多標本モデルにおける分布探索による統計解析法

横浜市大・国際総合科学　白石 高章

1　はじめに

ある要因 A があり, k 個の水準 $A_1, \ldots, A_k$ を考える. 水準は群ともよばれる. 水準 A_i における標本の観測値 $(X_{i1}, X_{i2}, \cdots, X_{in_i})$ は第 i 標本または第 i 群とよばれ, 平均 μ_i, 分散 σ^2 である同一の連続型分布関数 $F((x-\mu_i)/\sigma)$ をもつとする. すなわち,

$$P(X_{ij} \leqq x) = F\left(\frac{x-\mu_i}{\sigma}\right), \;\; E(X_{ij}) = \mu_i, \;\; V(X_{ij}) = \sigma^2.$$

さらにすべての X_{ij} は互いに独立であると仮定する. $F(x)$ を未知の分布関数であってもよいとする. 総標本数を $n \equiv n_1 + \cdots + n_k$ とおき, $\nu \equiv \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i$, $\tau_i \equiv \mu_i - \nu$ とおく. このとき $\sum_{i=1}^{k} n_i\tau_i = 0$ である. τ_i は要因 A の水準 A_i における 主効果 (main effect) または相対処理効果 (additive treatment effect), ν を全平均 (overall mean) と呼ばれている. パラメータ τ_i $(i = 1, \cdots, k)$ の点推定として, $F(x)$ が正規分布のときに最良な最小自乗法 $\tilde{\tau}_i \equiv \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}$ $(i = 1, \cdots, k)$, 頑健推定として, 順位推定

$$\hat{\tau}_i = \frac{1}{n}\sum_{i'=1}^{k} n_{i'}\hat{\delta}_{ii'} \quad (i = 1, \cdots, k) \tag{1.1}$$

と M 推定

$$\breve{\tau}_i = \frac{1}{n}\sum_{i'=1}^{k} n_{i'}\breve{\theta}_{ii'} \quad (i = 1, \cdots, k) \tag{1.2}$$

がある, ただし, $\hat{\delta}_{ii'} = (\{X_{ij} - X_{i'j'} : j = 1, \cdots, n_i, j' = 1, \cdots, n_{i'}\}$ の標本中央値), $\hat{\delta}_{ii} = 0$, $i \neq i'$ に対して

$$T_{ii'}(\theta) \equiv \frac{1}{n_i}\sum_{j=1}^{n_i}\psi\left(\frac{X_{ij} - \tilde{\nu}_{ii'} - (n_{i'}/N_{ii'})\cdot\theta}{\breve{\sigma}_n}\right) - \frac{1}{n_{i'}}\sum_{j=1}^{n_{i'}}\psi\left(\frac{X_{i'j} - \tilde{\nu}_{ii'} + (n_i/N_{ii'})\cdot\theta}{\breve{\sigma}_n}\right),$$

$$\tilde{\nu}_{ii'} \equiv \frac{n_i\bar{X}_{i\cdot} + n_{i'}\bar{X}_{i'\cdot}}{n_i + n_{i'}}, \;\; N_{ii'} \equiv n_i + n_{i'}, \;\; \breve{\sigma}_n \equiv \frac{\sqrt{\pi}}{\sqrt{2}n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}|X_{ij} - \bar{X}_{i\cdot}|,$$

とおき, $T_{ii'}(\theta) = 0$ の解を $\breve{\theta}_{ii'}$ とし, $\breve{\theta}_{ii} = 0$ とおく.

これら 3 つの推定量には $F(x)$ により特長がある. $F(x)$ に近い分布を探すことにより, 3 つの推定量の中の 1 つを選択する推定方式を提案し, 新しい推定法式は, 最小自乗法, 順位推定と M 推定よりも安定した良さをもつことを示す. 同様に

$$\text{帰無仮説 } H_0 : \; \mu_1 = \cdots = \mu_k \quad \text{vs. 対立仮説 } H_1 : \; \text{ある } i \neq i' \text{ について } \mu_i \neq \mu_{i'} \tag{1.3}$$

の検定として, F 検定, Kraskal-Wallis 検定, M 検定があり, これらの中の 1 つを選択する検定法が安定した検出力を持つことを示すことは可能である.

