数理解析研究所講究録
645巻 1988年 114-125

二元配置におけるノンパラメトリックT－法

阪大 基礎工　松田眞一 (Shin-ichi Matsuda)

1．　はじめに

T－法とは、Tukey [7] が提案した多重比較問題に対する一方法である。この方法は、データが正規分布に従っているときのみ有効であったので、Sen [3] は適用範囲を広げるべく、順位を使ったノンパラメトリックT－法を提案した。この2つの方法は共に、一元配置のデータを対象としている。

二元配置のデータに対して、Sen [4] はアラインメントという方法を使ってノンパラメトリックT－法を構築している。ここでは、欠測値を含む場合にも適用できる、別の方法によるノンパラメトリックT－法を提案する。

2．　モデル

p（≧3）個の処理，n個のブロックからなる二元配置 $\{X_{ij} : i=1,2,\ldots,n, j=1,2,\ldots,p\}$ を考える。欠測値を許すためj番目の処理において観測値のあるブロックの番号の集合を D_j と記すことにする。観測値は次のモデルに従っているとする。

(1)　　$X_{ij} = \mu_0 + \alpha_i + \tau_j + \varepsilon_{ij}$　　for $j=1,2,\ldots,p,\ i\in D_j$

ここで、右辺はすべて未知の値で、μ_0 は平均効果、$\alpha_1,\alpha_2,\ldots,\alpha_n$ はブロック効果（母数、または確率変数）、$\tau_1,\tau_2,\ldots,\tau_p$ は処理効果（興味のある母数、$\Sigma\tau_j = 0$ と仮定する）、$\varepsilon_{ij}, j=1,2,\ldots,p, i\in D_j$ は誤差成分である。また、$\{\varepsilon_{ij} : j=1,\ldots,p,\ i\in D_j\}$ はi．i．d．で、未知の連続分布 F(x) に従っているとする。

この時、検定したい帰無仮説は、

(2) $H : \tau_1 = \tau_2 = \cdots = \tau_p = 0$

である。

便宜上、次のようにおく。

$$D_{jk} = D_j \cap D_k \ , n_{jk} = \# D_{jk} \qquad \text{for } 1 \leqq j < k \leqq p$$

また、次を仮定する。

$$n_{jk} > 0, \quad n_{jk}/n \to d_{jk} > 0 \quad \text{as } n \to \infty$$

3. 統計量の構成

j, k番目の処理 ($1 \leqq j < k \leqq p$) に対し、

(3) $$Y_{jk}^{(i)} = (X_{ij} - X_{ik}) /2$$
$$= \Delta_{jk}/2 + e_{jk}^{(i)} \qquad \text{for } i \in D_{jk}$$

とおく。但し、

(4) $$\Delta_{jk} = \tau_j - \tau_k \ , \quad e_{jk}^{(i)} = (\varepsilon_{ij} - \varepsilon_{ik}) /2$$

ここで、$e_{jk}^{(i)}$ は $\{\varepsilon_{ij}\}$ に対する仮定により i, j, kによらない共通の分布に従う。これを $G(x)$ とおく。この分布は、連続で原点について対称となる。

そこで、絶対順位を用いた統計量を構成しよう。スコア E_{na} を

(5) $$E_{na} = J_n(a/(n+1)) \qquad \text{for } a = 1,2,\ldots,n$$

で定義する。但し、J_n は Sen and Puri [5] と同様な次の仮定を満たしているとする。

<u>仮定 1</u> ある定数でない関数 $J(u)$ が存在し、$0 < u < 1$ に対して

$$\lim_{n\to\infty} J_n(u) = J(u) \text{ で、 } J(0) = 0 \text{ 。}$$

<u>仮定 2</u>

$$\int_0^\infty \left[J_{n_{jk}}\left(\frac{n_{jk}}{n_{jk}+1} H_{jk}^{(n_{jk})}(x)\right) - J\left(\frac{n_{jk}}{n_{jk}+1} H_{jk}^{(n_{jk})}(x)\right)\right] dG_{jk}^{(n_{jk})}(x)$$
$$= o_p(n^{-1/2})$$
$$\text{for } 1 \leqq j < k \leqq p$$

但し、

$G_{j,k}^{(n_k)}(x) = \#\{ i \in D_{jk} : Y_{j,k}^{(i)} \leqq x \}/n_{jk}$: 経験分布関数,

$H_{j,k}^{(n_k)}(x) = G_{j,k}^{(n_k)}(x) - G_{j,k}^{(n_k)}(-x-) \qquad (x \geqq 0)$

