クラスター分析に関するノート

情報学部 堀田 敬介

2004/7/31（2008/7/1 改訂, 2009/10/31 改訂）

1　類似度の測定

まずはじめに，各データ間の距離を測るが，尺度毎に様々な方法が提案されている．尺度に対応した類似度測定の距離を示す.

1.1　間隔尺度による類似度の測定

n 個の対象があり，各対象は間隔尺度で m 個の属性（変量）が測定されているとする．このとき対象 p と q を

$$\boldsymbol{x}_p = [x_{p1}, \cdots, x_{pm}]^T, \quad \boldsymbol{x}_q = [x_{q1}, \cdots, x_{qm}]^T$$

とし，その距離 d_{pq} を測ることを考える．

ユークリッド距離
$$d_{pq} = \sqrt{\sum_{i=1}^{m} (x_{pi} - x_{qi})^2}$$

ユークリッド平方距離
$$d_{pq} = \sum_{i=1}^{m} (x_{pi} - x_{qi})^2$$

重み付きユークリッド距離
$$d_{pq} = \sqrt{\sum_{i=1}^{m} w_i (x_{pi} - x_{qi})^2}$$
w_i $(i = 1, \ldots, m)$ は重み (weight)

マンハッタン距離
$$d_{pq} = \sum_{i=1}^{m} |x_{pi} - x_{qi}|$$

ミンコフスキー距離
$$d_{pq} = \sqrt[k]{\sum_{i=1}^{m} |x_{pi} - x_{qi}|^k}$$
（$k = 1$ ならばマンハッタン距離，$k = 2$ ならばユークリッド距離となる）

キャンベラ距離
$$d_{pq} = \sum_{i=1}^{m} \frac{|x_{pi} - x_{qi}|}{|x_{pi}| + |x_{qi}|}$$

マハラノビス距離
$$d_{pq} = (\boldsymbol{x}_p - \boldsymbol{x}_q)^T \Sigma^{-1} (\boldsymbol{x}_p - \boldsymbol{x}_q)$$
(Σ ： 分散共分散行列)

内積
$$d_{pq} = \boldsymbol{x}_p^T \boldsymbol{x}_q = \sum_{i=1}^{m} x_{pi} x_{qi}$$
（平均 0，長さ $||\boldsymbol{x}_p|| = ||\boldsymbol{x}_q|| = 1$ に標準化した際の Pearson の積率相関係数に一致）

Pearson の積率相関係数
$$r_{pq} = \frac{\sum_{i=1}^{m} (x_{pi} - \bar{x}_p)(x_{qi} - \bar{x}_q)}{\sqrt{\sum_{i=1}^{m} (x_{pi} - \bar{x}_p)^2} \sqrt{\sum_{i=1}^{m} (x_{qi} - \bar{x}_q)^2}}$$

1.2　順序尺度による類似度の測定

n 個の対象があり，各対象は順序尺度で m 個の属性（変量）が測定されているとする．このとき対象 p と q を

$$\boldsymbol{x}_p \ = \ [x_{p1}, \cdots, x_{pm}]^T, \quad \boldsymbol{x}_q \ = \ [x_{q1}, \cdots, x_{qm}]^T$$

とし，各変量は各固体の順位をあらわす数値 $(1, 2, \ldots, m)$ になっている場合，どの程度対応する変量の順位が一致しているかを，順位相関係数 r_{pq} で計ることを考える．

Spearman の順位相関係数　$r_{pq} \ = \ 1 - \dfrac{6}{m(m+1)(m-1)} \displaystyle\sum_{i=1}^{m} (x_{pi} - x_{qi})^2$

（上記の順位尺度変量に Pearson の積率相関係数を計算した結果）

Kendall の順位相関係数　$r_{pq} \ = \ 1 - \dfrac{F - R}{n(n-1)/2}$

Kendall の相関係数について，F, R はそれぞれ，

F は $\forall i, j \in \{1, \ldots, m\} : (i < j)$　$x_{pi} > x_{pj}$ かつ $x_{qi} > x_{qj}$ ，または $x_{pi} < x_{pj}$ かつ $x_{qi} < x_{qj}$ の個数
R は $\forall i, j \in \{1, \ldots, m\} : (i < j)$　$x_{pi} > x_{pj}$ かつ $x_{qi} < x_{qj}$ ，または $x_{pi} < x_{pj}$ かつ $x_{qi} > x_{qj}$ の個数

