多標本比率モデルにおける順序制約がある場合の Jonckheere-Terpstra 型検定法

2012SE195 岡田実咲　2012SE222 千田真緒
指導教員：白石高章

1　はじめに

統計学の基礎となる確率や事象を学んできた．それを基に統計学の分布論や二項分布に関係した 1，2 標本モデルを統計的解析法, 分散安定化変換等を学んだ．統計学が実際にどのように用いられているかが明確になるにつれて Jonckheere-Terpstra 検定に興味を持ち，この検定手法について研究することにした．

そこで本論文では，比率に順序制約がある場合の多標本比率モデルにおける Jonckheere-Terpstra 型検定について考察する．

2　正規母集団での Jonckheere-Terpstra 型検定

$i = 1, \cdots, k$，$j = 1, \cdots, n_i$ に対して $Y_{ij} \sim N(\mu_i, \sigma^2)$ とし，すべての Y_{ij}は互いに独立であると仮定する．ただし，$Y_{ij} \sim N(\mu_i, \sigma^2)$ は Y_{ij}が$N(\mu_i, \sigma^2)$ に従うの意味である．帰無仮説 $\mathcal{H}_0$:$\mu_1 = \cdots = \mu_k$ vs. 対立仮説 $\mathcal{H}_A$: $\mu_1 \leqq \cdots \leqq \mu_k$ (少なくとも 1 つの $\leqq$ は $\lneqq$ である．)

$$T_J \equiv \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} n_i n_{i'} (\bar{Y}_{i'\cdot} - \bar{Y}_{i\cdot}), \tag{1}$$

$$\hat{\sigma}^2 \equiv \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2,$$

$$\bar{Y}_{i\cdot} \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

とおき，

$$S_N \equiv \frac{T_J}{\sqrt{\hat{\sigma}^2 \sum_{i'=2}^{k} n_{i'} (\sum_{i=1}^{i'-1} n_i)(\sum_{i=1}^{i'} n_i)}}$$

とする．自由度 $n-k$ の t 分布の上側 100α%点を $t(n-k;\alpha)$ とすると，水準αの検定は，

$$\begin{cases} S_N > t(n-k;\alpha) \text{ のとき } \mathcal{H}_0\text{を棄却する} \\ S_N < t(n-k;\alpha) \text{ のとき } \mathcal{H}_0\text{を棄却しない} \end{cases}$$

となる．この検定法は文献 [2] により導かれた．

3　k 標本比率モデル

ある要因 A が k 個の水準 $A_1, \cdots, A_k$ を考える．水準は群とも呼ばれる．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 標本とよばれ，成功の確率が p_i 失敗の確率が $1-p_i$ のベルヌーイ試行を X_{ij} とする．すなわち，$X_{ij} \sim B(1, p_i)$ である．さらにすべての X_{ij} は互いに独立であると仮定する．帰無仮説 $H_0 : p_1 = \cdots = p_k$ vs. 対立仮説 $H_1 : p_1 \leqq \cdots \leqq p_k$（少なくとも 1 つの不等式は $\lneqq$ である．）H_0の下で，$p_i \equiv p_0 \quad (i = 1, \cdots, k)$ とする．

表 1　k 標本比率モデル

標本	サイズ	データ	成功の回数
第 1 標本	n_1	$X_{11}, \cdots, X_{1n_1}$	W_1
第 2 標本	n_2	$X_{21}, \cdots, X_{2n_2}$	W_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
第 k 標本	n_k	$X_{k1}, \cdots, X_{kn_k}$	W_k

$$(c1) \qquad \lim_{n\to\infty} \frac{n_i}{n} = \lambda_i, 0 < \lambda_i < 1$$

文献 [1] の 214 ページの式 (7.27) を用いると，条件 (c1) の下で，$n \to \infty$ として，

$$\widehat{Z}_i \equiv \sqrt{n_i} \left\{ \arcsin(\sqrt{\widehat{p}_i}) - \arcsin(\sqrt{p_0}) \right\} \xrightarrow{\mathcal{L}} Z_i \sim N\left(0, \frac{1}{4}\right) \tag{2}$$

となる．ただし，

$$\widehat{p}_i \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},\ W_i = \sum_{j=1}^{n_i} X_{ij} \sim B(n_i,\ p_i)$$

とする．

$$V\left(\bar{Y}_{i'\cdot}\right) = \frac{\sigma^2}{n_{i'}},\ V\left(\bar{Y}_{i\cdot}\right) = \frac{\sigma^2}{n_i}$$