仮定 3 $J(u)$ は絶対連続で、ある $K < \infty$, $h > 0$ に対し、

$$|J^{(i)}(u)| = |d^i J(u)/du^i| \leqq K[u(1-u)]^{h-i-1/2} \quad \text{for } i = 0,1$$

さらに、次のようなn次元ベクトル空間上の関数 $h(W)$ を定義する。

(6) $$h(W) = \left(\sum_{a=1}^{n} E_{nR_a} Z_a\right)/n$$

但し、 $W = (W_1, W_2, \ldots, W_n)'$,

R_a : $|W_a|$ の $\{|W_i| : i=1,2,\ldots,n\}$ における順位,

$$Z_a = \begin{cases} 1 & \text{if } W_a > 0 \\ 0 & \text{otherwise} \end{cases}$$

この関数によって、統計量を次のように定義する。

(7) $$T_{j,k}^{(n_k)} = h(Y_{jk})$$

但し、Y_{jk} は $\{Y_{j,k}^{(i)} : i \in D_{jk}\}$ を並べたベクトル。

4. 統計量の性質

まず、Sen and Puri [5] の定理4．1より次を得る。

補題 1 $N(T_n - \mu)$ は漸近的に平均0のp′次元多変量正規分布に従う。

但し、

(8) $$N = \mathrm{diag}(n_{12}^{1/2}, \ldots, n_{p-1p}^{1/2}),$$

$$T_n = (T_{1,2}^{(n_{12})}, \ldots, T_{1,p}^{(n_{1p})}, T_{2,3}^{(n_{23})}, \ldots\ldots, T_{p-1,p}^{(n_{p-1p})})' ,$$

$$\mu = (\mu_{12}, \ldots, \mu_{1p}, \mu_{23}, \ldots\ldots, \mu_{p-1p})' ,$$

$$\mu_{jk} = \int_0^\infty J[H(x;\Delta_{jk}/2)]\, dG(x - \Delta_{jk}/2),$$

$$H(x;\delta) = G(x-\delta) - G(-x-\delta) \qquad (x \geqq 0),$$

$$p' = p(p-1)/2$$

証明　Sen and Puri [5] の定理４．１より次が分かる。

$$T_{j,k}^{(n_{jk})} - \mu_{jk} \sim \left(\sum_{i \in D_{jk}} B_{j,k}^{(i)}\right) / n_{jk} \quad \text{for } 1 \leqq j < k \leqq p$$

但し、$B_{j,k}^{(i)}$,$1 \leqq j < k \leqq p, i \in D_{jk}$ は i が異なる間では独立同分布で、すべて平均０，分散有限の確率変数である。ここで、$\sim$ は漸近的確率同等を示す。

この時、任意の $\{a_{jk} : 1 \leqq j < k \leqq p\}$ (not all zero) に対し、

$$\sum_{j,k} a_{jk} \left(T_{j,k}^{(n_{jk})} - \mu_{jk}\right) \sim \sum_{j,k} \left(\sum_{i \in D_{jk}} a_{jk} B_{j,k}^{(i)}\right) / n_{jk}$$

となる。そこで、

$$C_i = \sum_{j,k} \frac{a_{jk} n}{n_{jk}} \tilde{B}_{j,k}^{(i)}$$

但し、

$$\tilde{B}_{j,k}^{(i)} = \begin{cases} B_{j,k}^{(i)} & \text{if } i \in D_{jk} \\ 0 & \text{otherwise} \end{cases}$$

とおくと、

$$\sum_{j,k} \left(\sum_{i \in D_{jk}} a_{jk} B_{j,k}^{(i)}\right) / n_{jk} = \sum_{i \in \cup D_{jk}} C_i / n$$

と書ける。

C_i はその定義より、$\tilde{B}_{j,k}^{(i)}$ の組合わせによる高々$2^{p'}$通りの確率変数にしかなり得ない。そこで、$n_{jk} \neq 0$ であることを用いると、ある K_1, K_2 に対し、

$$K_1 > \mathrm{Var}\, C_i > K_2 > 0 \qquad \text{for } i \in \cup D_{jk}$$

が分かる。仮定より$O(n_{jk}) = O(n)$ であることを用いると、

$$\mathrm{Var}\left(\sum_{i \in \cup D_{jk}} C_i\right) = O(n)$$

を得る。

一方、任意の $\varepsilon > 0, s > 0$ に対し $d_{jk} > 0$ より、

$$\lim_{n \to \infty} \int_{\{|C_i| \geqq \varepsilon s\}} C_i^2 dP = \int_{\{|\bar{C}_i| \geqq \varepsilon s\}} \bar{C}_i^2 dP$$

