詳細解答

STEP II 2005，第 14 題

概率與統計 • 混合分佈

> **STEP II 2005，第 14 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/05-s2-q14

關鍵思路

這個密度是兩個我們早已熟悉的密度的加權混合：一個是標準正態分佈（standard normal distribution），另一個是 $[0,\lambda]$ 上的均勻分佈（uniform distribution）。因此，f 的每個積分都可拆成正態部分和均勻部分，而每一部分都可以直接由標準事實（總概率為 1、均值、$\mathrm{E}(X^2)$）讀出，完全不必做任何艱深的積分。方差（variance）則由 $\mathrm{E}(X^2)-\mu^2$ 求得。

解答

常數 k、均值與方差

由於 ϕ 和 g 都是概率密度函數（probability density function），兩者在整條實數線上的積分都是 1。所以

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

至於均值（mean），我們利用兩個成分各自的均值：標準正態分佈的均值為 0，而 $[0,\lambda]$ 上的均勻密度 g 的均值為 $\frac{1}{2}\lambda$。因此

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

同樣地，標準正態分佈滿足 $\int x^2\phi(x)\,\mathrm{d}x=1$（由於其均值為 0，這就是它的方差），而

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

所以

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

想一想 為甚麼 $\mathrm{E}(X^2)$ 可以這樣拆開，方差卻不可以？

期望值（expectation）是積分，而積分是線性的，所以 $\mathrm{E}(X^2)$ 是各成分二階矩（second moment）的加權組合，權重與 f 由各成分密度組合而成時所用的完全相同。方差並非線性（它涉及 μ^2），所以我們必須先把二階矩組合起來，然後才減去 μ^2。這樣做便得

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

這正是所要求的結果。

$\lambda = 2$ 的情況

此時 $k = \frac{1}{3}$，而在 $[0,2]$ 上 $\lambda g(x) = 1$，所以

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

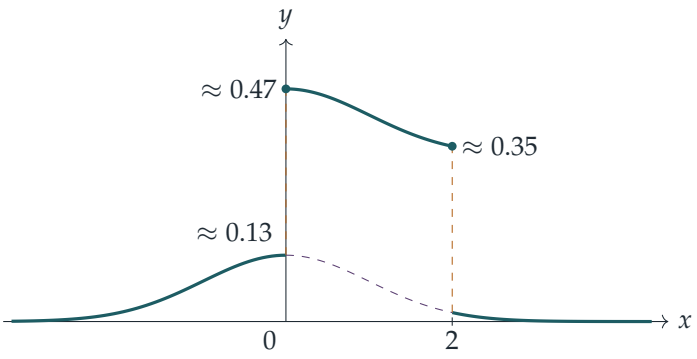
此外

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

第 (i) 部分

這條曲線是標準正態鐘形曲線的三分之一，並在區間 $[0,2]$ 上向上提升了 $\frac{1}{3}$。所以它在 $x = 0$ 處向上跳躍，在 $x = 2$ 處向下跳躍，跳躍的高度都是 $\frac{1}{3}$。利用 $\phi(0) = 1/\sqrt{2\pi} \approx 0.399$ 和 $\phi(2) \approx 0.054$：

- 在 0 的左邊緊貼處，$f \approx 0.133$；在 0 的右邊緊貼處，$f \approx 0.466$（最大值）；
- 在 2 的左邊緊貼處，$f \approx 0.351$；在 2 的右邊緊貼處，$f \approx 0.018$；
- 在 $[0,2]$ 上曲線遞減，形狀與正態曲線一致；
- 當 $x \to \pm\infty$ 時 $f \to 0$，而且處處 $f > 0$。



$[0,2]$ 上的紫紅色虛線是 $\frac{1}{3}\phi(x)$，畫出來以作比較；實線則是 $y = f(x)$。（縱軸的比例已被拉長。）

第 (ii) 部分

設 $F(x) = \mathrm{P}(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$。對 $\frac{1}{3}\phi$ 積分得 $\frac{1}{3}\Phi$，而 $[0,2]$ 上額外的 $\frac{1}{3}$ 所貢獻的，是 $\frac{1}{3}$ 乘以 $[0,2]$ 中位於 x 左邊那一段的長度。所以

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

各段在 $x = 0$ 和 $x = 2$ 處互相吻合，這是必然的（沒有點質量（point mass）時，累積分佈函數（CDF）是連續的），而且當 $x \to \infty$ 時 $F(x) \to \frac{1}{3}(1+2) = 1$。

第 (iii) 部分

這裏 $\mu + 2\sigma = \frac{2}{3} + \frac{2}{3}\sqrt{7} \approx 2.43$，大於 2，所以在該處要用 F 的最後一個分支。由於 X 是連續的，

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

驗算

- 當 $\lambda \to 0$ 時，均勻部分消失：公式給出 $\mu \to 0$ 和 $\sigma^2 \to \frac{12}{12} = 1$，正是標準正態分佈的數值，合乎預期。

- 對於 $\lambda = 2$，全方差定律（law of total variance，見下文）以另一種方法給出相同的答案。各成分的權重為 $\frac{1}{3}, \frac{2}{3}$，方差為 $1, \frac{1}{3}$，均值為 $0, 1$。方差的平均值是 $\frac{1}{3} + \frac{2}{9} = \frac{5}{9}$，均值的方差是 $\frac{2}{3} - \frac{4}{9} = \frac{2}{9}$，總和是 $\frac{7}{9}$。
- 以數值方法對 f 積分（中點法則，在 0 和 λ 處分段），取 $\lambda = 0.5, 1, 2, 3.7$，所得的總概率 1、均值和方差公式全部吻合，誤差小於 10^{-6}；對於 $\lambda = 2$，結果是 $\mu = 0.6667$ 和 $\sigma^2 = 0.7778$。
- 直接計算 $\Phi(2.4305)$ 得 0.99246，而非題目所引用的 0.9921；概率則是 0.8308，而非 0.8307。這微小的差異源自題目所給的數值，而我們應該採用題目的數值。

常見錯誤

- 忘記了因子 $k = \frac{1}{3}$，以致總概率變成 3。
- 把方差算成 $k[1 + \lambda \cdot \frac{\lambda^2}{12}]$，即把各成分的方差組合起來。混合分佈的方差不能線性地組合；二階矩才可以。
- 把 f 畫成連續曲線：它在 $x = 0$ 和 $x = \lambda$ 處有跳躍。

STEP 之外

形如 $\sum_i w_i f_i(x)$ 的密度，其中權重 $w_i \geq 0$ 且總和為 1，稱為**混合分佈**（mixture distribution）：以概率 w_i 選出成分 i，然後從 f_i 中抽樣。這裏的權重是 $\frac{1}{1+\lambda}$ 和 $\frac{\lambda}{1+\lambda}$。混合分佈的方差其實就是全方差定律的化身：各成分方差的平均值，加上各成分均值的方差。混合分佈是統計學和機器學習中聚類（clustering）方法的基礎。Blitzstein 與 Hwang 的 *Introduction to Probability*（概率論入門教材）對條件期望值（conditional expectation）和全方差定律有清晰的講解。

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。