解答例

# STEP I 2005、第 13 問

確率・統計 ・ 対称な分布

> **STEP I 2005、問 13**
>
> 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-s1-q13

## 鍵となる考え方

分布（distribution）の形については、2 つの累積確率（cumulative probability）と対称性（symmetry）以外に何も与えられていない。したがって、すべてをこれらから組み立てなければならない。$\mu$ に関する対称性からは、さらに 2 つの事実がただちに得られる。すなわち、確率の半分ずつが $\mu$ の両側にあること、そして $\mu - k\sigma$ より下の裾の確率が $\mu + k\sigma$ より上の裾の確率に等しいことである。小問 (ii) での工夫は、登場する容量（495, 500, 505 ml）がいずれも、どちらかの種類の牛乳にとって $\mu$、$\mu \pm \frac{1}{2}\sigma$、$\mu + \sigma$ のいずれかに位置していることに気づく点にある。

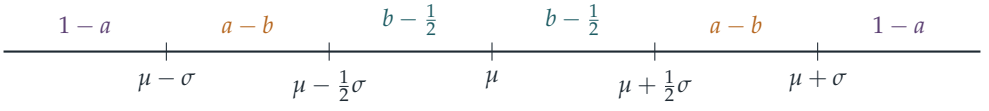
## 解答

$X$ は容量を表すモデルなので、全体を通して連続（continuous）確率変数として扱う。したがって 1 点での確率は 0 であり、$<$ と書いても $\leq$ と書いても違いはない。

$\mu$ に関する対称性より、任意の $k$ に対して $\mathrm{P}(X \leq \mu) = \frac{1}{2}$ および $\mathrm{P}(X \leq \mu - k\sigma) = \mathrm{P}(X \geq \mu + k\sigma)$ が成り立つ。特に

$$\mathrm{P}(X \leq \mu - \tfrac{1}{2}\sigma) = 1 - b, \qquad \mathrm{P}(X \leq \mu - \sigma) = 1 - a.$$

$\mu$ から $\sigma$ を単位として測った各区間の確率を書き留めておくと見通しがよい。




たとえば $\mathrm{P}(\mu < X \leq \mu + \frac{1}{2}\sigma) = b - \frac{1}{2}$、$\mathrm{P}(\mu + \frac{1}{2}\sigma < X \leq \mu + \sigma) = a - b$ である。左側

の区間は対称性から従う。

**小問 (i)**

累積確率の差をとって、

$$\mathrm{P}\left(\mu - \tfrac{1}{2}\sigma \le X \le \mu + \sigma\right) = \mathrm{P}(X \le \mu + \sigma) - \mathrm{P}\left(X < \mu - \tfrac{1}{2}\sigma\right) = a - (1-b) = a + b - 1.$$

条件付き確率（conditional probability）については、分子の事象は $\{\mu - \frac{1}{2}\sigma \le X \le \mu + \frac{1}{2}\sigma\}$ であり、その確率は $b - (1-b) = 2b - 1$ である。一方、$\mathrm{P}(X \ge \mu - \frac{1}{2}\sigma) = 1 - (1-b) = b$ である。よって

$$\mathrm{P}\left(X \le \mu + \tfrac{1}{2}\sigma \mid X \ge \mu - \tfrac{1}{2}\sigma\right) = \frac{2b-1}{b}.$$

**小問 (ii)(a)**

紙パックが無脂肪乳（skimmed milk）である事象を $S$、全乳（full-fat milk）である事象を $F$ とすると、$\mathrm{P}(S) = 0.6$、$\mathrm{P}(F) = 0.4$ である。容量（ml）を $V$ とする。

> **考えてみよう** それぞれの種類の牛乳にとって、500 ml と 505 ml は $\sigma$ を単位としてどこに位置するか。

無脂肪乳（$\mu = 500, \sigma = 10$）では、500 は $\mu$、505 は $\mu + \frac{1}{2}\sigma$ である。全乳（$\mu = 495$, $\sigma = 10$）では、500 は $\mu + \frac{1}{2}\sigma$、505 は $\mu + \sigma$ である。区間の確率を読み取ると、

$$\mathrm{P}(V < 505 \mid S) = b, \qquad \mathrm{P}(500 < V < 505 \mid S) = b - \tfrac{1}{2},$$
$$\mathrm{P}(V < 505 \mid F) = a, \qquad \mathrm{P}(500 < V < 505 \mid F) = a - b.$$

