テューキー・クレーマー法の中間点の評価に関する研究

M2012MM006 日比野真之
指導教員：松田眞一

1　はじめに

私は次節で述べる不等式に興味を持ち，この研究を始めることにした．具体的には，不等式の中間点の評価に関して研究をする．前半で研究してきた理論，後半でさまざまなデータで行ったシミュレーションについて述べる．

2　本論文での不等式の概要

私が興味を持った不等式は，白石 [4] に書かれていたテューキー・クレーマーの多重比較検定の式

$$TA(t) \leq P_0\left(\max_{i<i'} |T_{ii'}| \leq t\right) \leq TA^*(t) \qquad (1)$$

である．ここで検定統計量 $T_{ii'}$ は標本 X_{ij} $(1 \leq i \leq k; 1 \leq j \leq n_i)$ に対して

$$T_{ii'} \equiv \frac{\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}}{\sqrt{V_E\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}} \qquad (1 \leq i < i' \leq k)$$

で与えられ

$$TA(t) \equiv k\int_0^\infty \left[\int_{-\infty}^\infty \{\Phi(x) - \Phi(x - \sqrt{2}ts)\}^{k-1} d\Phi(x)\right] g(s)ds$$

$$TA^*(t) \equiv \sum_{j=1}^k \int_0^\infty \left[\int_{-\infty}^\infty \prod_{\substack{i=1\\ i\neq j}}^k \left\{\Phi\left(\sqrt{\frac{\lambda_{ni}}{\lambda_{nj}}}x\right) - \Phi\left(\sqrt{\frac{\lambda_{ni}}{\lambda_{nj}}}x - \sqrt{\frac{\lambda_{ni}+\lambda_{nj}}{\lambda_{nj}}}ts\right)\right\} d\Phi(x)\right] g(s)ds$$

であり，n は総データ数，k は群の数である．

$$\lambda_{ni} \equiv \frac{n_i}{n}(i = 1, \cdots, k)$$

$$\begin{aligned} g(s) &\equiv \frac{m^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)2^{\frac{m}{2}-1}} s^{m-1}\exp\left(\frac{-ms^2}{2}\right) \\ &= \frac{mse^{-s^2}c^{\left(\frac{m}{2}-1\right)}}{\Gamma\left(\frac{m}{2}\right)} \\ c &\equiv \frac{ms^2e^{-s^2}}{2}, \quad m \equiv n-k \end{aligned}$$

である．
このうち，左側の不等式は，Hayter[1] に，右側の不等式は，白石 [4] に詳しい証明が載せられている．それぞれの理論について，次節以降で述べる．

3　Hayter(1984) の定理

Hayter[1] では，以下の定理について証明がかなり詳細に述べられている．

定理　$1 \leq i \leq k$ に対して $X_i \sim N(0, \sigma_i^2)$ は独立であるとする．ただし，$0 < \sigma_i < \infty$ であり，いくつかの固定した $q > 0$ に対し，$\xi_{ij} = q(\sigma_i^2 + \sigma_j^2)^{\frac{1}{2}} (1 \leq i, j \leq k)$ としよう．このとき関数

$$F = F(\sigma_1, \cdots, \sigma_k) = P\{|X_i - X_j| \leq \xi_{ij}; 1 \leq i, j \leq k\}$$

は σ_i がすべて等しいときに最小化される．

証明

$$\psi_k(x_1, \cdots, x_k) = \left(\prod_{i=1}^k f_{\sigma_i}(x_i)\right) I_{\{|x_i - x_j| \leq \xi_{ij}; 1 \leq i,j \leq k\}}$$

としよう．ただし，$f_{\sigma_i}(\cdot)$ は $N(0, \sigma_i^2)$ に従う確率変数の密度関数で，I_A は集合 A の指示関数である．ψ の添字 k はその定義域の次元を表す．このとき，

$$\begin{aligned} F =& \int_{-\infty \leq x_1, \cdots, x_k \leq \infty} \cdots \int \psi_k(x_1, \cdots, x_k) dx_1 \cdots dx_k \\ =& \int_{-\infty \leq x_1, \cdots, x_k \leq \infty} \cdots \int \psi_{k-1}(x_1, \cdots, x_{k-1}) f_{\sigma_k}(x_k) \left(\prod_{i=1}^{k-1} I_{\{|x_k - x_i| \leq \xi_{ik}\}}\right) dx_1 \cdots dx_k \end{aligned}$$