但し、

$$\bar{C}_i = \sum_{j,k} \frac{a_{jk}}{d_{jk}} \tilde{B}^{(i)}_{j,k}$$

が成り立ち、さらに s については単調減少なので、

$$\int_{\{|C_i| \geqq \varepsilon s_n\}} C_i^2 dP \to 0 \qquad \text{as } n \to \infty$$

但し、

$$s_n^2 = \mathrm{Var}\left(\sum_{i \in \cup D_{jk}} C_i\right)$$

が成立する。

ここで、前と同じ議論により C_i は高々$2^{p'}$通りしかパターンがないので、

$$\max_{i \in \cup D_{jk}} \int_{\{|C_i| \geqq \varepsilon s_n\}} C_i^2 dP \to 0 \qquad \text{as } n \to \infty$$

ゆえに、

$$\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{i \in \cup D_{jk}} \int_{\{|C_i| \geqq \varepsilon s_n\}} C_i^2 dP$$

$$\leqq \lim_{n \to \infty} O(1) \cdot \max_{i \in \cup D_{jk}} \int_{\{|C_i| \geqq \varepsilon s_n\}} C_i^2 dP$$

$$= 0$$

よって、Lindeberg condition を満たすことになるので、中心極限定理により

$n^{-1/2} \sum_{i \in \cup D_{jk}} C_i$ は漸近的に平均０の正規分布に従う。

従って、$N(T_n - \mu)$ は漸近的に平均０のｐ′次元多変量正規分布に従う。 □

今、次のような対立仮説の列を考えよう。

ある固定した λ_{jk} $(1 \leqq j < k \leqq p)$ に対し、

(9) $\quad K_n : \Delta_{jk} = \lambda_{jk} / n^{1/2} \quad \text{for } 1 \leqq j < k \leqq p$

そのとき、次が分かる。

補題　2　補題1の条件が満たされているとき対立仮説列 $\{K_n\}$ の下で、$N(T_n - \mu)$ は漸近的に平均0，分散行列 Σ のp′次元多変量正規分布に従う。

但し、

$$\Sigma = (\sigma_{jk,j'k'}) : 1 \leqq j < k \leqq p, 1 \leqq j' < k' \leqq p,$$

$$\sigma_{jk,j'k'} = \begin{cases} \int_0^1 J^{*2}(x)dx \,/4 & \text{if } j = j', k = k' \\ \dfrac{\#(D_{jk} \cap D_{j'k'})}{4(n_{jk}\, n_{j'k'})^{1/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} J^*(G(x))J^*(G(y))dG_{jk,j'k'}(x,y) & \text{otherwise} \end{cases}$$

ここで、

$$J^*(x) = \begin{cases} J(2x-1) & \text{if } x \geqq 1/2 \\ -J(1-2x) & \text{otherwise.} \end{cases}$$

$G_{jk,j'k'}(x,y)$ ： $(e_{j,k}^{(n)}, e_{j',k'}^{(n)})$ の分布関数

証明は、Sen and Puri [5] の系4．1と同様である。

よって、Hocheberg [1] と同様にして次を得る。

定理　1

(10)　$$\lim_{n\to\infty} P_H\{|T_{j,k}^{(n)} - \mu| \leqq n_{jk}^{-1/2} A(J^*)\, |m|_{p',\alpha}/2,\ 1 \leqq j < k \leqq p\} \geqq 1 - \alpha$$

但し、

$$\mu = \int_0^\infty J[H(x;0)]dG(x) = \int_{1/2}^1 J^*(x)dx = \int_0^1 J(x)dx/2,$$

$$\{A(J^*)\}^2 = \int_0^1 J^{*2}(x)dx - \left(\int_0^1 J^*(x)dx\right)^2$$

で、$|m|_{p',\alpha}$ は標準正規分布からの大きさp′の標本の最大絶対値の分布の上側α点。

証明　H の下では $\mu_{jk} = \mu$ であるから補題2より

$$N(T_n - \mu I) \to N_{p'}(0, \Sigma) \quad \text{in law}$$

但し、$I = (1,\ldots,1)'$ 。

そこで、Šidák [6] の系1を用い、

$$\lim_{n\to\infty} P_H\{|T_{j,k}^{(n_{jk})} - \mu| \leqq n_{jk}^{-1/2} A(J^*) |m|_{p,\alpha}/2 \ , 1 \leqq j < k \leqq p\}$$