全確率の法則（law of total probability）より、

$$\mathrm{P}(V < 505) = 0.6b + 0.4a,$$
$$\mathrm{P}(500 < V < 505) = 0.6\left(b - \tfrac{1}{2}\right) + 0.4(a - b) = 0.4a + 0.2b - 0.3.$$

「500 ml より多く 505 ml より少ない」という事象が求めたい共通部分であるから、

$$\mathrm{P}(V > 500 \mid V < 505) = \frac{0.4a + 0.2b - 0.3}{0.4a + 0.6b} = \frac{4a + 2b - 3}{4a + 6b}.$$

**小問 (ii)(b)**

無作為に選んだ紙パックが非常に多数あれば、観測された割合は対応する確率の推定値となる。1 つ目の情報は $\mathrm{P}(V \le 505) = 0.7$ ということであり、小問 (a) の計算によれば

$$0.6b + 0.4a = 0.7, \quad \text{すなわち} \quad 4a + 6b = 7. \tag{1}$$

**考えてみよう** それぞれの種類の牛乳について、$\mathrm{P}(V \geq 495)$ はいくらか。

無脂肪乳では 495 は $\mu - \frac{1}{2}\sigma$ なので、$\mathrm{P}(V \geq 495 \mid S) = b$ である（これは小問 (i) の分母であった）。全乳では 495 は $\mu$ なので、対称性より $\mathrm{P}(V \geq 495 \mid F) = \frac{1}{2}$ である。ベイズ（Bayes）の定理（Bayes' theorem）より、

$$\mathrm{P}(F \mid V \geq 495) = \frac{0.4 \times \frac{1}{2}}{0.6b + 0.4 \times \frac{1}{2}} = \frac{0.2}{0.6b + 0.2}.$$

これを $\frac{1}{3}$ に等しいとおくと $0.6b + 0.2 = 0.6$ となり、$b = \frac{2}{3}$ である。すると (1) より $4a = 7 - 4 = 3$ となるので、

$$a \approx \tfrac{3}{4}, \qquad b \approx \tfrac{2}{3}.$$

観測された割合は真の確率を近似するにすぎないので、これらはあくまで推定値である。

## 検算

- 推定値は妥当である。$\mu < \mu + \frac{1}{2}\sigma < \mu + \sigma$ であるから $\frac{1}{2} < b < a < 1$ でなければならず、実際そうなっている。
- 代入して確かめると、$0.6 \cdot \frac{2}{3} + 0.4 \cdot \frac{3}{4} = 0.4 + 0.3 = 0.7$、また $0.2/(0.4 + 0.2) = \frac{1}{3}$ である。
- 正規分布（Normal distribution）では $a \approx 0.841$、$b \approx 0.691$ であり、これらを用いると小問 (i) の値は 0.533 と 0.553 になる。（正規分布でない）ラプラス（Laplace）分布から $10^5$ 個以上の標本をシミュレーションしたところ、小問 (i) と (ii)(a) のすべての式と一致した。たとえば (ii)(a) では、式の値 0.2515 に対して 0.2518 が得られた。

## よくある間違い

- $1 - b$ とすべきところを $\mathrm{P}(X \leq \mu - \frac{1}{2}\sigma) = b$ と書いてしまう。対称性で折り返されるのは **裾の確率**であって、累積確率ではない。
- 2 種類の牛乳の平均（mean）が異なることを忘れる。そのため、同じ容量（たとえば 500 ml）でも、平均から何標準偏差（standard deviation）離れているかはそれぞれで異なる。
- 両方の種類について $\mathrm{P}(V \geq 495) = b$ を使ってしまう。全乳では 495 ml は平均である。

## STEP のその先へ

この問題の要点は、分布がまったくわからなくても、対称性と全確率の法則だけでかなり先まで進めるということである。ここでの 2 つの母集団の混合は単純な**混合モデル**（mixture model）であり、小問 (ii)(b) はそのパラメータをデータから推定することの初歩版である。分布によらない関連する結果として、分散（variance）だけを用いて $\mathrm{P}(|X-\mu| \geq k\sigma)$ を上から評価するチェビシェフ（Chebyshev）の不等式（Chebyshev's inequality）がある。Blitzstein と Hwang の *Introduction to Probability*（確率論入門）、または Grimmett と Welsh の *Probability: An Introduction*（確率論入門）を参照せよ。

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。