統計 ・ 連続分布

STEP II 2004、問 12

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/04-s2-q12

鍵となる考え方

すべては一つの原始関数（antiderivative）から導かれる。xe^{-ax^2} は、定数倍を除けば e^{-ax^2} の導関数（derivative）である。これにより正規化定数（normalising constant）と累積分布関数（cumulative distribution function）の閉じた形が得られ、中央値（median）は一つの指数方程式を解くだけで求まる。最頻値（mode）は曲線が頂点をとる点である。ただし、その頂点が $[0,1]$ の内部にある場合に限る。そうでなければ、密度（density）は $x=1$ においてもまだ増加している。条件付き確率（conditional probability）は、m と h における累積分布関数の値だけで求められる。

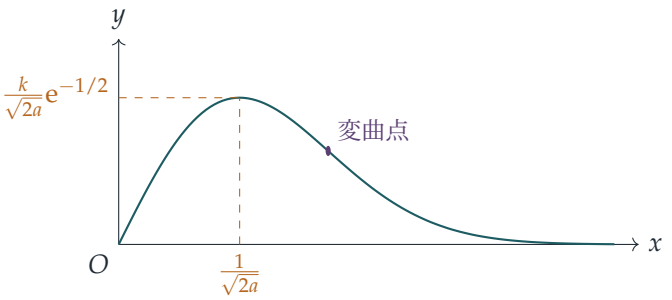
解答

概形

$x \geq 0$ に対して $g(x) = kxe^{-ax^2}$ とおく。このとき $g(0)=0$ であり、$x>0$ で $g(x)>0$ である。また積の微分法（product rule）により

$$g'(x) = ke^{-ax^2}(1-2ax^2),$$

となり、これは $0 \leq x < 1/\sqrt{2a}$ で正、それ以降で負である。よって g は原点から（そこでの傾きは k）増加し、$x = 1/\sqrt{2a}$ でただ一つの極大値 $y = \dfrac{k}{\sqrt{2a}}e^{-1/2}$ をとり、その後減少する。$x \to \infty$ のとき指数関数が一次の因子に勝つので $y \to 0$ となり、x 軸が漸近線（asymptote）である。（もう一度微分すると、$x = \sqrt{3/(2a)}$ に変曲点（point of inflection）があり、そこで曲線は上に凸（concave）から下に凸（convex）に変わることがわかる。）



定数 k

$\dfrac{\mathrm{d}}{\mathrm{d}x}e^{-ax^2} = -2axe^{-ax^2}$ であるから、$0 \le t \le 1$ に対して

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

全確率は 1 であるから、$t=1$ とすると $\dfrac{k}{2a}(1-e^{-a}) = 1$、すなわち

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

となり、題意が示された。（$a > 0$ より分母は正である。）これを代入し直すと、$[0,1]$ 上の累積分布関数は

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

最頻値

密度は先ほど概形を描いた曲線を $x=1$ で切り取ったものである。概形から、$[0,1]$ 上で密度が最大となるのは、極大点が区間内にあればその点であり、そうでなければ右端である。

- $a > \frac{1}{2}$ ならば $2a > 1$ であるから $1/\sqrt{2a} < 1$ である。頂点は区間の内部にあり、$m = \dfrac{1}{\sqrt{2a}}$ である。
- $a < \frac{1}{2}$ ならば $1/\sqrt{2a} > 1$ であるから、$[0,1)$ 全体で $f'(x) > 0$ である。密度は区間全体で増加し、$m = 1$ である。

（$a = \frac{1}{2}$ のときは、どちらの式でも $m = 1$ となる。）

中央値

中央値 h は $F(h)=\frac{1}{2}$ を満たす：

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

対数をとり、$h>0$ を用いると

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

$1+\mathrm{e}^{-a}<2$ より対数は正であるから、これは実数である。

$h>m$ となるのはいつか

$a\le\frac{1}{2}$ ならば $m=1$ であり、$h<1$（確率の半分は h より下にあり、1 より上には何もない）であるから $h<m$ である。したがって $h>m$ となりうるのは $a>\frac{1}{2}$ のときだけであり、そのとき $m=1/\sqrt{2a}$ である。2 乗どうしを比較すると（どちらも正である）、