$$\geqq \prod_{1\leqq j<k\leqq p} \lim_{n\to\infty} P_H\{|T_{j,k}^{(n_{jk})} - \mu| \leqq n_{jk}^{-1/2} A(J^*) |m|_{p,\alpha}/2\}$$

$$= 1 - \alpha$$

よって、成り立つ。 □

この定理によって、検定は次のようになされる。

与えられたデータに対して求めた $T_{j,k}^{(n_{jk})}$ が、

$$|T_{j,k}^{(n_{jk})} - \mu| > n_{jk}^{-1/2} A(J^*) |m|_{p,\alpha}/2$$

となるようなj，kの組は有意な差があるとみなす。

なお、検定の際必要な $|m|_{p,\alpha}$ の値に対する表は十分なものがない。コンピューターによって得た値を表．1に掲げておく。

5．漸近的相対効率

まず、Sen [4] の方法について述べる。Sen の方法では、アラインメントをブロック全体で考え、ブロック平均を引くことでブロック効果をなくそうとしている。実際には、

$$Y_{ij} = X_{ij} - \sum_{k=1}^{p} X_{ik}/p,$$

$$\tilde{e}_{ij} = \varepsilon_{ij} - \sum_{k=1}^{p} \varepsilon_{ik}/p \qquad \text{for} \quad i=1,2,..,n, j=1,2,..,p$$

とおくと、

$$Y_{ij} = \tau_j + \tilde{e}_{ij}$$

となって他の効果が消える。ここで、$\tilde{e}_{ij}$ は ε_i に対する仮定によりi，jによ

らない共通の分布に従う。これを $\tilde{G}(x)$ とおく。

スコア $\tilde{E}_{na}$ $(a = 1,2,\ldots,n)$ を

$$\tilde{E}_{na} = \tilde{J}_n(a/(n+1))$$

で定義する。但し、$\tilde{J}_n$ は Sen [4] の条件を満たすとする。

この時、次のようなn次元ベクトル空間の直積上の関数 $\tilde{h}(U,V)$ を定義する。

$$\tilde{h}(U,V) = (\sum_{a=1}^{n} \tilde{E}_{2nQa}) /n$$

但し、

$U = (U_1,U_2,\ldots,U_n)'$,

$V = (V_1,V_2,\ldots,V_n)'$,

Q_a ：U_a の $\{U_i,V_i : i = 1,2,\ldots,n\}$ における順位

この関数によって統計量を次のように定義する。

$$\tilde{T}^{(n)}_{j,k} = \tilde{h}(Y_j ,Y_k) \qquad \text{for} \quad 1 \leqq j < k \leqq p$$

但し、

$$Y_j = (Y_{1j} ,Y_{2j} ,\ldots,Y_{nj})' , \quad Y_k = (Y_{1k} ,Y_{2k} ,\ldots,Y_{nk})'$$

Sen [4] で次が分かっている。

補題 3

$$\lim_{n\to\infty} P_H\{ | \tilde{T}^{(n)}_{j,k} - \tilde{\mu} | \leqq n^{-1/2}(\frac{p}{p-1})^{1/2} A(\tilde{J}) R_{p\alpha} /2 \ , 1 \leqq j < k \leqq p\} \geqq 1 - \alpha$$

但し、

$$\tilde{J}(x) = \lim_{n\to\infty} \tilde{J}_n(x),$$

$$\tilde{\mu} = \int_0^1 \tilde{J}(x)dx,$$

$A(.)$ ：定理1で定義したもの，

$R_{p\alpha}$ ：標準正規分布からの大きさpの標本の範囲の分布の上側α点

そこで、欠測値のない二元配置に対し（即ち、$n_{jk} = n$ for $1 \leqq j < k \leqq p$）、提案した方法の Sen の方法に対する漸近的相対効率（ＡＲＥ）を調べよう。比較には Hodges and Lehmann [2] の推定量を用いる。

