解答例

STEP I 1995、第 14 問

確率・統計 ・ 比率の推定

STEP I 1995、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/95-s1-q14

鍵となる考え方

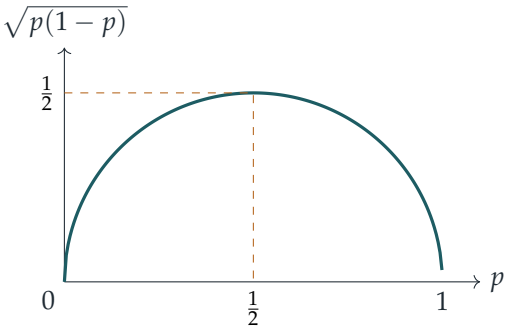
標本比率（sample proportion）は、二項分布（binomial distribution）に従う個数を n で割ったものである。n が大きいとき、これは近似的に平均 p、分散 $p(1-p)/n$ の正規分布（normal distribution）に従う。困るのは、分散がまさに未知の p に依存していることである。小問 (i) はこれを解決する。標準偏差（standard deviation）がとりうる最悪の値を与えるので、最悪の場合に合わせて選んだ標本の大きさは、p がどんな値であっても通用する。そのうえで小問 (iii) では、p についておおよその見当がすでについている場合に何が変わるかを問うている。

解答

小問 (i)

平方完成（completing the square）により、$f(p) = p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2$ とおく。よって $[0,1]$ 上で $f(p) \leq \frac{1}{4}$ であり、等号は $p = \frac{1}{2}$ のときに限り成り立つ。平方根は増加関数であるから、

$$\max_{0 \leq p \leq 1} \sqrt{p(1-p)} = \sqrt{\tfrac{1}{4}} = \tfrac{1}{2}, \qquad \text{最大値をとるのは} p = \tfrac{1}{2}.$$



小問 (ii)

標本に含まれる女性の人数を X とする。選ばれた各人は確率 p で女性である。母集団（population）は大きいと仮定し、n 回の選択は事実上独立（independent）であるとみなす（非常に大きな母集団からの非復元抽出は、復元抽出とほとんど同じである）。このとき

$$X \sim \mathrm{B}(n,p), \qquad \hat{p} = \frac{X}{n}, \qquad \mathrm{E}(\hat{p}) = p, \qquad \mathrm{Var}(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

n が大きいとき、中心極限定理（central limit theorem）（二項分布の正規近似）により

$$\hat{p} \approx \mathrm{N}\left(p,\, \frac{p(1-p)}{n}\right), \qquad \text{よって} \qquad Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx \mathrm{N}(0,1).$$

求めたいのは $\mathrm{P}(|\hat{p}-p| > 0.01) < 0.01$、すなわち

$$\mathrm{P}\left(|Z| > \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) < 0.01.$$

対称性により、この 1% は両側の裾に 0.5% ずつ分かれる。$\Phi(2.576) = 0.995$ であるから、

$$\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}} > 2.576 \quad \Longleftrightarrow \quad \sqrt{n} > 257.6\sqrt{p(1-p)}.$$

が必要である。

考えてみよう p はわからないのに、p を含む不等式からどうやって n を選べばよいのだろうか？

ここで小問 (i) が生きてくる。p がどんな値でも $\sqrt{p(1-p)} \le \frac{1}{2}$ である。したがって $\sqrt{n} > 257.6 \times \frac{1}{2} = 128.8$ ならば、すべての p に対して不等式が成り立つ。この条件は必要条件でもある。$p = \frac{1}{2}$ のとき（そして女性の割合がそれに近いとき）にこの上界が達成されるからである。よって

$$n > 128.8^2 = 16\,589.44,$$

が必要であり、求める標本の大きさは

$$\boxed{n = 16\,590} \qquad (\text{約}16\,600).$$

である。より精密なパーセント点 $z = 2.5758$ を用いると $n > 16\,587.2$ となる。いずれにせよ答えは「約 16 600」であり、有効数字 3 桁目は使う数表によって変わる。

小問 (iii)

