解答例

STEP II 2005、第 14 問

確率・統計 ・ 混合分布

STEP II 2005、問 14

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/05-s2-q14

鍵となる考え方

この密度関数（density）は、すでによく知っている二つの密度関数、すなわち標準正規分布（standard normal distribution）と $[0,\lambda]$ 上の一様分布（uniform distribution）の重み付き混合である。したがって f の積分はすべて正規分布の部分と一様分布の部分に分かれ、それぞれの部分は標準的な事実（全確率が 1 であること、平均（mean）、$\mathrm{E}(X^2)$）から読み取ることができ、難しい積分計算をする必要はない。そのうえで分散（variance）は $\mathrm{E}(X^2)-\mu^2$ から求まる。

解答

定数 k、平均、分散

ϕ と g はどちらも確率密度関数（probability density function）であるから、それぞれ実数全体で積分すると 1 になる。よって

$$1=\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = k\,[1+\lambda\cdot 1] \quad\Longrightarrow\quad k=\frac{1}{1+\lambda}.$$

平均については、二つの成分の平均を用いる。標準正規分布の平均は 0 であり、$[0,\lambda]$ 上の一様密度 g の平均は $\frac{1}{2}\lambda$ である。ゆえに

$$\mu=\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x = k\left[\int x\phi(x)\,\mathrm{d}x+\lambda\int xg(x)\,\mathrm{d}x\right]=\frac{1}{1+\lambda}\left[0+\lambda\cdot\frac{\lambda}{2}\right]=\frac{\lambda^2}{2(1+\lambda)}.$$

同様に、標準正規分布では $\int x^2\phi(x)\,\mathrm{d}x=1$ であり（平均が 0 なので、これはその分散である）、また

$$\int x^2 g(x)\,\mathrm{d}x=\frac{1}{\lambda}\int_0^{\lambda} x^2\,\mathrm{d}x=\frac{\lambda^2}{3}.$$

よって

$$\mathrm{E}(X^2) = \frac{1}{1+\lambda}\left[1 + \lambda \cdot \frac{\lambda^2}{3}\right] = \frac{3+\lambda^3}{3(1+\lambda)}.$$

考えてみよう なぜ $\mathrm{E}(X^2)$ はこのように分けられるのに、分散は分けられないのだろうか？

期待値は積分であり、積分は線形であるから、f が各成分の密度関数の重み付き和であるのと同じ重みで、$\mathrm{E}(X^2)$ は各成分の 2 次モーメント（second moment）の重み付き和になる。分散は線形ではない（μ^2 を含む）ので、まず 2 次モーメントを合成し、そのあとで μ^2 を引かなければならない。そうすると

$$\begin{aligned}\sigma^2 = \mathrm{E}(X^2) - \mu^2 &= \frac{3+\lambda^3}{3(1+\lambda)} - \frac{\lambda^4}{4(1+\lambda)^2} \\ &= \frac{4(3+\lambda^3)(1+\lambda) - 3\lambda^4}{12(1+\lambda)^2} \\ &= \frac{12 + 12\lambda + 4\lambda^3 + 4\lambda^4 - 3\lambda^4}{12(1+\lambda)^2} = \frac{\lambda^4 + 4\lambda^3 + 12\lambda + 12}{12(1+\lambda)^2},\end{aligned}$$

となり、求める結果が得られる。

$\lambda = 2$ の場合

このとき $k = \frac{1}{3}$ であり、$[0,2]$ 上で $\lambda g(x) = 1$ であるから

$$f(x) = \begin{cases} \frac{1}{3}\phi(x) & x < 0 \text{ または} x > 2, \\ \frac{1}{3}(\phi(x) + 1) & 0 \le x \le 2. \end{cases}$$

また

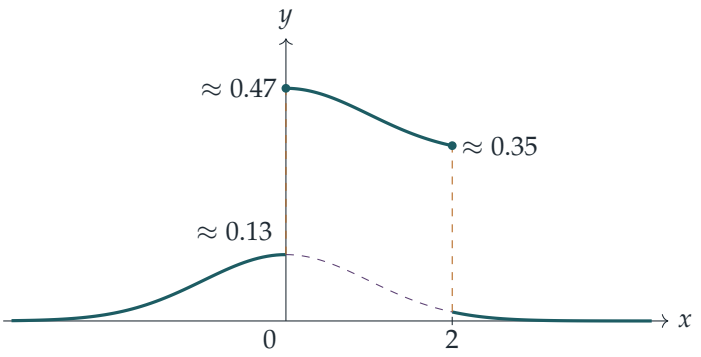
$$\mu = \frac{4}{6} = \frac{2}{3}, \qquad \sigma^2 = \frac{16 + 32 + 24 + 12}{12 \cdot 9} = \frac{84}{108} = \frac{7}{9}, \qquad \sigma = \frac{\sqrt{7}}{3}.$$

小問 (i)

この曲線は、標準正規分布の釣鐘型曲線を 3 分の 1 倍したものを、区間 $[0,2]$ 上で $\frac{1}{3}$ だけ持ち上げたものである。したがって $x = 0$ で上向きに、$x = 2$ で下向きに跳び、跳びの高さはいずれも $\frac{1}{3}$ である。$\phi(0) = 1/\sqrt{2\pi} \approx 0.399$ および $\phi(2) \approx 0.054$ を用いると、次のようになる。