であり，$y_k = \frac{x_k}{\sigma_k}$ という置換を行うことにより

$$\begin{aligned} F =& \int_{y_k=-\infty}^\infty \int_{|x_i - y_k\sigma_k| \leq \xi_{ik}, 1 \leq i \leq k-1} \cdots \int (2\pi)^{-\frac{1}{2}} \exp\left\{\frac{-y_k^2}{2}\right\} \psi_{k-1}(x_1, \cdots, x_{k-1}) dx_1 \cdots dx_{k-1} dy_k \end{aligned} \qquad (2)$$

となる．この後，証明は，補題 1〜4(補題 5 および系は補題 4 の証明時に使用) に分けられている．それぞれの補題の内容は以下のとおりである．

補題 1　$1 \leq i \leq k-1$ に対し，

$$x_i^* = y_k\sigma_k + \xi_{ik}, N_i = \{j : 1 \leq j \leq k-1, j \neq i\}$$

を定義し，集合 $V_i(\mathcal{Y}_k) \subseteq \mathbb{R}^{k-2}$ を
$V_i(\mathcal{Y}_k) = \{(x_j : j \in N_i) : |x_j - y_k\sigma_k| \leq \xi_{jk}, \forall j \in N_i\}$
とする．このとき，

$$\frac{\partial F}{\partial \sigma_k} = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \sum_{i=1}^{k-1} G_i$$

である．ただし，

$$G_i = \int_{y_k=-\infty}^\infty \int_{V_i(y_k)} \cdots \int \exp\left\{\frac{-y_k^2}{2}\right\} \psi_{k-1}(x_i, \cdots, x_i^*, \cdots, x_{k-1}) \times \left[y_k + \frac{\partial \xi_{ik}}{\partial \sigma_k}\right] \left(\prod_{j \in N_i} dx_j\right) dy_k$$

である．

補題 2 $1 \leq i \leq k-1$ に対し，補題 1 で定義した数量 G_i は

$$G_i = \int_{r=-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_i^2}+\frac{1}{\sigma_k^2}\right)r^2\right\} \int \cdots \int_{x_i - r \in R_{i,j}, j \in N_i} \psi_{k-2}(x_j : j \in N_i) \left(\prod_{j \in N_i} dx_j\right) dr$$

となる．ただし，$a_i > 0$ はある定数で集合 $R_{ij} \subset \mathbb{R}$ は

$$\left\{x_j : \left| x_j - r + \frac{\sigma_k^2 \xi_{ik}}{\sigma_i^2+\sigma_k^2} \right| \leq \xi_{jk}\right\} \cap \left\{x_j : \left| x_j - r - \frac{\sigma_i^2 \xi_{ik}}{\sigma_i^2+\sigma_k^2} \right| \leq \xi_{ij}\right\} \equiv \{x_j : x_j - r \in R_{ij}\}$$

で定義される．

補題 3 $1 \leq i \leq k-1, j \in N_i$ に対し，補題 2 で定義した集合 R_{ij} は，

$$R_{ij} = \left[\frac{\sigma_i^2 \xi_{ik}}{\sigma_i^2+\sigma_k^2} - \xi_{ij}, \frac{-\sigma_k^2 \xi_{ik}}{\sigma_i^2+\sigma_k^2} + \xi_{jk}\right]$$

であり 0 より大きい長さを持つ．また，もし R_{ij} の中間点を m_{ij} で表すとするならば，このとき $\forall j \in N_i, 1 \leq i \leq k-1$ に対し，$\sigma_i > (=,<)\sigma_k \Leftrightarrow m_{ij} > (=,<)0$ が成り立つ．

補題 4 補題 2 の $G_i (1 \leq i \leq k-1)$ に対し，
$\sigma_i > (=,<)\sigma_k \Leftrightarrow G_i > (=,<)0$ が成り立つ．

補題 5 有限集合 $J \subset \mathbb{N}$ とある固定された $r > 0, \delta_{ij} \geq 0, \tau_i > 0$ に対し，

$$g_{|J|}(x_i : i \in J) = \left(\prod_{i \in J} f_{\tau_i}(x_i + r)\right) I_{\{|x_i - x_j| \leq \delta_{ij}; \forall i,j \in J\}}$$