を得る．ここで，$\mathcal{H}_0$ の下で $\mu_i \equiv \mu_0$ とおき，

$$\widetilde{Z}_i \equiv \sqrt{n_i}(\bar{Y}_{i\cdot} - \mu_0) \sim N(0, \sigma^2) \iff \bar{Y}_{i\cdot} - \mu_0 = \frac{\widetilde{Z}_i}{\sqrt{n_i}}$$

とすると，式 (1) を $\widetilde{Z}_i$ を用いて以下のように表現できる．

$$T_J = \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} \left(n_i \sqrt{n_{i'}} \widetilde{Z}_{i'} - n_{i'} \sqrt{n_i} \widetilde{Z}_i \right)$$

$i=1,\cdots,k$ に対して，$\widetilde{Z}_i, \widetilde{Z}_{i'}$の代わりに$\widehat{Z}_i, \widehat{Z}_{i'}$を代入したものを$\widehat{T}_J$とすると，

$$\widehat{T}_J = \sum_{i'=2}^{k}\sum_{i=1}^{i'-1} n_i n_{i'} \left\{ \arcsin(\sqrt{\widehat{p}_{i'}}) - \arcsin(\sqrt{\widehat{p}_i}) \right\}$$

が導かれる．

$$\arcsin(\sqrt{\widehat{p}_i}) - \arcsin(\sqrt{p_0}) = \frac{\widehat{Z}_i}{\sqrt{n_i}}$$

であるので，

$$\widehat{T}_J = \sum_{i'=2}^{k}\sum_{i=1}^{i'-1} n_i n_{i'} \left(\frac{\widehat{Z}_{i'}}{\sqrt{n_{i'}}} - \frac{\widehat{Z}_i}{\sqrt{n_i}} \right)$$

と表現できる．$\widetilde{T}_j$ を

$$\widetilde{T}_J \equiv \sum_{i'=2}^{k}\sum_{i=1}^{i'-1} n_i n_{i'} \left(\frac{Z_{i'}}{\sqrt{n_{i'}}} - \frac{Z_i}{\sqrt{n_i}} \right)$$

と定義する．
$\widetilde{T}_J$の平均は，$E(Z_{i'})=0$，$E(Z_i)=0$ より，

$$E(\widetilde{T}_J) = \sum_{i'=2}^{k}\sum_{i=1}^{i'-1} n_i n_{i'} \left(\frac{E(Z_{i'})}{\sqrt{n_{i'}}} - \frac{E(Z_i)}{\sqrt{n_i}} \right) = 0$$

である．また，

$$T_{i'} = \left(\sqrt{n_{i'}} \sum_{i=1}^{i'-1} n_i \right) Z_{i'} - n_{i'} \sum_{i=1}^{i'-1} \sqrt{n_i} Z_i$$

とする．$\ell < \ell'$に対して T_ℓ，$T_{\ell'}$を求めると，

$$T_\ell = \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_\ell \right) Z_\ell - \left(n_\ell \sum_{i=1}^{\ell-1} \sqrt{n_i} Z_i \right),$$

$$T_{\ell'} = \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) Z_{\ell'} - \left(n_{\ell'} \sum_{i=1}^{\ell'-1} \sqrt{n_i} Z_i \right)$$

が求められる．T_ℓ，$T_{\ell'}$の平均はそれぞれ，

$$E(T_\ell)=0,\ E(T_{\ell'})=0$$

である．さらに，

$$\begin{aligned} \mathrm{Cov}\left(T_\ell, T_{\ell'}\right) &= E\left[\left\{T_\ell - E\left(T_\ell\right)\right\}\left\{T_{\ell'} - E\left(T_{\ell'}\right)\right\}\right] \\ &= 0 \end{aligned}$$

となる．また，T_ℓの分散は，

$$V(T_\ell) = \frac{1}{4} n_\ell \left(\sum_{i=1}^{\ell-1} n_i \right) \left\{ \sum_{i=1}^{\ell-1} n_i + n_\ell \right\}$$

である．さらに，

$$\widetilde{T}_{i'} \equiv \left(\sqrt{n_{i'}} \sum_{i=1}^{i'-1} n_i \right) Z_i - \left(n_{i'} \sum_{i=1}^{i'-1} \sqrt{n_i} Z_i \right)$$