- 0 のすぐ左では $f \approx 0.133$、0 のすぐ右では $f \approx 0.466$（最大値）である。
- 2 のすぐ左では $f \approx 0.351$、2 のすぐ右では $f \approx 0.018$ である。
- $[0,2]$ 上では、曲線は正規分布の曲線の形に沿って減少する。

▶ $x \to \pm\infty$ のとき $f \to 0$ であり、いたるところで $f > 0$ である。



$[0,2]$ 上のプラム色の破線は比較のために示した $\frac{1}{3}\phi(x)$ であり、実線が $y = f(x)$ である。（縦方向の目盛りは引き伸ばしてある。）

小問 (ii)

$F(x) = \mathrm{P}(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ とおく。$\frac{1}{3}\phi$ を積分すると $\frac{1}{3}\Phi$ となり、$[0,2]$ 上の追加分 $\frac{1}{3}$ は、$[0,2]$ のうち x より左にある部分の長さの $\frac{1}{3}$ 倍だけ寄与する。よって

$$F(x) = \begin{cases} \frac{1}{3}\Phi(x) & x < 0, \\ \frac{1}{3}\big(\Phi(x) + x\big) & 0 \le x \le 2, \\ \frac{1}{3}\big(\Phi(x) + 2\big) & x > 2. \end{cases}$$

各部分は $x = 0$ と $x = 2$ で一致する。これは当然であり（点質量（point mass）がなければ累積分布関数（CDF）は連続である）、また $x \to \infty$ のとき $F(x) \to \frac{1}{3}(1+2) = 1$ となる。

小問 (iii)

ここで $\mu + 2\sigma = \frac{2}{3} + \frac{2}{3}\sqrt{7} \approx 2.43$ であり、これは 2 より大きいので、そこでは F の最後の場合の式を用いる。X は連続型であるから

$$\begin{aligned} \mathrm{P}(0 < X < \mu + 2\sigma) &= F\big(\tfrac{2}{3} + \tfrac{2}{3}\sqrt{7}\big) - F(0) \\ &= \tfrac{1}{3}\big(\Phi(\tfrac{2}{3} + \tfrac{2}{3}\sqrt{7}) + 2\big) - \tfrac{1}{3}\Phi(0) \\ &= \tfrac{1}{3}\big(0.9921 + 2 - 0.5\big) = \frac{2.4921}{3} = 0.8307. \end{aligned}$$

検算

- $\lambda \to 0$ とすると一様分布の部分は消える。公式からは $\mu \to 0$ および $\sigma^2 \to \frac{12}{12} = 1$ が得られ、当然そうあるべきように標準正規分布の値と一致する。
- $\lambda = 2$ のとき、全分散の公式（law of total variance、下記参照）を使うと別の方法で同じ答えが得られる。各成分の重みは $\frac{1}{3}, \frac{2}{3}$、分散は $1, \frac{1}{3}$、平均は $0, 1$ である。分散の平均は $\frac{1}{3} + \frac{2}{9} = \frac{5}{9}$、平均の分散は $\frac{2}{3} - \frac{4}{9} = \frac{2}{9}$ であり、合計は $\frac{7}{9}$ となる。
- $\lambda = 0.5, 1, 2, 3.7$ について f を数値積分（中点則（midpoint rule）、0 と λ で区間を分割）したところ、全確率 1、平均、分散の公式が 10^{-6} より良い精度で再現された。$\lambda = 2$ では $\mu = 0.6667$ および $\sigma^2 = 0.7778$ が得られた。
- $\Phi(2.4305)$ を直接計算すると、問題文の 0.9921 ではなく 0.99246 となり、確率は 0.8307 ではなく 0.8308 となる。このわずかな違いは問題文で与えられた値によるものであり、解答ではその値を用いるべきである。

よくある間違い

- 係数 $k = \frac{1}{3}$ を忘れ、全確率が 3 になってしまう。
- 分散を $k[1 + \lambda \cdot \frac{\lambda^2}{12}]$ と計算する、つまり各成分の分散を合成してしまう。混合分布の分散は線形には合成できない。線形に合成できるのは 2 次モーメントである。
- f を連続な曲線として描いてしまう。実際には $x = 0$ と $x = \lambda$ で跳びがある。

STEP のその先へ

総和が 1 となる重み $w_i \geq 0$ を用いた $\sum_i w_i f_i(x)$ という形の密度関数は**混合分布**（mixture distribution）である。これは、確率 w_i で成分 i を選び、そのあと f_i から標本を取り出すことに相当する。ここでの重みは $\frac{1}{1+\lambda}$ と $\frac{\lambda}{1+\lambda}$ である。混合分布の分散は、姿を変えた全分散の公式にほかならない。すなわち、各成分の分散の平均に、各成分の平均の分散を加えたものである。混合分布は、統計学や機械学習（machine learning）におけるクラスタリング（clustering）手法の基礎となっている。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）は、条件付き期待値（conditional expectation）と全分散の公式をわかりやすく解説している。

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。