

On a non-parametric discriminant procedure

阪大 基礎工 山口 光 代

2変数の2標本 location問題に対して, S. K. Chatterjee
と P. K. Sen (1964) が提供したある rank-sum statistics
を判別問題に応用して, 判別関数を作ることを試み, その場
合の判別方程式の consistency が成立していることを示し
た。又, 標本の大きさについての考察を加えた。

§1. 序

次のような2変量判別問題を考える。 $\pi_i (i=1, \dots, k)$ は,
 k 個の異った母集団とする。 (X_i, Y_i) は, それぞれ, 命令関
数 $F_i(x, y)$ とする確率変数の k 個の組とする。 $F_i(x, y)$ は,
 π_i における命令関数を表わし, 連続関数でありとする。各 i
($i=1, \dots, k$) について, この $F_i(x, y)$ については, 何等の知
識がなく, 母集団 π_i から取り出されたことから, 事前に判って
いる大きさ m_i の標本値を持つだけとする。この標本値を
 $(x_{i1}, y_{i1}), \dots, (x_{im_i}, y_{im_i})$ とする。そして, 他の母集団

π_0 において、連続な分布関数 $F_0(x, y)$ を持つと仮定している確率変数 (X_0, Y_0) がある。この母集団 π_0 は、 $\pi_1, \dots, \pi_k$ の k 個の母集団のうちの一つと一致していることが解っている。 π_0 から、又 $2n_0$ の標本組を得るとして、その標本組を、 $(x_0, y_0), \dots, (x_{n_0}, y_{n_0})$ とする。これらの得られた観測値の組と比較して、 π_0 は、どの母集団と一致しているかを判定しない。

このような、non-parametric な判別問題は、最初、Felix と Hodges (1951, 1952) によって試みられた。

Stoller (1954) は、1変数の場合の2母集団への判別を行っている。Das Gupta (1964) は、 p 変数の場合の k 個の母集団への判別を、経験分布関数の間の、Kolmogorov の距離を用いて行ない、その判別方法の consistency を示している。多変数の場合、tolerance region を用いたことを試みている Quessenberry と Gessaman (1968) の論文、そして、1変数2母集団への判別で、2種の判別誤差の最大のものである、与えた他でお22のような sequential な方法を導入した Woinsky と Kurz (1969) の論文は興味深いものである。その他にも、Ryzin (1966), Hudimoto (1963, 1964, 1968), Peltó (1969) 等、いくつかの論文がある。

§2. 判別方式.

母集団 π_i ($i=1, \dots, k$) にあつる分布関数 $F_i(x, y)$ は、次の条件をみたすものとす。

(1) $F_1(x, y), \dots, F_k(x, y)$ は、連続関数である。

(2) ある単調関数 $g(x)$ が存在して、 $Y_i = g(X_i), \dots, Y_k = g(X_k)$ とする確率は、それぞれ 0 である。

(3) $i \neq j = 1, \dots, k$ に対して、

$$P(X_i > X_j) \neq P(X_i \leq X_j)$$

$$P(Y_i > Y_j) \neq P(Y_i \leq Y_j)$$

観測値の大きさ $n_0, n_1, \dots, n_k$ は、 $N_i = n_0 + n_i$ ($i=1, \dots, k$) とすると、 $n_0/N_i \rightarrow X_i$ ($0 < X_i < 1$) ($i=1, \dots, k$) を条件をみたすように大きくするとす。

各 i ($=1, \dots, k$) に対して、観測値の組、

$$\pi_0: (x_{01}, y_{01}), \dots, (x_{0n_0}, y_{0n_0})$$

$$\pi_i: (x_{i1}, y_{i1}), \dots, (x_{in_i}, y_{in_i})$$

を比較する。 $(x_{01}, \dots, x_{0n_0}, x_{i1}, \dots, x_{in_i})$ を大きさの順に並べ、その順位を、 $(r_{i1}, \dots, r_{in_i}, r_{n_0+i}, \dots, r_{n_0+i+n_i})$ とする。同様に $(y_{01}, \dots, y_{0n_0}, y_{i1}, \dots, y_{in_i})$ を大きさの順に並べ、その順位を、 $(s_{i1}, \dots, s_{in_i}, s_{n_0+i}, \dots, s_{n_0+i+n_i})$ とする。

$$S_{1i} = \sum_{\alpha=1}^{n_0} r_{\alpha i}, \quad S_{2i} = \sum_{\alpha=1}^{n_0} s_{\alpha i}$$

$$C_i = \frac{12}{N_i(N_i^2-1)} \sum_{\alpha=1}^{N_i} \left(x_{\alpha i} - \frac{N_i+1}{2} \right) \left(x_{\alpha i} - \frac{N_i+1}{2} \right)$$

とおく.