を定義する．ただし，$f_{\tau_i}(x)$ は $N(0, \tau_i^2)$ に従う確率変数の密度である．また，ある固定された $d_i > 0$ と $m_i \in \mathbb{R}$ について，$n \in \mathbb{N}$ に対し，
$A_n = \{(x_1, \cdots, x_n) : m_i - d_i \leq x_i \leq m_i + d_i, 1 \leq i \leq n\} \subseteq R^n$
と
$-A_n = \{(x_1, \cdots, x_n) : -m_i - d_i \leq x_i \leq -m_i + d_i, 1 \leq i \leq n\} \subseteq R^n$
を定義する．このとき，もし $1 \leq i \leq n$ に対し $m_i \geq 0$ ならばすべての $n \in \mathbb{N}$ に対し，

$$\int \cdots \int_{A_n} g_n(x_1, \cdots, x_n) dx_1 \cdots dx_n \leq \int \cdots \int_{-A_n} g_n(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

が成り立つ．

系 (a) $d_i, m_i > 0 \ \forall i$
かつ
(b) $|(m_i + d_i) - (m_j + d_j)| < \delta_{ij} \ \forall i,j$
であると仮定せよ．このとき，すべての $n \in \mathbb{N}$ に対し

$$\int \cdots \int_{A_n} g_n(x_1, \cdots, x_n) dx_1 \cdots dx_n < \int \cdots \int_{-A_n} g_n(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

が成り立つ．

これらを使い，証明を行う．補題 1 と 4 より，$\frac{\partial F}{\partial \sigma_k}$ は，$b_{ki} > (=,<)0 \Leftrightarrow \sigma_k > (=,<)\sigma_i$ を満たすような b_{ki} に対し，

$$\frac{\partial F}{\partial \sigma_k} = \sum_{i=1}^{k-1} b_{ki}$$

のように表すことができる．しかし，σ_k は σ の中から任意に選んだので，より一般的には $1 \leq i \leq k$ に対し，$\frac{\partial F}{\partial \sigma_i}$ は，$b_{ij} > (=,<)0 \Leftrightarrow \sigma_i > (=,<)\sigma_j$ を満たすような b_{ij} に対し，

$$\frac{\partial F}{\partial \sigma_i} = \sum_{\substack{j=1 \\ j \neq i}}^{k-1} b_{ij}$$

のように表される．
この結果を使うと，すべての σ_i が等しいとき F が最小であるという結果を得ることができる．その手順は以下のとおりである．
$\sigma_{[1]} \leq \sigma_{[2]} \leq \cdots \leq \sigma_{[k]}$ を順序づけられた σ_i を表すものとしよう．もし
(1)$\sigma_{[1]}$ から $\sigma_{[2]}$ に増加させる
(2)$\sigma_{[1]}$ と $\sigma_{[2]}$ から $\sigma_{[3]}$ に増加させる ($\sigma_{[1]} = \sigma_{[2]}$ のまま)
⋮
(k-1)$\sigma_{[1]}, \cdots, \sigma_{[k-1]}$ から $\sigma_{[k]}$ に増加させる．($\sigma_{[1]} = \cdots = \sigma_{[k-1]}$ のまま)
とするならば，F は増加が必要な段階ごとに (すべての σ_i が等しくなければそのような段階は存在する)，狭義単調減少していく．
これで定理の証明が完結した．

4 シミュレーション

(1) 式の中間点が真値とどのような関係にあるか，シミュレーションで評価する．まず，以下の記号を定義する．

$TA(t) = 1-\alpha$を満たす t の解を $ta(k, m; \alpha)$

$P_0\left(\max_{i<i'} |T_{ii'}| \leq t\right) = 1-\alpha$を満たす t の解を $te(k, m, \lambda_{n1}, \cdots, \lambda_{nk}; \alpha)$

$TA^*(t) = 1-\alpha$を満たす t の解を $ta^*(k, m, \lambda_{n1}, \cdots, \lambda_{nk}; \alpha)$

とする．(1) より，

$$ta^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};\alpha) \le te(k,m,\lambda_{n1},\cdots,\lambda_{nk};\alpha) \le ta(k,m;\alpha)$$

の関係が成り立つ．
(1) の両端点 $TA(t)$，$TA^*(t)$ は二重積分で表現されているため，精度保証された数値積分が可能である．しかし，真ん中の式

$$P_0\left(\max_{i<i'}|T_{ii'}| \le t\right)$$

は，$n_1 = \cdots = n_k$ でなければ複雑な多重積分でしか表現できない．このため，精度保証された数値積分が行えず $te(k,m,\lambda_{n1},\cdots,\lambda_{nk};\alpha)$ の代わりに $ta(k,m;\alpha)$ が使われ，保守的な方法になっている．そこで，私は

$$te^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};\alpha) \equiv \frac{1}{2}\left(ta^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};\alpha) + ta(k,m;\alpha)\right)$$