となる．即ち，全変量対（$n(n-1)/2$ 個）に対する，順序が同じ個数と逆になる個数の差の割合を意味する．

1.3　名義尺度（[0,1]-データ）による類似度の測定

n 個の対象があり，各対象は名義尺度で m 個の属性（変量）が測定されており，その値は 1 または 0 であるとする．このとき対象 p, q

$$\boldsymbol{x}_p \ = \ [x_{p1}, \cdots, x_{pm}]^T, \quad \boldsymbol{x}_q \ = \ [x_{q1}, \cdots, x_{qm}]^T$$

間の類似度 S_{pq} を測定することを考える．

類似比 (the coefficient of Jaccard)	$S_{pq} \ = \ a/(a+b+c)$
一致係数 (the simple matching coefficient)	$S_{pq} \ = \ (a+d)/m$
Russel-Rao 係数	$S_{pq} \ = \ a/m$
Rogers-Tanimoto 係数	$S_{pq} \ = \ (a+d)/(m+b+c)$
Hamann 係数	$S_{pq} \ = \ \{(a+d)-(b+c)\}/m$
ファイ係数	$S_{pq} \ = \ (ad-bc)/\{(a+b)(c+d)(a+c)(b+d)\}^{\frac{1}{2}}$

ただし，

$a \ = \ \sum_{k=1}^{m} x_{pi}x_{qi}$	対象 p, q がともに 1 をとる変量の個数
$b \ = \ \sum_{k=1}^{m} x_{pi}(1-x_{qi})$	対象 p が 1，q が 0 をとる変量の個数
$c \ = \ \sum_{k=1}^{m} (1-x_{pi})x_{qi}$	対象 p が 0，q が 1 をとる変量の個数
$d \ = \ \sum_{k=1}^{m} (1-x_{pi})(1-x_{qi})$	対象 p, q がともに 0 をとる変量の個数

であり，

$$a+b+c+d \ = \ m$$

となる．

1.4 名義尺度による類似度の測定（変量間類似度）

n 個の対象があり，各対象は名義尺度で m 個の属性（変量）の度数が測定されているとする．対象 p の属性 i の度数を n_{pi} で表す.

対象 \ 属性	1	...	i	$\cdots$	m
1	f_{11}	$\cdots$	f_{1i}	$\cdots$	f_{1m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$
p	f_{p1}	$\cdots$	f_{pi}	$\cdots$	f_{pm}
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$
n	f_{n1}	$\cdots$	f_{ni}	$\cdots$	f_{nm}

平均平方根一致係数 $$C = \sqrt{\chi^2/(\chi^2+nm)}$$

（χ^2 は 2 つの変量間の独立性検定のためのカイ 2 乗統計量）

グッドマン・クラスカルの λ $$\lambda = \frac{\sum_p \max_i f_{pi} + \sum_i \max_p f_{pi} - \max_p \sum_i f_{pi} - \max_i \sum_p f_{pi}}{2nm - \max_p \sum_i f_{pi} - \max_i \sum_p f_{pi}}$$

2 クラスター分析：クラスター化の方法

クラスター p とクラスター q があわさり一つのクラスター t を作る場合，新しくできるクラスター t と p, q 以外のクラスタ達（r と呼ぼう）との類似度 S_{tr} を求める必要がある．今，p と q，p と r，q と r の元の類似度をそれぞれ S_{pq}，S_{pr}, S_{qr} としたときに，これらを基にして S_{tr} を求める方法は，例えば以下のようなものがある．

1. 最短距離法 (nearest neighbor method)
2. 最長距離法 (furthest neighbor method)
3. 群平均法 (group average method)
4. 重心法 (centroid method)
5. 中央値法 (median method)
6. ウォード法 (Ward method)

ここでは，上記 6 つのうち，重心法とウォード法についてのみ記す.

