平成19年度計算機支援数学

目標　Mathematica を利用した統計的推測法の比較実験の技術を習得する

題材　独立標本を用いた，位置母数に関する点推定・検定・信頼区間

点推定　以下の３つの推定量の平均絶対誤差、，平均二乗誤差を比較

- 標本平均
- Hodges–Lehmann 推定量
- 中央値

検定　以下の３つの検定の有意水準，検出力を比較

- t–検定：正規母集団を仮定
- Wilcoxon 符号付ランク検定：対称性を仮定
- Tukey の符号検定：中央値における分布関数の連続性のみ仮定

信頼区間　以下の３つの信頼区間の被覆確率を比較

- t–分布に基づく信頼区間
- Wilcoxon 符号付きランク検定の採択域
- Tukey の符号付検定の採択域

比較実験のための母集団分布　標準正規分布，Caucy 分布，指数分布の対数

実験のパラメータ　標本数，中央値の値（検出力の比較ため），有意水準・信頼係数

1　準備

独立標本

$$X_1, X_2, \cdots, X_n : \text{独立に同じ確率分布 (母集団分布) に従う確率変数}$$

母集団分布

標準正規分布

$$\text{確率密度関数}: f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$\mathrm{E}[X] = 0, \quad \text{対称分布}$$

Caucy 分布

$$\text{確率密度関数}: f_2(x) = \frac{1}{\pi(1+x^2)}$$

$$\mathrm{E}[x] = \infty, \quad \text{対称分布}$$

指数分布の対数

$Y \sim g(y) = e^{-y} \quad (y > 0)$ とするときの $X = \log Y - \log\log 2$ の分布

$$\text{確率密度関数}: f_3(x) = \log 2 \exp(x - e^x \log 2)$$

$$\mathrm{E}[X] = -C - \log(\log 2) \approx -0.210703,\ C:\ \text{Euler 定数}, \quad \text{非対称分布}$$

確率分布の中央値　次を満たす θ の値を X の確率分布の中央値と呼ぶ。

$$\Pr\{X \le \theta\} = \Pr\{X \ge \theta\}$$

＊上記３つの母集団分布の中央値（母中央値）は、いずれも 0。

位置・尺度分布族 (Location–scale family)

確率密度関数 $f(x)$ を持つ確率変数 X に対して、$Y = \sigma(X - \theta) \quad (\sigma > 0)$ の確率密度関数は、

$$f(y; \theta, \sigma) := \frac{1}{\sigma} f\left(\frac{y-\theta}{\sigma}\right)$$

となる。ことのき、

$$\mathcal{F} = \{f(y; \theta, \sigma) : -\infty < \theta < \infty, \sigma > 0\}$$

を、f によって生成される位置・尺度分布族と呼び、θ を位置母数、σ を尺度母数と呼ぶ。

2　母中央値の点推定

標本平均

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 任意の母集団分布に対して、母平均の不偏推定量。
- 対称分布はと平均と中央値は一致するので、中央値の推定量として使える。
- 正規分布の平均の最尤推定量

中央値

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}: \text{ 順序統計量}$$
$$\theta_m = \begin{cases} X_{(k+1)} & n = 2k+1 \\ \frac{X_{(k)}+X_{(k+1)}}{2} & n = 2k \end{cases}$$

- 母集団分布型が不明であるとき、中央値の推定として最も良く用いられる。
- はずれ値（極端に大きな、あるいは、小さな値）の影響をもっとも受けにくい

Hodges–Lehmann 推定量

$$\frac{X_i + X_j}{2} \quad (i \le j = 1, 2, \cdots, n) \text{ を小さい順に並べた値を}$$
$$W_{(1)} \le W_{(2)} \le \cdots \le W_{(M)}, \quad M = \frac{n(n+1)}{2}$$

とするとき、

$$\theta_{HL} = \begin{cases} W_{(k+1)} & M = 2k+1 \\ \frac{W_{(k)}+W_{(k+1)}}{2} & M = 2k \end{cases}$$

- 母集団分布が母中央値に関して対称であるときに用いられる。
- 標本平均より、はずれ値の影響を受けにくい。

2.1　実験

上記 3 つの点推定の推定精度を、前述の 3 つの母集団に対して比較する。

推定精度の基準

- 平均 2 乗誤差 (MSE)　$\mathrm{E}[(\hat{\theta} - \theta)^2]$
- 平均絶対誤差 (MAE)　$\mathrm{E}[|\hat{\theta} - \theta|]$

Monte–Carlo Simulation

g : n–変数関数

$\mathrm{E}[g(X_1, \cdots, X_n)]$ の値を推定する方法

手順

(1) 疑似乱数を用いて $X_1, \cdots, X_n$ を生成し、$g(X_1, \cdots, X_n)$ の値を計算する。

(2) (1) を K 回繰り返し、K 個の g の値から、平均値 $\bar{g}$, 標準偏差 $s(g)$ を計算する。

(3) $\bar{g}$ によって、$\mathrm{E}[g(X_1, \cdots, X_n)]$ の値を推定する。このときの推定は、$\frac{2s(g)}{\sqrt{K}}$、すなわち、

$$\mathrm{E}[g(X_1, \cdots, X_n)] = \bar{g} \pm \frac{2s(g)}{\sqrt{K}}$$

比較方法　確率密度関数 $f_i(y; \theta, \sigma)$ を持つ母集団分布の θ の推定量として、標本平均、中央値、Hodges–Lehmann 推定値を用いた時の、MSE, MAE を Monte–Carlo Simulation によって推定し、大小を比較する。

レポート問題１　f_1, f_2, f_3 の選び方、尺度母数、位置母数の値の選び方によって、３つの推定量の優劣（基準値の大小）にどのように変わるかを調べよ。