$$R_i = \frac{12}{(1-C_i^2) n_o n_i (N_i+1)} \left[\left(S_{1i} - \frac{n_o(N_i+1)}{2} \right)^2 + \left(S_{2i} - \frac{n_o(N_i+1)}{2} \right)^2 - 2C_i \left(S_{1i} - \frac{n_o(N_i+1)}{2} \right) \left(S_{2i} - \frac{n_o(N_i+1)}{2} \right) \right]$$

とおく. この統計量に基づいて.

$$R_i = \min_{1 \leq j \leq k} R_j \quad \text{ならば} \quad \pi_o = \pi_i$$

と判別する判別方式を考える。

§3. 判別方式の consistency

$$A(x) = \begin{cases} 1 & : x > 0 \\ 0 & : x = 0 \\ -1 & : x < 0 \end{cases} \quad \text{とする.}$$

$$M_{j1}(1,1) = \sum_{\alpha}^{n_o} \sum_{\beta}^{n_j} A(x_{\alpha\alpha} - x_{j\beta})^2 / n_o n_j$$

$$M_{j2}(1,1) = \sum_{\alpha \neq \alpha'}^{n_o} \sum_{\beta}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(x_{\alpha'\alpha} - x_{j\beta}) / n_o(n_o-1) n_j$$

$$M_{j3}(1,1) = \sum_{\alpha}^{n_o} \sum_{\beta \neq \beta'}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(x_{\alpha\alpha} - x_{j\beta'}) / n_o n_j (n_j-1)$$

$$M_{j4}(1,1) = \sum_{\alpha \neq \alpha'} \sum_{\beta \neq \beta'}^{n_0} \sum_{\beta'}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(x_{\alpha\alpha'} - x_{j\beta'}) / n_0 n_j (n_0 - 1) (n_j - 1)$$

$$M_{j1}(2,2) = \sum_{\alpha} \sum_{\beta}^{n_0} \sum_{\beta'}^{n_j} A(y_{\alpha\alpha} - y_{j\beta})^2 / n_0 n_j$$

$$M_{j2}(2,2) = \sum_{\alpha \neq \alpha'} \sum_{\beta}^{n_0} \sum_{\beta'}^{n_j} A(y_{\alpha\alpha} - y_{j\beta}) A(y_{\alpha\alpha'} - y_{j\beta'}) / n_0 (n_0 - 1) n_j$$

$$M_{j3}(2,2) = \sum_{\alpha} \sum_{\beta \neq \beta'}^{n_0} \sum_{\beta'}^{n_j} A(y_{\alpha\alpha} - y_{j\beta}) A(y_{\alpha\alpha} - y_{j\beta'}) / n_0 n_j (n_j - 1)$$

$$M_{j4}(2,2) = \sum_{\alpha \neq \alpha'} \sum_{\beta \neq \beta'}^{n_0} \sum_{\beta'}^{n_j} A(y_{\alpha\alpha} - y_{j\beta}) A(y_{\alpha\alpha'} - y_{j\beta'}) / n_0 n_j (n_0 - 1) (n_j - 1)$$

$$M_{j1}(1,2) = \sum_{\alpha} \sum_{\beta}^{n_0} \sum_{\beta'}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(y_{\alpha\alpha} - y_{j\beta'}) / n_0 n_j$$

$$M_{j2}(1,2) = \sum_{\alpha \neq \alpha'} \sum_{\beta}^{n_0} \sum_{\beta'}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(y_{\alpha\alpha'} - y_{j\beta'}) / n_0 (n_0 - 1) n_j$$

$$M_{j3}(1,2) = \sum_{\alpha} \sum_{\beta \neq \beta'}^{n_0} \sum_{\beta'}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(y_{\alpha\alpha} - y_{j\beta'}) / n_0 n_j (n_j - 1)$$

$$M_{j4}(1,2) = \sum_{\alpha \neq \alpha'} \sum_{\beta \neq \beta'}^{n_0} \sum_{\beta'}^{n_j} A(x_{\alpha\alpha} - x_{j\beta}) A(y_{\alpha\alpha'} - y_{j\beta'}) / n_0 n_j (n_0 - 1) (n_j - 1)$$

$\varepsilon \propto \varepsilon$ R_j は.