2.1 重心法 centroid method

重心法は，クラスタ間の類似度を各クラスタの重心間の距離で測る方法．クラスタ p, q 間，p, r 間，q, r 間，及び t, r 間の類似度 S_{pq}, S_{pr}, S_{qr} 及び S_{tr} について，

$$S_{tr} = \frac{n_p}{n_p+n_q}S_{pr} + \frac{n_q}{n_p+n_q}S_{qr} - \frac{n_p n_q}{(n_p+n_q)^2}S_{pq}$$

と更新する方法である．クラスタ $p, q, r,$ 及び t の重心をそれぞれ $\boldsymbol{x}_p, \boldsymbol{x}_q, \boldsymbol{x}_r,$ 及び $\boldsymbol{x}_t$ とし，クラスタ $p, q, r,$ 及び t 内の対象数をそれぞれ $n_p, n_q, n_r,$ 及び n_t とする．すると，

$$\begin{aligned} n_t &= n_p + n_q, \\ \bar{\boldsymbol{x}}_t &= \frac{n_p}{n_p+n_q}\bar{\boldsymbol{x}}_p + \frac{n_q}{n_p+n_q}\bar{\boldsymbol{x}}_q \end{aligned}$$

である．クラスタ t, r 間のユークリッド平方距離を d_{tr}^2 とすると，

$$\begin{aligned} d_{tr}^2 &= ||\bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}}_r||^2 = \left\| \left(\frac{n_p}{n_p+n_q}\bar{\boldsymbol{x}}_p + \frac{n_q}{n_p+n_q}\bar{\boldsymbol{x}}_q \right) - \bar{\boldsymbol{x}}_r \right\|^2 \\ &= \left\| \frac{n_p}{n_p+n_q}(\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r) + \frac{n_q}{n_p+n_q}(\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r) \right\|^2 \\ &= \frac{n_p^2}{(n_p+n_q)^2}\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r\|^2 + \frac{n_q^2}{(n_p+n_q)^2}\|\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r\|^2 - \frac{2n_pn_q}{(n_p+n_q)^2}|(\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r, \bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r)| \\ &= \frac{n_p(n_p+n_q) - n_pn_q}{(n_p+n_q)^2}\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r\|^2 + \frac{n_q(n_p+n_q) - n_pn_q}{(n_p+n_q)^2}\|\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r\|^2 - \frac{2n_pn_q}{(n_p+n_q)^2}|(\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r, \bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r)| \\ &= \frac{n_p}{n_p+n_q}\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r\|^2 + \frac{n_q}{n_p+n_q}\|\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r\|^2 - \frac{n_pn_q}{(n_p+n_q)^2}\left\{\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r\|^2 + \|\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r\|^2 - 2\,|(\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r, \bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r)|\right\} \\ &= \frac{n_p}{n_p+n_q}\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r\|^2 + \frac{n_q}{n_p+n_q}\|\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r\|^2 - \frac{n_pn_q}{(n_p+n_q)^2}\|(\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r) - (\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r)\|^2 \\ &= \frac{n_p}{n_p+n_q}\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_r\|^2 + \frac{n_q}{n_p+n_q}\|\bar{\boldsymbol{x}}_q - \bar{\boldsymbol{x}}_r\|^2 - \frac{n_pn_q}{(n_p+n_q)^2}\|\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_q\|^2 \\ &= \frac{n_p}{n_p+n_q}d_{pr}^2 + \frac{n_q}{n_p+n_q}d_{qr}^2 - \frac{n_pn_q}{(n_p+n_q)^2}d_{pq}^2 \end{aligned}$$

となる．従って，重心法では，類似度 S_{tr} としてユークリッド平方距離 d_{tr}^2 をとったときのみ妥当．

2.2 ウォード法 Ward method

クラスタ p とクラスタ q を一つのクラスタ t に統合するとき，他のクラスタ r との類似度 S_{tr} を決める方法の一つ．クラスタ p に属する対象 j を $\boldsymbol{x}_{pj}$ と表すことにする．

$$\boldsymbol{x}_{pj} = \begin{pmatrix} x_{1pj} \\ \vdots \\ x_{mpj} \end{pmatrix} \in \mathbb{R}^m$$