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

最初の段階では $a>0$ を掛け、3 番目では exp が増加関数であること、最後では ln が増加関数であることを用いた。最後の段階では $2\mathrm{e}^{-1/2}-1>0$ も必要であるが、これは $\mathrm{e}^{1/2}<2$（$\mathrm{e}<4$ による）から成り立つ。

> **考えてみよう** 上の同値変形の連鎖は $a>\frac{1}{2}$ を仮定していた。結果にはこれを追加の条件として加える必要はないのか？

必要はない。だからこそ次の小問でそれを示させているのである。以下で $c=-\ln\left(2\mathrm{e}^{-1/2}-1\right)$ が $\frac{1}{2}$ より大きいことを示す。よって $a>c$ ならば自動的に $a>\frac{1}{2}$ となり、$m=1/\sqrt{2a}$ の式が適用でき、上の連鎖から $h>m$ が従う。$a\le\frac{1}{2}$ の場合と合わせて、$h>m$ となるのはちょうど $a>c$ のときであることが示された。

定数は $\frac{1}{2}$ より大きい

ln は増加関数であるから

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

であり、最後の不等式は正しい。（数値的には $c\approx1.546$ である。）

条件付き確率

$a > c$ とする。このとき $m < h$ であるから、事象 $\{X > m\} \cap \{X < h\}$ は事象 $\{m < X < h\}$ に等しい。$P(X < h) = \frac{1}{2}$ であるから

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

$m = 1/\sqrt{2a}$ のとき $am^2 = \frac{1}{2}$ であるから、$F(m) = \dfrac{1 - \mathrm{e}^{-1/2}}{1 - \mathrm{e}^{-a}}$ である。したがって

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

となり、題意が示された。

検算

- $a = c$ のとき分子 $2\mathrm{e}^{-1/2} - \mathrm{e}^{-a} - 1$ はちょうど 0 となり、そこで $m = h$ となることと合致する。また $a > c$ では正であり、確率として当然そうあるべきである。
- $a \to \infty$ のとき答えは $2\mathrm{e}^{-1/2} - 1 \approx 0.213$ に近づき、これは 1 より小さい。
- $a = 2.5$ のとき：$k \approx 5.447$、$m \approx 0.447$、$h \approx 0.496$ であり、密度を数値積分すると $P(X > m \mid X < h) \approx 0.1427$ となって、公式と一致する。$a = 1$ のときは $m \approx 0.707 > h \approx 0.616$ となり、$1 < c$ と整合する。

よくある間違い

- どの場合にも $m = 1/\sqrt{2a}$ と答えてしまう：$a < \frac{1}{2}$ のとき極大点は $x = 1$ より先にあり、そこでは密度は 0 である。
- F の中の k は $F(1) = 1$ となるように正規化した値でなければならないことを忘れる。
- $P(X > m \mid X < h) = P(X > m)/P(X < h)$ と書いてしまう：分子は**両方の**事象が起こる確率でなければならない。

STEP のその先へ

この密度は切断レイリー分布（truncated Rayleigh distribution）である。$x = 1$ での切断がなければ、$k = 2a$ としたときの $kx\mathrm{e}^{-ax^2}$ は、二つの座標が独立（independent）な正規分布（normal distribution）に従う確率変数である点の、原点からの距離の分布で

ある。平均（mean）、中央値、最頻値を比較し、切断によってそれらがどのようにずれるかを調べることは、最初の統計学の講義で繰り返し現れるテーマである。連続確率変数（continuous random variable）とその密度関数・分布関数についてさらに学ぶには、Grimmett と Welsh による *Probability: An Introduction*（確率論入門）がよい出発点である。

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。