$$R_j = \frac{3}{(1 - \varepsilon_j^2) n_0 n_j (N_j + 1)} \left[n_0 n_j (M_{j1}(1,1) + M_{j1}(2,2) - 2 \varepsilon_j M_{j1}(1,2)) \right]$$

$$\begin{aligned}
& + n_0(n_0-1)n_j (M_{j2}(1,1) + M_{j2}(2,2) - 2C_j M_{j2}(1,2)) \\
& + n_0 n_j (n_j-1) (M_{j3}(1,1) + M_{j3}(2,2) - 2C_j M_{j3}(1,2)) \\
& + n_0 n_j (n_0-1)(n_j-1) (M_{j4}(1,1) + M_{j4}(2,2) - 2C_j M_{j4}(1,2))] \\
& \text{と、U-統計量を用いて書き直せる。U-統計量の性質より、}
\end{aligned}$$

$$A_{j1} = M_{j1}(1,1) + M_{j1}(2,2) - 2C_j M_{j1}(1,2), A_{j2} = M_{j2}(1,1) + M_{j2}(2,2) - 2C_j M_{j2}(1,2), A_{j3} = M_{j3}(1,1) + M_{j3}(2,2) - 2C_j M_{j3}(1,2),$$

$$A_{j4} = M_{j4}(1,1) + M_{j4}(2,2) - 2C_j M_{j4}(1,2) \text{ である。これより、}$$

$$\bar{A}_{ji} = 2 \left(1 - \bar{F}_j \int_{E^2} (2F_j(x, \infty) - 1)(2F_j(\infty, y) - 1) dF_0(x, y) \right),$$

$$\bar{A}_{j2} = \int_{E^1} (2F_0(x, \infty) - 1)^2 dF_j(x, \infty) - 2\bar{F}_j \int_{E^2} (2F_0(x, \infty) - 1)(2F_0(\infty, y) - 1)$$

$$dF_j(x, y) + \int_{E^1} (2F_0(\infty, y) - 1)^2 dF_j(\infty, y),$$

$$\bar{A}_{j3} = \int_{E^1} (2F_j(x, \infty) - 1)^2 dF_0(x, \infty) + \int_{E^1} (2F_j(\infty, y) - 1)^2 dF_0(\infty, y)$$

$$- 2\bar{F}_j \int_{E^2} (2F_j(x, \infty) - 1)(2F_j(\infty, y) - 1) dF_0(x, y),$$

$$\begin{aligned}
\bar{A}_{j4} &= \left[\int_{E^1} (2F_j(x, \infty) - 1) dF_0(x, \infty) - \bar{F}_j \int_{E^1} (2F_j(\infty, y) - 1) dF_0(\infty, y) \right]^2 \\
&+ \left[\int_{E^1} (2F_j(\infty, y) - 1) dF_0(\infty, y) \right]^2 (1 - \bar{F}_j^2)
\end{aligned}$$

に、 $\Pi_0 = \Pi_1$ の下で、確率収束することがいえる。ここで、

C_j については、Chatterjee と Sen の論文 [9] で、

$$\bar{F}_j = 3 \int_{\mathbb{R}^2} (2\bar{F}_j(x, \infty) - 1)(2\bar{F}_j(\infty, y) - 1) d\bar{F}_j(x, y)$$

ただし, $\bar{F}_j(x, y) = X_j F_0(x, y) + (1 - X_j) F_j(x, y)$ である。
 とすると, C_j は $\bar{F}_j$ に確率収束することの証明が出来る。
 ここで, $\bar{F}_j = \pm 1$ となるのは, 単調関数 $g(x)$ が存在して,
 $Y_j = g(X_j)$, $Y_0 = g(X_0)$ と書ける場合, かつ, この場合に限る。
 以上により, 条件 (2), (3) の下では,

$$R_j = \frac{3A_{j1}}{(1-C_j^2)(N_j+1)} + \frac{3(n_0-1)A_{j2}}{(1-C_j^2)(N_j+1)} + \frac{3(n_j-1)A_{j3}}{(1-C_j^2)(N_j+1)} \\ + \frac{3(n_0-1)(n_j-1)}{(1-C_j^2)(N_j+1)} A_{j4}$$

の第1項が0, 第2項が $\frac{3\bar{A}_{j2} X_j}{(1-\bar{F}_j^2)}$, 第3項が $\frac{3\bar{A}_{j3} (1-X_j)}{(1-\bar{F}_j^2)}$
 第4項は $\frac{3\bar{A}_{j4}}{(1-\bar{F}_j^2)}$ に確率収束することから $\frac{(n_0-1)(n_j-1)}{N_j+1} \rightarrow \infty$ になり,
 任意の定数 C に対して, $\pi_0 = \pi_i$ の下では,

$$\text{Prob}(R_j > C) \rightarrow 1 \quad (j \rightarrow \infty)$$

となる。

一方, R_i は, $F_i(x, y)$ の下では, その極限分布として, 自由度2の χ^2 -分布を持つことが [9] で証明されている。

よって, $\pi_0 = \pi_i$ の下では,

$$\text{Prob}(R_i = \min_{1 \leq j \leq k} R_j) \rightarrow 1$$

となり, この判別方式は, consistent である。

§4. 標本の大きさについて.