とすると，$\widetilde{T}_J$の分散は，

$$V(\widetilde{T}_J) = \sum_{i'=2}^{k} \frac{1}{4} n_{i'} \left(\sum_{i=1}^{i'-1} n_i \right) \left(\sum_{i=1}^{i'} n_i \right)$$

である．以上より，

$$\begin{aligned} \widetilde{T}_J &\equiv \sum_{i'=2}^{k}\sum_{i=1}^{i'-1} n_i n_{i'} \left(\frac{Z_{i'}}{\sqrt{n_{i'}}} - \frac{Z_i}{\sqrt{n_i}} \right) \\ &\sim \quad N\left(0, \sum_{i'=2}^{k} \frac{1}{4} n_{i'} \left(\sum_{i=1}^{i'-1} n_i \right) \left(\sum_{i=1}^{i'} n_i \right) \right) \end{aligned}$$

が成立する．

$$E\left(\frac{\widetilde{T}_J - E\left(\widetilde{T}_J\right)}{\sqrt{V\left(\widetilde{T}_J\right)}} \right) = 0,\ V_0\left(\frac{\widetilde{T}_J - E\left(\widetilde{T}_J\right)}{\sqrt{V\left(\widetilde{T}_J\right)}} \right) = 1$$

より，

$$\frac{\widetilde{T}_J - E\left(\widetilde{T}_J\right)}{\sqrt{V\left(\widetilde{T}_J\right)}} = \frac{\widetilde{T}_J}{\sqrt{V\left(\widetilde{T}_J\right)}} \quad \sim \quad N(0,1)$$

が導かれる．文献 [1] の系 3.6 を用いて，H_0の下で，

$$\frac{\widehat{T}_J}{\sqrt{V\left(\widetilde{T}_J\right)}} \xrightarrow{\mathcal{L}} \frac{\widetilde{T}_J}{\sqrt{V\left(\widetilde{T}_J\right)}} \quad \sim \quad N(0,1)$$

となる．$\widetilde{T}_J$ の分散を代入すると，H_0の下で，

$$S_B \equiv \frac{\widehat{T}_J}{\sqrt{\frac{1}{4}\sum_{i'=2}^{k} n_{i'} \left(\sum_{i=1}^{i'-1} n_i \right) \left(\sum_{i=1}^{i'} n_i \right)}}$$

$$\xrightarrow{\mathcal{L}} \frac{\widetilde{T}_J}{\sqrt{V\left(\widetilde{T}_J\right)}} \quad \sim \quad N(0,1)$$

となる．

検定方式

$$\boldsymbol{X}_i \equiv (X_{i1}, \cdots, X_{in_i}) \qquad (i=1,\cdots,k)$$

とおく．帰無仮説 $H_0 : p_1 = \cdots = p_k$ vs. 対立仮説 $H_1 : p_1 \leqq \cdots \leqq p_k$ （少なくとも 1 つの $\leqq$ は $\lneqq$ である．）
標準正規分布の上側 $100\alpha\%$ 点を $z(\alpha)$ とすると，検定統計量 S_B より，条件 (c1) の下で，

$$\lim_{n\to\infty} P_0(S_B > z(\alpha)) = \alpha$$

であるので検定関数 $\phi(\cdot)$ を

$$\phi(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_k)=\begin{cases}1 & (S_B>z(\alpha)\ \text{のとき})\\ 0 & (S_B<z(\alpha)\ \text{のとき})\end{cases}$$

で定義すれば,

$$\lim_{n\to\infty}E_0\{\phi(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_k)\}=1\times\lim P_0(S_B>z(\alpha))+0\times\lim P_0(S_B<z(\alpha))=\alpha$$

が得られる．よって，$\phi(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_k)$ は，水準 α の漸近的な検定であるといえる．

4　プログラム内容

文献 [4] より，比率モデルにおける順序制約がある場合の Jonckheere-Terpstra 型検定による検定結果を出力するプログラムを C 言語により作成した．以下のプログラムは今回作成したプログラムの main プログラムである．

```
int main(void)
{
    input();
    keisan();
    keisan2();
    keisan3();
    keisan4();
    float ALPHA,XN;
    printf("0 より大きく 0.5 より小さなアルファの値を入力しれください\n");
    scanf("%f",&ALPHA);
    XN=KAI(ALPHA);
    printf("誤差　%f の標準正規分布の上側 %f パーセント点は %f　\n",ERR,100*ALPHA,XN);

    printf("帰無仮説 H0:p1= … =pk vs. t 対立仮説 H1:p1<= … <=pk\n
    (少なくとも 1 つの<=は<≠である )\n");
        if(SB>XN)
        {
            printf("H0 を棄却する\n");
        }
        else{printf("H0 を棄却しない\n");
        }

    return(0);
}
```