Hodges and Lehmann の推定量は以下のように構成される。

$$\Delta^*_{jk} := \sup\{\Delta : h(Y_{jk} - \Delta I/2) > \bar{E}_n/2\},$$

$$\Delta^{**}_{jk} := \inf\{\Delta : h(Y_{jk} - \Delta I/2) < \bar{E}_n/2\},$$

$$\hat{\Delta}_{jk} := (\Delta^*_{jk} + \Delta^{**}_{jk})/2,$$

$$\hat{\lambda}_{jk} := n^{1/2}\hat{\Delta}_{jk}$$

但し、

$$\bar{E}_n = \left(\sum_{a=1}^{n} E_{na}\right)/n,$$

I：定理１で定義したもの

この推定量に対し次が分かる。

定理　２　適当な条件の下で次が成り立つ。

$$\lim_{n\to\infty} P_{(Kn)}\{|\hat{\lambda}_{jk} - \lambda_{jk}| \leqq 2\frac{A(J^*)}{B(J^*,G)}|m|_{p,\alpha}, 1 \leqq j < k \leqq p\} \geqq 1 - \alpha$$

但し、

$$B(J^*,G) = \int_{-\infty}^{\infty} \frac{d}{dx} J^*(G(x))dG(x)$$

Sen の場合も同様の推定量 $\hat{\tilde{\lambda}}_{jk}$ に対し次の定理を得る。

定理　３　適当な条件の下で、次が成り立つ。

$$\lim_{n\to\infty} P_{(Kn)}\{|\hat{\tilde{\lambda}}_{jk} - \lambda_{jk}| \leqq \left(\frac{p}{p-1}\right)^{1/2}\frac{A(\tilde{J})}{B(\tilde{J},\tilde{G})}R_{p\alpha}, 1 \leqq j < k \leqq p\} \geqq 1 - \alpha$$

よって信頼区間にもとずくＡＲＥ（Pitman 効率）は、

$$\frac{p}{4(p-1)}\left(\frac{A(\tilde{J})\ B(J^*,G)\ R_{p\alpha}}{A(J^*)\ B(\tilde{J},\tilde{G})\ |m|_{p,\alpha}}\right)^2$$

となる。

Sen の方法で用いるスコア関数 $\tilde{J}$ の極限は J^* と同じとし、分布関数 F(x) は密度 f(x) を持つとする。

次の場合についてＡＲＥを調べた。

・f(x) ：正規分布、一様分布、両側指数分布、Cauchy 分布。

・スコア：正規スコア、Wilcoxon 型スコア。

・α = 0.01, 0.05; p = 3,...,8

そのうち Wilcoxon 型スコアの結果が表．２である。これらの組合せに対し求めたＡＲＥは Cauchy 分布を除いてだいたい .9～1 である。従って、欠測値が殆どない場合以外では提案した方法が有力と思われる。

表．１　$|m|_{p,\alpha}$

		α = 0.05	α = 0.01
3	3	2.38774	2.93416
4	6	2.63104	3.14276
5	10	2.79963	3.28926
6	15	2.92780	3.40165
7	21	3.03074	3.49253
8	28	3.11648	3.56861

表. 2 ARE ($J^*(x) = 2x - 1$)

	Normal	D.exp.	Uniform	Cauchy
p = 3	.9632 .9858	.9812 1.0043	.9711 .9939	1.2842 1.3144
4	.9533 .9814	.9480 .9759	.9679 .9964	1.4300 1.4721
5	.9495 .9792	.9215 .9503	.9659 .9961	1.5192 1.5667
6	.9473 .9778	.9008 .9298	.9631 .9941	1.5789 1.6297
7	.9466 .9770	.8848 .9133	.9606 .9915	1.6227 1.6748
8	.9457 .9765	.8715 .8999	.9575 .9887	1.6549 1.7088

上段 : $\alpha = 0.05$

下段 : $\alpha = 0.01$

References

[1] Hocheberg,Y.(1974). Some generalizations of the T-method in simultaneous inference. J.Multivar.Anal. 4,224-234.

[2] Hodges,J.L.,Jr. and Lehmann,E.L.(1963). Estimates of location based on rank tests. Ann.Math.Statist. 34, 598 - 611.

[3] Sen,P.K.(1966). On nonparametric simultaneous confidence regions and tests for the one criterion analysis of variance problem. Ann.Inst. Statist.Math. 18, 319 - 336.

[4] Sen,P.K.(1969). On nonparametric T-method of multiple comparisons for randomized blocks. Ann.Inst.Statist.Math. 21, 329 - 333.

[5] Sen,P.K. and Puri,M.L.(1967). On the theory of rank order tests for location in the multivariate one sample problem. Ann.Math.Statist. 38, 1216 - 1228.

[6] Šidák,Z.(1967). Rectangular confidence regions for the means of multivariate normal distributions. J.Amer.Statist.Assoc. 62,626-633.

[7] Tukey,J.W.(1953). The problem of multiple comparisons, Unpublished manuscript, Princeton University.