解答例

STEP I 1989、第 14 問

確率・統計 ・ 条件付き分布

> **STEP I 1989、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/89-s1-q14

鍵となる考え方

すべては一組の確率、すなわち種が幅 1 リーグの各帯状地帯に落ちる確率に基づいている。種が木に育つのは、公園地か農地に落ちたときであり、そのときに限る。したがって木の本数は二項分布（binomial distribution）に従う。木が育つ距離 Y は、陸地に落ちたという**条件のもとでの**（conditioned）L にほかならない。その確率密度関数（probability density function）は、f を陸地の帯だけに制限し、全面積が 1 になるように拡大したものである。最後の小問は仮説検定（hypothesis test）である。二項分布に従う個数として、農地の木が 23 本というのは驚くほど多いのだろうか。

解答

各帯の確率

L の累積分布関数（cumulative distribution function）は、$0 \le x \le 4$ に対して

$$\mathrm{F}(x) = \int_0^x \left(\tfrac{1}{2} - \tfrac{1}{8}t\right)\mathrm{d}t = \frac{x}{2} - \frac{x^2}{16},$$

であるから、$\mathrm{F}(1) = \frac{7}{16}$、$\mathrm{F}(2) = \frac{3}{4}$、$\mathrm{F}(3) = \frac{15}{16}$、$\mathrm{F}(4) = 1$ である。よって

$$\mathrm{P}(\text{公園地}) = \mathrm{P}(0 \le L \le 1) = \tfrac{7}{16}, \qquad \mathrm{P}(\text{湖}) = \tfrac{3}{4} - \tfrac{7}{16} = \tfrac{5}{16},$$
$$\mathrm{P}(\text{農地}) = \mathrm{P}(2 \le L \le 3) = \tfrac{15}{16} - \tfrac{3}{4} = \tfrac{3}{16}, \qquad \mathrm{P}(\text{海}) = \tfrac{1}{16}.$$

（L は連続型なので境界の確率は 0 であり、境界がどちらの帯に属するかは問題にならない。）種が木に育つ確率は

$$p = \tfrac{7}{16} + \tfrac{3}{16} = \tfrac{5}{8}.$$

である。1 年間に育つ木の本数を T とする。96 個の種は互いに独立に風で運ばれ、それぞれ確率 $\frac{5}{8}$ で育つので、$T \sim \mathrm{B}(96, \frac{5}{8})$ であり、毎年新たに育つフィーバーツリーの本数の平均は

$$\mathrm{E}(T) = 96 \times \tfrac{5}{8} = 60.$$

である。

小問 (i)

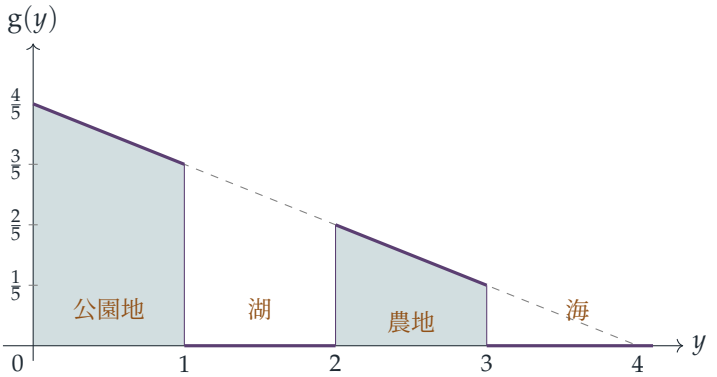
Y は、種が陸地に落ちたという条件のもとでの L の分布をもつ。陸地を $A = [0,1] \cup [2,3]$ と書く。$y \in A$ に対して

$$\mathrm{P}(y \le Y \le y + \delta y) = \frac{\mathrm{P}(y \le L \le y + \delta y)}{\mathrm{P}(L \in A)} \approx \frac{\mathrm{f}(y)\,\delta y}{5/8},$$

であり、Y が A の外の値をとることはない。したがって Y の確率密度関数は

$$\mathrm{g}(y) = \begin{cases} \frac{8}{5}(\frac{1}{2} - \frac{1}{8}y) = \dfrac{4-y}{5} & \text{条件：} 0 \le y \le 1 \text{ または} 2 \le y \le 3, \\ 0 & \text{それ以外のとき} \end{cases}$$

である。確認として、その面積は $\frac{8}{5}(\frac{7}{16} + \frac{3}{16}) = 1$ となる。グラフは同じ右下がりの直線の 2 つの部分からなる。公園地の上では $\frac{4}{5}$ から $\frac{3}{5}$ まで下がり、湖の上では何もなく、農地の上では $\frac{2}{5}$ から $\frac{1}{5}$ まで下がる。



破線は比較のために示した $\frac{8}{5}\mathrm{f}$ である。Y の確率密度関数は陸地の上でだけこれに一致する。

平均については、$\int y(4-y)\,\mathrm{d}y = 2y^2 - \frac{1}{3}y^3$ であるから

$$\begin{aligned} \mathrm{E}(Y) &= \frac{1}{5}\left[2y^2 - \tfrac{1}{3}y^3\right]_0^1 + \frac{1}{5}\left[2y^2 - \tfrac{1}{3}y^3\right]_2^3 \\ &= \frac{1}{5}\left(\frac{5}{3}\right) + \frac{1}{5}\left(9 - \frac{16}{3}\right) = \frac{1}{5}\cdot\frac{5}{3} + \frac{1}{5}\cdot\frac{11}{3} = \frac{16}{15}. \end{aligned}$$