4.1　プログラムの流れ

1．input 関数により，標本数，標本サイズ，観測値の個数を入力する．
2．keisan 関数により，$\arcsin\sqrt{(\widehat{p}_i)}$ の値を定義する．
3．keisan2 関数により，$\widehat{T}_J$ の値を定義する．
4．keisan3 関数により，$\widetilde{T}_J$ の分散の値を計算，出力する．
5．keisan4 関数により，S_B の値を計算，出力する．
6．main 関数にて以上のプログラムを実行し，有意水準 α(ALPHA) を入力，結果が出力される．

5　呼吸器疾患のデータに関する検定内容とその考察

飲酒，喫煙がそれぞれ呼吸器疾患に関係があるかどうかを調べるにあたり，男女合わせて 918 人分のデータ（文献 [3]）を使用した.このとき，男女ともに,『飲酒，喫煙ともになし』,『飲酒なし，喫煙あり』,『飲酒，喫煙ともにあり』の 3 標本についての検定を行った．それぞれの人数に関しては表 2 で表す．

表 2　男女における飲酒，喫煙の人数

飲酒 喫煙	なし なし	なし あり	あり あり	合計
男	141	133	185	459
（疾患あり）	(38)	(102)	(135)	(275)
女	118	74	124	316
（疾患あり）	(31)	(51)	(92)	(174)
合計	259	207	309	775
（疾患あり）	(69)	(153)	(227)	(449)

p_1 を『飲酒，喫煙ともになし』の場合の疾患の母比率，p_2を『飲酒なし，喫煙あり』の場合の母比率，p_3を『飲酒，喫煙ともにあり』の場合の疾患の母比率とする．上記の検定の結果，有意水準 α=0.01，0.05 のどちらの場合でも帰無仮説 H_0 は棄却された．ゆえに，飲酒および喫煙は呼吸器疾患に $p_1\leqq p_2\leqq p_3$(少なくとも 1 つの $\leqq$ は $\lneqq$ である) の関係があることがわかった．また，男，女，男女での結果が同じだったので性別はあまり関係していないことが考えられる．

6　食物アレルギーデータの検定内容とその考察

食物アレルギーが年齢に関係があるかどうかを調べるにあたり，20 歳代から 50 歳代あわせて 15766 人分のデータ (文献 [5]) を使用した．それぞれの人数に関しては表 3 で表す．

表 3　食物アレルギーの発症率

年代	人数	アレルギーあり
50 歳代	684	60
40 歳代	7877	756
30 歳代	7643	795
20 歳代	438	53

上記の検定の結果，有意水準 α=0.01，0.05 のどちらの場合でも帰無仮説 H_0 は棄却された．ゆえに，年齢が若い人ほど食物アレルギーの発症率が高いことがわかった．点

推定量 $\widehat{p}_1 = 0.088, \widehat{p}_2 = 0.096, \widehat{p}_3 = 0.104, \widehat{p}_4 = 0.121$ であり，増加関係がある．近年テレビゲームの普及が盛んになり，外で遊ぶ機会が減った．つまり，清潔すぎる環境で育ち，雑菌に触れる機会が減っている．腸内細菌の増殖が阻止されると，経口免疫寛容が形成されず食物アレルギーを誘発する可能性が高くなるので，若者の食物アレルギー発症率が高いのではないかと考察できる．

7　同時信頼区間

文献 [6] の 66 ページより，

$$(c.8) \qquad p_i^{n_i} \leqq \{1-(1-\alpha)^{\frac{1}{k}}\}/2, \quad (1-p_i)^{n_i} \leqq \{1-(1-\alpha)^{\frac{1}{k}}\}/2 \quad (i=1,\cdots,k)$$

を仮定する．(c.8) の仮定の下に，同時信頼区間は

$$\frac{L_i}{K_i \cdot F_{L_i}^{K_i}\left(\frac{1-(1-\alpha)^{\frac{1}{k}}}{2}\right) + L_i} < p_i < \frac{K_i^* \cdot F_{L_i^*}^{K_i^*}\left(\frac{1-(1-\alpha)^{\frac{1}{k}}}{2}\right)}{K_i^* \cdot F_{L_i^*}^{K_i^*}\left(\frac{1-(1-\alpha)^{\frac{1}{k}}}{2}\right) + L_i^*} \quad (i=1,\cdots,k). \quad (3)$$