クラスタ p 内の変動 D_p は

$$D_p := \sum_{i=1}^{m}\sum_{j=1}^{n_p}(x_{ipj} - \bar{x}_{ip\cdot})^2 \quad \text{ただし} \quad \bar{x}_{ip\cdot} = \frac{1}{n_p}\sum_{j=1}^{n_p}x_{ipj}$$

であり，全クラスタ内変動 D は，クラスタ数を K として，

$$D := \sum_{p=1}^{K} D_p$$

で与えられる．またこのとき，

$$D_t = D_p + D_q + \Delta D_{pq} \quad \text{ただし} \quad \Delta D_{pq} = \frac{n_pn_q}{n_p+n_q}\sum_{i=1}^{m}(\bar{x}_{ip\cdot} - \bar{x}_{iq\cdot})^2$$

となる．なぜなら，

$$
\begin{aligned}
n_t &= n_p + n_q, \\
\bar{x}_{it\cdot} &= \frac{1}{n_t}\sum_{j=1}^{n_t} x_{itj} \;=\; \frac{1}{n_t}\left(\sum_{j=1}^{n_p} x_{ipj} + \sum_{j=1}^{n_q} x_{iqj}\right) \\
&= \frac{n_p}{n_t}\sum_{j=1}^{n_p} x_{ipj} + \frac{n_q}{n_t}\sum_{j=1}^{n_q} x_{iqj} \;=\; \frac{n_p}{n_t}\bar{x}_{ip\cdot} + \frac{n_q}{n_t}\bar{x}_{iq\cdot}
\end{aligned}
$$

であることに注意すると，

$$
\begin{aligned}
D_t &= \sum_{i=1}^{m}\sum_{j=1}^{n_t}(x_{itj} - \bar{x}_{it\cdot})^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_t}\{x_{itj}^2 - 2x_{itj}\bar{x}_{it\cdot} + \bar{x}_{it\cdot}^2\} \\
&= \sum_{i=1}^{m}\left\{\sum_{j=1}^{n_p}\{x_{ipj}^2 - 2x_{ipj}\bar{x}_{it\cdot} + \bar{x}_{it\cdot}^2\} + \sum_{j=1}^{n_q}\{x_{iqj}^2 - 2x_{iqj}\bar{x}_{it\cdot} + \bar{x}_{it\cdot}^2\}\right\} \\
&= \sum_{i=1}^{m}\left\{\sum_{j=1}^{n_p}(x_{ipj} - \bar{x}_{ip\cdot})^2 + 2\bar{x}_{ip\cdot}\sum_{j=1}^{n_p} x_{ipj} - n_p\bar{x}_{ip\cdot}^2 - 2\bar{x}_{it\cdot}\sum_{j=1}^{n_p} x_{ipj} + n_p\bar{x}_{it\cdot}^2\right. \\
&\qquad \left. + \sum_{j=1}^{n_q}(x_{iqj} - \bar{x}_{iq\cdot})^2 + 2\bar{x}_{iq\cdot}\sum_{j=1}^{n_q} x_{iqj} - n_q\bar{x}_{iq\cdot}^2 - 2\bar{x}_{it\cdot}\sum_{j=1}^{n_q} x_{iqj} + n_q\bar{x}_{it\cdot}^2\}\right\} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_p}(x_{ipj} - \bar{x}_{ip\cdot})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_q}(x_{iqj} - \bar{x}_{iq\cdot})^2 \\
&\qquad + \sum_{i=1}^{m}\left\{2n_p\bar{x}_{ip\cdot}^2 - n_p\bar{x}_{ip\cdot}^2 - 2n_p\bar{x}_{ip\cdot}\bar{x}_{it\cdot} + n_p\bar{x}_{it\cdot}^2 + 2n_q\bar{x}_{iq\cdot}^2 - n_q\bar{x}_{iq\cdot}^2 - 2n_q\bar{x}_{iq\cdot}\bar{x}_{it\cdot} + n_q\bar{x}_{it\cdot}^2\right\} \\
&= D_p + D_q + \sum_{i=1}^{m}\left\{n_p\bar{x}_{ip\cdot}^2 + n_q\bar{x}_{iq\cdot}^2 - n_t\bar{x}_{it\cdot}^2\right\} \\
&= D_p + D_q + \sum_{i=1}^{m}\left\{n_p\bar{x}_{ip\cdot}^2 + n_q\bar{x}_{iq\cdot}^2 - \frac{1}{n_t}\left(n_p^2\bar{x}_{ip\cdot}^2 + 2n_pn_q\bar{x}_{ip\cdot}\bar{x}_{iq\cdot} + n_q^2\bar{x}_{iq\cdot}^2\right)\right\} \\
&= D_p + D_q + \sum_{i=1}^{m}\left\{\frac{(n_t - n_p)n_p}{n_t}\bar{x}_{ip\cdot}^2 - \frac{n_pn_q}{n_t}\bar{x}_{ip\cdot}\bar{x}_{iq\cdot} + \frac{(n_t - n_q)n_q}{n_t}\bar{x}_{iq\cdot}^2\right\} \\
&= D_p + D_q + \frac{n_pn_q}{n_p + n_q}\sum_{i=1}^{m}(\bar{x}_{ip\cdot} - \bar{x}_{iq\cdot})^2 \\
&= D_p + D_q + \Delta D_{pq}
\end{aligned}
$$

