13.4 平均か中央値か？

ワークシート

問題

STEP II 2000、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/00-s2-q14

ヒント

まずは自力で問題に取り組んでみよう。行き詰まったら、ヒントを**一つずつ**読み、そのたびに問題に戻ること。**手を止めて、やってみよう**と記したヒントが重要な段階である。各ヒントの後には計算用の余白がある。

? **考えてみよう**　中央値の密度（density）が $y^n(1-y)^n$ に比例するのはなぜか。（Y が y の近くにあるためには、$2n+1$ 個の値のうちいくつが y より小さく、いくつが y より大きくなければならないか。）

段階 1　定数

扱う内容：定数 k

? **考えてみよう**　確率密度関数（density）は積分すると 1 になる。問題で与えられた積分を、二つの指数をともに n として用いよ。

段階 2 平均（mean）と分散（variance）

扱う内容：平均、および分散

考えてみよう 密度関数は $\frac{1}{2}$ に関して対称（symmetric）か。そこから $\mathrm{E}(Y)$ が得られる。

手を止めて、やってみよう $\mathrm{E}(Y^2)$ については、与えられた積分を指数 $n+2$ と n で再び用い、階乗（factorial）を整理せよ。

段階 3　平均との比較

扱う内容：中央値（median）と平均の比較

考えてみよう　$\bar{X}$ の平均と分散はいくらか。n が大きいときの二つの分散を比較せよ。

手を止めて、やってみよう　Y と $\bar{X}$ をともに平均 $\frac{1}{2}$ の正規分布（normal distribution）に従うものとみなせ。それぞれの確率を Φ を用いて表せ。どちらが大きいか、またそれはなぜか。

背後にある数学

効率

二つの不偏推定量（unbiased estimator）の**相対効率**（relative efficiency）とは、それらの分散の比である。これは、一方が他方に匹敵するには何倍の観測値が必要かを表す。一様分布のデータでは、中央値の平均に対する効率は $\frac{1}{3}$ である。正規分布のデータでは $\frac{2}{\pi} \approx 0.64$ であり、中央値には約 57% 多くの観測値が必要になる。しかし裾の重いデータでは立場が逆転する。コーシー分布（Cauchy distribution、Section 12.22）では平均は役に立たない。n 個の観測値の平均は観測値一つと比べて少しもよくならな

いからである。一方、中央値はうまく機能する。ラプラス分布（Laplace distribution、Section 12.24）では、中央値の効率は平均の 2 倍である。

頑健統計

中央値は**頑健**（robust）である。データのほぼ半分を任意の値に変えても、中央値は大きくは動かない。平均にはそのような防御がなく、誤った値が一つあるだけで台なしになりうる。1960 年代から 1970 年代にかけて、ジョン・テューキー（John Tukey）、Peter Huber らは**頑健統計**（robust statistics）を発展させた。これは、きれいなデータに対しては平均とほぼ同じ効率をもちながら、外れ値（outlier）には強い推定量であり、たとえば**刈り込み平均**（trimmed mean、最大と最小の 10% を捨てて残りを平均したもの）がある。これは飛び込みや体操などの競技の採点に用いられている。

ラプラスと中央値

ラプラス（Laplace）は 1774 年に、誤差が二重指数分布（ラプラス分布）に従うならば、真の値の最良の推定値は中央値であることを示した。ガウス（Gauss）は 1809 年に、誤差が正規分布に従うならば最良の推定値は平均であることを示し、これが最小二乗法（method of least squares）へとつながった。どちらが正しいかは誤差のふるまいによって決まり、これは新しい種類のデータが現れるたびに統計学者がいまなお議論している問題である。

次の一歩

1. **中央値の密度。** 値が一つ $[y, y+\delta y]$ にあり、n 個がそれより小さく、n 個がそれより大きいという数え上げによって、密度 $\frac{(2n+1)!}{(n!)^2} y^n (1-y)^n$ を導け。

2. **三つの観測値。** $n = 1$（一様分布の観測値が三つ）のとき、$\mathrm{Var}(Y)$ と $\mathrm{Var}(\bar{X})$ を正確に求めよ。どちらが小さいか。

3. **ミッドレンジ。**中心のもう一つの推定量に、ミッドレンジ（midrange）$\frac{1}{2}(\max + \min)$ がある。一様分布のデータで観測値が $m = 2n+1$ 個のとき、その分散が $\frac{1}{2(m+1)(m+2)}$ であることを示せ。平均と比べてどうか。（Section 13.3 と比較せよ。）

4. **外れ値。** 11 個の測定値が 9.8、9.9、10.0、10.0、10.1、10.1、10.2、10.2、10.3、10.4、104（入力ミス）である。平均と中央値を比較せよ。

さらに学ぶために

- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。順序統計量と効率について。
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods***（Springer）。頑健な手法がなぜ重要なのかを、専門家でない読者向けに解説している。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***（Harvard University Press）。集約、平均、中央値を歴史的な視点から論じている。

詳しい解説付きの解答は第 13.4 節にある：

https://beyondstep.keithmcnulty.org/ja/ch13/mean-or-median.html

毎週 1 問の STEP の問題を丁寧に解き明かす The STEP Up。メールで無料配信中：keithmcnulty.substack.com