で与えられる．ただし，$i = 1,\cdots,k$ に対して，$K_i \equiv 2(n_i - X_i + 1), L_i \equiv 2X_i, K_i^* \equiv 2(X_i+1), L_i^* \equiv 2(n_i - X_i)$ とおく．

7.1　呼吸器疾患のデータに関する解析結果

信頼区間について $\alpha = 0.05$（男）のときを考える．

$i = 1$ の場合は，$0.187 < p_1 < 0.365$
$i = 2$ の場合は，$0.670 < p_2 < 0.850$
$i = 3$ の場合は，$0.650 < p_3 < 0.802$

となる．このとき p_1 と p_2 は信頼区間が交わっていないため，p_1 と p_2 は異なることがわかる．同様に，p_1と p_3 も異なる．
また，$\alpha = 0.01$ のときも $0.165 < p_1 < 0.390$，$0.643 < p_2 < 0.865$，$0.623 < p_3 < 0.820$ なので結果は同じである．
$\alpha = 0.05$（女）のときを考える．

$i = 1$ の場合は，$0.173 < p_1 < 0.367$
$i = 2$ の場合は，$0.570 < p_2 < 0.808$
$i = 3$ の場合は，$0.640 < p_3 < 0.828$

となる．p_1 と p_2 は信頼区間が交わっていないため，p_1 と p_2 は異なることがわかる．同様に，p_1と p_3 も異なる．
また，$\alpha = 0.01$ のときも $0.153 < p_1 < 0.834$，$0.511 < p_2 < 0.834$，$0.611 < p_3 < 0.850$ なので結果は同じである．つづいて，$\alpha = 0.05$（男女）のときを考える．

$i = 1$ の場合は，$0.204 < p_1 < 0.340$
$i = 2$ の場合は，$0.662 < p_2 < 0.810$
$i = 3$ の場合は，$0.671 < p_3 < 0.790$

となる．このとき p_1 と p_2 は信頼区間が交わっていないため，p_1 と p_2 は異なることがわかる．同様に，p_1と p_3 も異なる．
また，$\alpha = 0.01$ のときも $0.190 < p_1 < 0.355$，$0.640 < p_2 < 0.824$，$0.654 < p_3 < 0.810$ なので結果は同じである．

7.2　食物アレルギーデータの解析結果

信頼区間について $\alpha = 0.05$ のときを考える．

$i = 1$ の場合は，$0.062 < p_1 < 0.120$
$i = 2$ の場合は，$0.088 < p_2 < 0.104$
$i = 3$ の場合は，$0.095 < p_3 < 0.113$
$i = 4$ の場合は，$0.084 < p_4 < 0.166$

となる．このとき p_1, p_2, p_3, p_4 すべての信頼区間が交わっている．Jonckheere-Terpstra 型検定では棄却されたので，どこかの群の間に違いがある．しかし，多重比較法では群と群の違いを見つけることはできなかった．
また，$\alpha = 0.01$ のとき $0.059 < p_1 < 0.124$，$0.086 < p_2 < 0.106$，$0.094 < p_3 < 0.115$，$0.079 < p_4 < 0.170$ なので結果は同じである．

8　おわりに

本論文では，多標本比率モデルにおける順序制約がある場合の Jonckheere-Terpstra 型検定法について考察した．C 言語プログラムを作成し結果を得ることができた．実際にプログラムを作成し現実のデータを用いることによって深めることができた．

参考文献

[1] 白石高章：『統計科学の基礎』日本評論社，東京，2012.

[2] 野澤慎：『多標本正規分布モデルにおける順序制約がある場合の Jonckheere-Terpstra 型検定法』南山大学情報理工学部情報システム数理学科卒業論文，愛知，2015 年 1 月.

[3] Daniel，W．W：『Biostatistics』Wiley，2008.

[4] 早川由宏：『Mathematica と C 言語による統計プログラミングの基礎』南山大学情報理工学部情報システム数理学科卒業論文，愛知，2013 年 1 月.

[5] 神奈川県衛生研究所：「神奈川県食物アレルギー実態調査」の概要，
http://www.eiken.pref.kanagawa.jp/001_event/pdf/060304pamphlet.pdf#search'

[6] 白石高章：『多群の 2 項モデルとポアソンモデルにおけるすべてのパラメータの多重比較法』日本統計学会誌，第 42 巻，第 1 号，55〜90 項，2012.