先に述べた判別式は, 1 変数の場合には,

$$R_i = \frac{12}{n_0 n_i (N_i + 1)} \left(S_i - \frac{n_0 (N_i + 1)}{2} \right)^2 \quad i=1, \dots, k$$

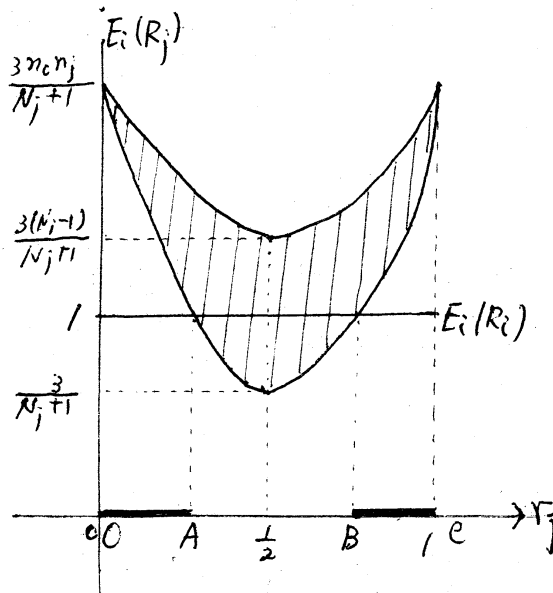
となる。\$S_i\$ は, \$(x_{01}, \dots, x_{0n_0}, x_{i1}, \dots, x_{in_i})\$ の順位 \$(r_1, \dots, r_{n_0+n_i})\$ の最初の \$n_0\$ 個の和である。

\$i \neq j\$ のとき, \$R_j\$ は 分布関数 \$F_i(x)\$ の下で平均をとると,

$$\frac{3}{N_j+1} [4(n_0 n_j - 1)(r_j^2 - r_j) + n_0 n_j] \leq E_i(R_j) \leq$$

$$\frac{3}{N_j+1} [4(n_0 - 1)(n_j - 1)(r_j^2 - r_j) + n_0 n_j]$$

となる。ここに \$R_j = \int_{E_i} F_j(x) dF_i(x)\$。よって, \$R_j\$ の分布関数 \$F_i(x)\$ の下での平均値は, 下図の斜線部分にある。



A, B の座標は,

$$\frac{1}{2} - \sqrt{\frac{N_j - 2}{12(n_0 n_j - 1)}}, \quad \frac{1}{2} + \sqrt{\frac{N_j - 2}{12(n_0 n_j - 1)}}$$

である。

\$E_i(R_i) = 1\$ であるから,

平均値にみれば, 図の OA

BC の部分に \$r_j\$ が落ち

るように, 標本の大きさ

\$(n_0, n_j)\$ を決めればよい。

思われる。これをすべての $j=1, \dots, k$ に行なって、標本の大きさ $(n_0, n_1, \dots, n_k)$ を決めるのが適当であろう。non-parametric な場合、 r_j は未知であるが、標本の大きさが大きくなると、 OA, BC の長さが大きくなるので、十分な標本の大きさをとれば、未知の r_j が、この区間に落ちる確率も大きくなる。これは、又、§3 で証明された consistency から当然である。

参考文献

- [1]. E. Fix and J. L. Hodges (1951)
Discriminatory Analysis, Non-Parametric Discrimination: Consistency Properties. USAF. school of aviation Medicine, Report No.4
- [2]. E. Fix and J. L. Hodges (1952)
Discriminatory Analysis, Non-Parametric Discrimination: Small Sample Performance. USAF. school of aviation Medicine, Report No.11
- [3]. S. Das-Gupta (1964)
Non-Parametric Classification Rules. Sankhyā A 24. p.p. 25-30
- [4]. H. Hudimoto (1963)
分類について - I. 二群への分類 AISM. p.p. 31-38
- [5]. H. Hudimoto (1964)
On a Distribution-free Two-Way Classification. AISM. 16 p.p. 247-253
- [6]. H. Hudimoto (1968)
On the Empirical Bayes Procedure (1). AISM. 20 p.p. 169-185
- [7]. C. R. Peltó (1969)
Adaptive Non-Parametric Classification. Technometrics 11.
- [8]. C. P. Queenberry and M. P. Gassaman (1968) Non-Parametric
Discrimination Using Tolerance regions. AMS. 39 p.p. 664-673
- [9]. J. Van Ryzin (1966)
Bayes Risk Consistency of Classification Procedures Using Density

Estimation. Sankhyā A 26. p.p. 261-270

[10]. S. K. Chatterjee and P. K. Sen (1964)

Non-Parametric Tests for the Bivariate Two-Sample Location Problems.

CSAB 14 p.p. 18-58

[11]. D. S. Stoller (1954)

Univariate Two-Population Distribution-Free Discrimination.

TASA 49 p.p. 770-777

[12]. M. N. Woinsky and L. Kurz (1969)

Sequential Non-Parametric Two-Way Classification with a Prescribed

Maximum Asymptotic Error Probability. AMS. 40 445-455.