一般に、条件は

$$n > \left(\frac{2.576}{0.01}\right)^2 p(1-p) \approx 66\,360\,p(1-p).$$

である。16 590 という値は最悪の場合 $p = \frac{1}{2}$ から得られたものであった。p が $\frac{1}{2}$ から十分離れていることがわかっていれば、$p(1-p)$ に対してより小さい上界を使える。

(a) 左利き。人々のおよそ 10% が左利きである。$p \le 0.15$ と仮定して差し支えなければ、$p(1-p)$ は $[0, \frac{1}{2}]$ で増加するので $p(1-p) \le 0.15 \times 0.85 = 0.1275$ となる。これから $n > 8461$ を得る。$p = 0.1$ ちょうどならば $n > 66\,360 \times 0.09 \approx 5970$ となり、約 6000 である。いずれにしても、何もわからないときに必要な 16 590 よりずっと小さな標本で済む。np は数百のオーダーなので、正規近似も依然として良い近似である。

(b) 百万長者。今度は p が非常に小さく、おそらく 0.001 から 0.01 程度である。文字どおりに受け取れば、$p \le 0.01$ から $p(1-p) \le 0.0099$、$n > 657$ を得る。しかしこれは要点を外している。

- p 自体が約 0.001 のとき、誤差 0.01 には意味がない。推定値 $\hat{p} = 0$ でさえ真の値から 0.01 以内にあるからである。本当に必要なのは**相対的な**精度、たとえば誤差を p の 1% 以下に抑えることである。そうすると $n > 2.576^2(1-p)/(p \times 0.01^2)$ が必要となり、$p = 0.001$ ではこれは約 6.6×10^7 である。桁違いに大きい。
- p がごく小さく n が数百程度しかないと、np は 1 より小さい。このとき二項分布は大きく歪んでおり、正規近似はもはや妥当でない。ポアソン近似（Poisson approximation）のほうがよいモデルとなるだろう。
- 実際には百万長者に接触するのは難しく、人々は自分の資産について本当のことを言わないかもしれない。したがって無作為抽出が可能かどうかも、正直な回答が得られるかどうかも疑わしい。

このように、左利きの場合は p のおおよその値がわかっていれば標本を小さくできる。一方、百万長者の場合は、意味のある精度を求めようとするとはるかに大きな標本が必要になるか、まったく別の抽出方法が必要になる。

検算

- $p = \frac{1}{2}$、$n = 16\,590$ のとき、正規近似により $\mathrm{P}(|\hat{p} - p| > 0.01) = 2\big(1 - \Phi(2.5762)\big) \approx 0.00999$ となり、1% をわずかに下回る。
- $n = 16\,590$、$p = \frac{1}{2}$ に対して二項分布の確率を正確に足し合わせると約 0.0102 となる。このわずかな超過は二項分布の離散性によるものである（連続性補正（continuity correction）なしの正規近似を用いたため）。$p = 0.3$ では正確な値は約 0.005 であり、小問 (i) の予想どおり十分に条件を満たしている。
- $p = 0.1$、$n = 6000$ に対する二項分布の正確な確率は約 0.0098 であり、(iii)(a) の値が確かめられる。

よくある間違い

- 2.576 の代わりに 1.96 や 2.326 を使ってしまう。1% は両側の裾に分かれるので、$\Phi(z) = 0.995$ が必要である。
- n を p の式のまま残してしまう。問題が求めているのはどんな p に対しても通用するただ一つの数であり、そのために小問 (i) がある。
- 「$\sqrt{p(1-p)} \le \frac{1}{2}$」から $n > 128.8$ が得られると読み違える。128.8 を超えなければならないのは n ではなく $\sqrt{n}$ である。

STEP のその先へ

これは世論調査で示される「誤差の幅（margin of error）」の背後にある標準的な計算である。約 1000 人の調査では、信頼度 95% でおよそ ±3 パーセントポイントの誤差となるが、これはまさに小問 (i) の最悪の場合の議論による。大学では、これは信頼区間（confidence interval）へとつながり、さらにまれな事象の比率をどう推定するかという問題へとつながる。そこではポアソン近似や正確な二項分布の方法が正規近似に取って代わる。Blitzstein and Hwang, *Introduction to Probability*（確率論入門）は、中心極限定理とこれらの近似を非常に読み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。