2　分布探索

適合度検定は検出力が低いので, 経験分布関数を使って, 観測値の従っている分布に近い分布を探す. $\sum_{i'=0}^{i-1} n_{i'}+1 \leqq m \leqq \sum_{i'=0}^{i} n_{i'}$ となる整数 m と $j=m-\sum_{i'=0}^{i-1} n_{i'}$ に対して, $Z_1,\cdots,Z_n$ を

$$Z_m=(X_{ij}-\bar{X}_{i\cdot})/\sqrt{1-1/n_i},$$

によって定義する, ただし, $n_0=0$ とする. このとき, $E(Z_i)=0$, $V(Z_i)=\sigma^2$ が成り立つ.

$$\hat{G}_n(x)\equiv\frac{1}{n}\#\{i: Z_i\leqq x,\ 1\leqq i\leqq n\}=(Z_1,\cdots,Z_n \text{ の経験分布関数 })$$

とする. $\hat{G}_n(x)$ は, $F((x-\boldsymbol{\tau})/\sigma)$ の不偏かつ一致推定量である. 明細に表現した F_0 に対して,

$$D_{F_0}=\sup_{-\infty<x<\infty}|\hat{G}_n(x)-F_0(x/\breve{\sigma}_n)|,\quad \breve{\sigma}_n=\sum_{i=1}^{n}|Z_i|/\left(n\int_{-\infty}^{\infty}|x|f_0(x)dx\right)$$

とおく.

$$D_{F_0}=\max_{1\leqq i\leqq n}\left[\max\left\{\left|\frac{i}{n}-F_0\left(\frac{Z_{(i)}}{\breve{\sigma}_n}\right)\right|,\ \left|F_0\left(\frac{Z_{(i)}}{\breve{\sigma}_n}\right)-\frac{i-1}{n}\right|\right\}\right].$$

とも表現され, D_{F_0} は観測値の従っている分布関数 F と明細に与えた F_0 の距離である. F_0 として対称な分布として, 標準正規分布 $N(0,1)$, 汚れた正規分布 $CN(\varepsilon)=(1-\varepsilon)N(0,1/(1+8\varepsilon))+\varepsilon N(0,9/(1+8\varepsilon))$, ロジスティック分布 $LG(0,\sqrt{3}/\pi)$, 両側指数分布 $DE(0,1/\sqrt{2})$ を選ぶ. 非対称な分布として指数分布, 対数正規分布, ワイブル分布, 非対称な汚れた正規分布 $ACN=0.98N(-0.1,1/1.47)+0.02N(4.9,0.01/1.47)$ を選ぶ. D_{F_0} を最小にする F_0 を探す. $\breve{\sigma}_n$ を計算するために, $\int_{-\infty}^{\infty}|x|f_0(x)dx$ の値が必要とされる.

分布の特徴をみる指標として, 歪度: $\ell_1=\int_{-\infty}^{\infty}x^3dF(x)$ と尖度: $\ell_2=\int_{-\infty}^{\infty}x^4dF(x)-3$ がある. SAS system や Microsoft Excel では, 歪度 ℓ_1 と尖度 ℓ_2 は

$$\begin{aligned}\hat{\ell}_1 &= \frac{n\sqrt{n-1}}{n-2}\cdot\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})^3}{\{\sum_{i=1}^{n}(Z_i-\bar{Z})^2\}^{3/2}},\\ \hat{\ell}_2 &= \frac{n(n+1)(n-1)}{(n-2)(n-3)}\cdot\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})^4}{\{\sum_{i=1}^{n}(Z_i-\bar{Z})^2\}^2}-\frac{3(n-1)^2}{(n-2)(n-3)}.\end{aligned}$$

によって推定されている.

D_{F_0}, $\hat{\ell}_1$, $\hat{\ell}_2$ の値を使って, $(\tau_1,\cdots,\tau_k)$ の推定量として, $(\tilde{\tau}_1,\cdots,\tilde{\tau}_k)$, $(\hat{\tau}_1,\cdots,\hat{\tau}_k)$, $(\breve{\tau}_1,\cdots,\breve{\tau}_k)$ のいずれかを選択する手法について, シミュレーションによる結果を当日報告する.

参考文献

Shiraishi, T. (1990). R-estimators and confidence regions in one-way MANOVA. J. Statist. Plan. Infer., 24, p203-214.

Shiraishi, T. (1996). On scale-invariant M-statistics in multivariate k samples. J. Japan Statist. Soc., 26, p241-253.