だからである．

さて，ウォード法は，全クラスタ内変動 D を最も小さくするのが好ましいとする方法，即ち，各反復のクラスタリング（例えば p, q を統合して t にする）について，その増分（例では ΔD_{pq}）を最小にすることを考える方法である．

この増分（ΔD_{pq}）を類似度 S_{pq} とし，類似度が最も小さいクラスタ同士を統合する．なお，初期状態（すべてのクラスタが 1 つの対象から構成されている状態）では，この値はユークリッド平方距離の $\frac{1}{2}$ となる．

また，類似度の更新は，次式で与えられる．

$$
\begin{aligned}
(\Delta D_{tr} \ =) \quad S_{tr} &= \frac{n_t n_r}{n_t + n_r} \sum_{i=1}^{m} (\bar{x}_{it\cdot} - \bar{x}_{ir\cdot})^2 \\
&= \frac{n_t n_r}{n_t + n_r} \sum_{i=1}^{m} \left(\frac{n_p}{n_t} \bar{x}_{ip\cdot} + \frac{n_q}{n_t} \bar{x}_{iq\cdot} - \bar{x}_{ir\cdot} \right)^2 \\
&= \frac{n_t n_r}{n_t + n_r} \sum_{i=1}^{m} \left(\frac{n_p^2}{n_t^2} \bar{x}_{ip\cdot}^2 + \frac{n_q^2}{n_t^2} \bar{x}_{iq\cdot}^2 + \bar{x}_{ir\cdot}^2 + \frac{2 n_p n_q}{n_t^2} \bar{x}_{ip\cdot} \bar{x}_{iq\cdot} - \frac{2 n_p}{n_t} \bar{x}_{ip\cdot} \bar{x}_{ir\cdot} - \frac{2 n_q}{n_t} \bar{x}_{iq\cdot} \bar{x}_{ir\cdot} \right) \\
&= \frac{n_r}{n_t + n_r} \sum_{i=1}^{m} \left\{ n_p (\bar{x}_{ip\cdot}^2 - 2 \bar{x}_{ip\cdot} \bar{x}_{ir\cdot} + \bar{x}_{ir\cdot}^2) + n_q (\bar{x}_{iq\cdot}^2 - 2 \bar{x}_{iq\cdot} \bar{x}_{ir\cdot} + \bar{x}_{ir\cdot}^2) - \frac{n_p n_q}{n_t} (\bar{x}_{ip\cdot}^2 + 2 \bar{x}_{ip\cdot} \bar{x}_{iq\cdot} + \bar{x}_{iq\cdot}^2) \right. \\
&= \frac{1}{n_t + n_r} \left\{ n_p n_r \sum_{i=1}^{m} (\bar{x}_{ip\cdot} - \bar{x}_{ir\cdot})^2 + n_q n_r \sum_{i=1}^{m} (\bar{x}_{iq\cdot} - \bar{x}_{ir\cdot})^2 - n_r \frac{n_p n_q}{n_p + n_q} \sum_{i=1}^{m} (\bar{x}_{ip\cdot} - \bar{x}_{iq\cdot})^2 \right\} \\
&= \frac{1}{n_t + n_r} \{ (n_p + n_r) \Delta D_{pr} + (n_q + n_r) \Delta D_{qr} - n_r \Delta D_{pq} \} \\
&= \frac{n_p + n_r}{n_t + n_r} S_{pr} + \frac{n_q + n_r}{n_t + n_r} S_{qr} - \frac{n_r}{n_t + n_r} S_{pq}
\end{aligned}
$$