と定義し，これがどの程度保守的なのかを評価する．
まず，求める点は 5%点とするので，$\alpha = 0.05$ である．R でのプログラムを用いて，はじめに $ta(k,m;0.05)$，$ta^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};0.05)$ を導出する．導出した値の妥当性は白石 [3] で手に入るプログラム，白石 [4] の数表と見比べながら評価した．探索方法は二分法を用いた．二分法とは，

$Step1$	最小値と最大値を決める
$Step2$	中間値 $=\frac{1}{2}$(最小値 + 最大値) とする
$Step3$	中間値 < 探索値ならば
	中間値を新たな最小値とする
	中間値 > 探索値ならば
	中間値を新たな最大値とする
$Step4$	十分な精度が保証されるまで
	$Step2$ と $Step3$ を繰り返す

というものである．本研究では，上記の最小値を 2，最大値を 3，探索値は 0.9500，十分な精度は 最大値 − 最小値 ≤ 0.001 として探索を行った．
ここで求まった近似値を用いることで，

$$te^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};0.05) \equiv \frac{1}{2}\left(ta(k,m;0.05) + ta^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};0.05)\right)$$

が求まる．その後，set.seed() 関数の初期値を 0 から 100000 まで 10000 刻みで各 20 万回繰り返しを行った．繰り返しの内容は，まずデータサイズを決め，rnorm() 関数を使った乱数でできた 3 群を作る．後に堀内 [2] のプログラムを使用するので，長さは最も長い群に合わせるよう，不足分には NA を入れておく．各棄却点を以下のように決める．

中間点の棄却点	$\sqrt{2}te^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};0.05)$
$TA(t)$ の棄却点	$\sqrt{2}ta(k,m;0.05)$
$TA^*(t)$ の棄却点	$\sqrt{2}ta^*(k,m,\lambda_{n1},\cdots,\lambda_{nk};0.05)$

ここで，堀内 [2] で得られた検定統計量と上記の棄却点を比較し，棄却された回数を出力するようにする．最終的に，出力結果を見ることで，リベラル (保守的でない) か保守的かを知ることができる．すなわち，20 万回中 5%(10000 回) 以上棄却されていればリベラル，そうでなければ保守的ということである．
以下では，3 つの群サイズを (a,b,c) のように表記することにする．実際に行った実験であるが，大標本の場合として，(30,45,60)，(20,40,80)，(20,50,100)，(40,60,80) の 4 通り，小標本の場合として，(14,21,28)，(10,15,20) の 2 通りを行った．群サイズの最大値が最小値の 2 倍以上の例があるが，これは，群サイズの差が大きくなった時の $TA(t)$，中間点，$TA^*(t)$ の傾向を調べるためである．
結果，棄却された回数の 95% 信頼区間は，表 1，図 1，図 2 のようになった．信頼区間は

$$\left[\bar{X} - 1.96\sqrt{\frac{S^2}{11}},\quad \bar{X} + 1.96\sqrt{\frac{S^2}{11}}\right]$$

で導出した．

表 1　棄却された回数の 95% 信頼区間

(30,45,60)	中間点	[10070.7571, 10109.0611]
	TA(t)	[9971.5485, 10008.6333]
	$TA^*(t)$	[10161.4846, 10198.6972]
(20,40,80)	中間点	[10070.1693, 10123.2852]
	TA(t)	[9768.4642, 9816.8085]
	$TA^*(t)$	[10393.7163, 10447.3746]
(20,50,100)	中間点	[10109.5908, 10166.4092]
	TA(t)	[9738.6487, 9781.5331]
	$TA^*(t)$	[10517.4329, 10572.3853]
(40,60,80)	中間点	[10007.1114, 10055.0704]
	TA(t)	[9921.8635, 9973.5910]
	$TA^*(t)$	[10110.0058, 10158.7214]
(14,21,28)	中間点	[10024.4434, 10061.5566]
	TA(t)	[9910.7826, 9949.0356]
	$TA^*(t)$	[10109.5821, 10150.9633]
(10,15,20)	中間点	[10039.2576, 10092.0151]
	TA(t)	[9961.1856, 10010.0872]
	$TA^*(t)$	[10137.0193, 10187.7080]