となる。したがって新しい木は、平均して城の南 $\frac{16}{15}$ リーグの位置に育つ。これは公園地を少し越えた湖の中であるが、それで問題はない。平均は確率変数が実際にとりうる値である必要はないからである。

小問 (ii)

1 年間に農地で育つ木の本数を X とする。しきたりが守られていれば 96 個の種が放たれ、それぞれが独立に確率 $\frac{3}{16}$ で農地に落ちるので

$$X \sim \mathrm{B}\left(96, \tfrac{3}{16}\right), \qquad \mathrm{E}(X) = 18, \qquad \mathrm{Var}(X) = 96 \cdot \tfrac{3}{16} \cdot \tfrac{13}{16} = \tfrac{117}{8} \approx 14.6,$$

であり、標準偏差（standard deviation）はおよそ 3.82 となる。

> **考えてみよう** 正確には何を検定しているのか。そして、どちらの向きに検定するのか。

帰無仮説（null hypothesis）は、しきたりが守られた、すなわち種は 96 個で $X \sim \mathrm{B}\left(96, \frac{3}{16}\right)$ である、というものである。農地の木が通常より**多い**ことから自然に生じる疑いは、96 個より多くの種が放たれたというものである。そこで片側検定（one-tailed test）を用い、帰無仮説のもとで 23 本**以上**となる確率がどれほどかを調べる。

$np = 18$ と $n(1-p) = 78$ はともに十分大きいので、連続性補正（continuity correction）を施した正規近似（normal approximation）$X \approx \mathrm{N}(18,\, 14.625)$ を用いる。

$$\mathrm{P}(X \geq 23) \approx \mathrm{P}\left(Z \geq \frac{22.5 - 18}{3.824}\right) = \mathrm{P}(Z \geq 1.18) \approx 0.12.$$

（二項分布による正確な値は 0.121 である。）したがって通常の条件のもとでも、農地の木が 23 本以上になることはおよそ 8 年に 1 度は起こる。これは 5% の有意水準（significance level）（さらには 10%）にもまったく届かない。この観測結果はしきたりが守られていたこととよく整合しており、この証拠だけでは王の疑いは正当化**されない**。

比較のために述べると、5% の水準であっても、王が妥当な根拠をもつには本数が少なくとも 25 本（正確には $\mathrm{P}(X \geq 25) \approx 0.049$）でなければならない。もし王が公園地の木の本数も知っていれば、木の総数 $T \sim \mathrm{B}\left(96, \frac{5}{8}\right)$ を検定することができ、そのほうが放たれた種の個数についてはるかによい証拠となるだろう。

検算

- 4 つの帯の確率 $\frac{7}{16}, \frac{5}{16}, \frac{3}{16}, \frac{1}{16}$ の和は 1 であり、Y の確率密度関数の全面積も 1 である。

- 中点則（midpoint rule）による数値積分では $\mathrm{E}(Y) = 1.06667$ となり、$\frac{16}{15}$ と一致する。400 000 個の種のシミュレーション（U を $[0,1]$ 上の一様分布として $L = 4 - 4\sqrt{1-U}$ を用いる）では、62.5% が陸地に落ち、それらの落下距離の平均は 1.069 であった。
- 二項分布の確率を正確に足し合わせると $\mathrm{P}(X \geq 23) = 0.1213$ となり、正規近似による推定値 0.120 に近い。

よくある間違い

- f そのものを Y の確率密度関数として用いてしまう。陸地の上での f の面積は $\frac{5}{8}$ しかないので、$\frac{5}{8}$ で割らなければならない。
- $\mathrm{E}(L\,;\, L \in A) = \int_A y\,\mathrm{f}(y)\,\mathrm{d}y = \frac{2}{3}$ を計算して、$\mathrm{P}(L \in A)$ で割るのを忘れてしまう。
- 裾の確率 $\mathrm{P}(X \geq 23)$ ではなく $\mathrm{P}(X = 23)$ を検定してしまう。X がある特定の 1 つの値をとる確率は、どの値であってもかなり小さい。

STEP のその先へ

小問 (i) は**条件付き分布**（conditional distribution）の一例である。すなわち、確率変数をある事象に制限し、全確率が 1 になるように正規化し直す（renormalising）操作である。同じ操作が、確率密度関数に対するベイズの定理（Bayes' theorem）や、統計学のいたるところで用いられる切断分布（truncated distribution）の基礎にある。小問 (ii) は仮説検定と p 値（p-value）への最初の入門であり、大学 1 年次の講義ではネイマン–ピアソン（Neyman–Pearson）の枠組みによってこれを厳密に扱う。Blitzstein and Hwang の *Introduction to Probability*（確率論入門）は、この問題のような具体的な状況設定の例題を豊富に用いて条件付き分布を扱っている。

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。