2.3 6つの方法を統合する式

G.N. Lance & W.T.Williams による統一的に扱うための式．各方法の違いを，パラメータ（$\alpha_p, \alpha_q, \beta, \gamma$）の違いで決定できると示した．平方距離を使用する場合は2番目の式となる．

$$\begin{cases} S_{tr} = \alpha_p S_{pr} + \alpha_q S_{qr} + \beta S_{pq} + \gamma |S_{pr} - S_{qr}| \\ S_{tr}^2 = \alpha_p S_{pr}^2 + \alpha_q S_{qr}^2 + \beta S_{pq}^2 + \gamma |S_{pr}^2 - S_{qr}^2| \quad \text{（平方距離用）} \end{cases}$$

(1) 最短距離法 (nearest neighbor method)

$$\alpha_p = \alpha_q := \frac{1}{2},\ \beta := 0,\ \gamma := -\frac{1}{2}$$

$$\begin{aligned} S_{tr} &= \frac{1}{2}S_{pr} + \frac{1}{2}S_{qr} - \frac{1}{2}|S_{pr} - S_{qr}| \\ &= \begin{cases} S_{pr} & \text{for} \quad S_{pr} \le S_{qr}, \\ S_{qr} & \text{for} \quad S_{pr} > S_{qr} \end{cases} \\ &= \min\{S_{pr}, S_{qr}\} \end{aligned}$$

(2) 最長距離法 (furthest neighbor method)

$$\alpha_p = \alpha_q := \frac{1}{2},\ \beta := 0,\ \gamma := \frac{1}{2}$$

$$\begin{aligned} S_{tr} &= \frac{1}{2}S_{pr} + \frac{1}{2}S_{qr} + \frac{1}{2}|S_{pr} - S_{qr}| \\ &= \begin{cases} S_{pr} & \text{for} \quad S_{pr} \ge S_{qr}, \\ S_{qr} & \text{for} \quad S_{pr} < S_{qr} \end{cases} \\ &= \max\{S_{pr}, S_{qr}\} \end{aligned}$$

(3) 重心法 (centroid method)

$$\alpha_p := \frac{n_p}{n_t},\ \alpha_q := \frac{n_q}{n_t},\ \beta\frac{n_p n_q}{n_t^2},\ \gamma := 0$$

$$S_{tr}^2 = \frac{n_p}{n_t}S_{pr}^2 + \frac{n_q}{n_t}S_{qr}^2 - \frac{n_p n_q}{n_t^2}S_{pq}^2$$

(4) 中央値法 (median method)

$$\alpha_p = \alpha_q := \frac{1}{2},\ \beta := -\frac{1}{4},\ \gamma := 0$$

$$S_{tr} = \frac{1}{2}S_{pr} + \frac{1}{2}S_{qr} - \frac{1}{4}S_{pq}$$

(5) 群平均法 (group average method)

$$\alpha_p := \frac{n_p}{n_t},\ \alpha_q := \frac{n_q}{n_t},\ \beta = \gamma := 0$$

$$S_{tr}^2 = \frac{n_p}{n_t}S_{pr}^2 + \frac{n_q}{n_t}S_{qr}^2$$

(6) ウォード法 (Ward method)

$$\alpha_p := \frac{n_p + n_r}{n_t + n_r},\ \alpha_q := \frac{n_q + n_r}{n_t + n_r},\ \beta := -\frac{n_r}{n_t + n_r},\ \gamma := 0$$

$$S_{tr}^2 = \frac{n_p + n_r}{n_t + n_r}S_{pr}^2 + \frac{n_q + n_r}{n_t + n_r}S_{qr}^2 - \frac{n_r}{n_t + n_r}S_{pq}^2$$