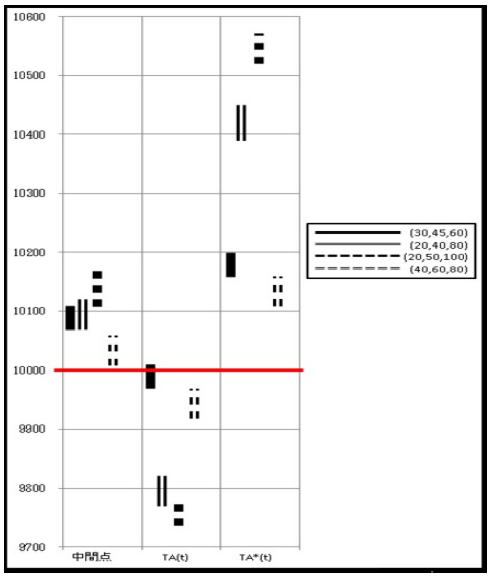

図 1　95% 信頼区間のグラフ (大標本)

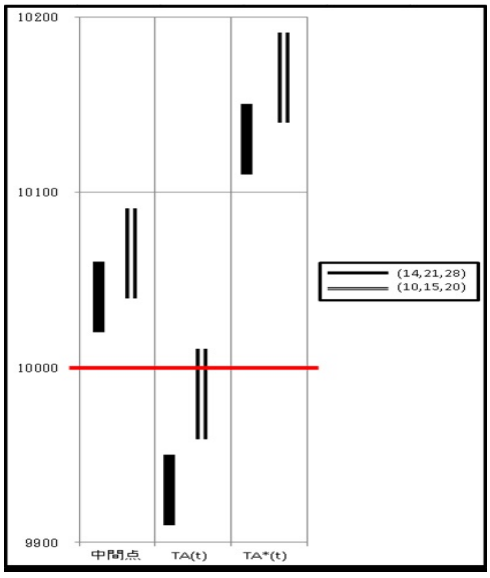

図 2　95% 信頼区間のグラフ (小標本)

5　考察

大標本の場合，(20,40,80) と (20,50,100) との比較では，サンプルサイズが大きくなると $TA(t)$ はより保守的に，$TA^*(t)$ と中間点はよりリベラルになるということが分かった．(30,45,60) と (40,60,80) ではサンプルサイズが大きくなると全体的に有意水準は 0.05 に近づく．(30,45,60) の $TA(t)$ は 5%に近いが，これは偶然の要素があると考える．

小標本の場合，サンプルサイズの差が小さくなると中間点と $TA^*(t)$ がよりリベラルになっていくことが分かった．一方，$TA(t)$ の保守度は下がった．先と同様，(10,15,20) の $TA(t)$ が 5%に近いのも，偶然の要素があると考えるが，リベラルに向かう端緒である可能性もある．$TA^*(t)-TA(t)$ は，大標本，小標本問わず，サンプルサイズの差が大きくなると大きくなった．

6　おわりに

今回の実験で，中間点は保守的，$TA^*(t)$ はリベラル，すなわち棄却された回数が

$$TA(t) \leq \text{中間点} \leq 10000 \leq TA^*(t)$$

となるような例を見つけたかったが，残念ながら見つけることができなかった．私の提案した中間点はリベラルであるということが分かった．

本研究では，特に大標本の場合は 1 回につき 3 時間程度もの実験時間がかかるので，実験に当てた時間が足りていなかったのも，うまくいく例を見つけられなかった原因であると考える．また，中間点の提案も 1 パターンのみではなく，様々なものを提案してみると変わった結果が得られたかもしれない．

一方，小標本の場合は，大標本の場合ほど実験時間を必要としないので，上に書いたことを実行してみるのには最適であると考える．

謝辞

指導教員の松田眞一先生には，数学理論の質問を中心に様々な質問に答えて下さり，感謝致します．また，木村美善先生，白石高章先生には，中間発表でのご指摘や統計の知識を教えて下さったことを感謝致します．

参考文献

[1] Hayter, A. J. : A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. Ann. Statist., **12**. 61-75, 1984.

[2] 堀内賢太郎：『R による多重比較法の研究』， 南山大学数理科学科卒業論文， 2003.

[3] 白石高章： プログラムソフト，
http://www.seto.nanzan-u.ac.jp/˜marble/tbook/softp2.html.

[4] 白石高章：『多群連続モデルにおける多重比較法— パラメトリック，ノンパラメトリックの数理統計』， 共立出版，2011.