# 多標本比率モデルにおける順序制約がある場合の線形型検定法

2013SE105 桑原侑史　2013SE104 草野元樹
指導教員：白石高章

## 1　はじめに

3 年次までで統計学の基礎となる確率を学んできた．それを基に 4 年次から統計学の分布論や二項分布に関係した 1，2 標本モデルにおける，統計的解析法等を学んだ．線形型検定は薬学分野などでも使用されている．統計学が実際にどのように用いられているか知っていくにつれて線形型検定に興味を持ち，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線形型検定について考察する．

## 2　線形型順位検定

ある要因Aがあり，$k$ 個の水準 $A_1,\cdots,A_k$を考える．水準 $A_i$における標本の観測値 $(Y_{i1},Y_{i2},\cdots,Y_{in_i})$ を第 $i$ 標本とし $P(Y_{ij}\leqq x)=F(x-c_i\Delta), E(Y_{ij})=c_i\Delta$とする．ただし，$F(x)$ は連続型の分布関数とする．また，すべての $Y_{ij}$は互いに独立であると仮定する．（表 1 参照）．

表 1　$k$ 標本モデル

| 標本 | サイズ | データ | 平均 | 分布関数 |
|---|---|---|---|---|
| 第 1 標本 | $n_1$ | $Y_{11},\cdots,Y_{1n_1}$ | $c_1\Delta$ | $F(x-c_1\Delta)$ |
| 第 2 標本 | $n_2$ | $Y_{21},\cdots,Y_{2n_2}$ | $c_2\Delta$ | $F(x-c_2\Delta)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 第 $k$ 標本 | $n_k$ | $Y_{k1},\cdots,Y_{kn_k}$ | $c_k\Delta$ | $F(x-c_k\Delta)$ |

総標本サイズ：$n\equiv n_1+\cdots+n_k$(すべての観測値の個数)，$\Delta$ はすべて未知パラメータとする．なお，本研究に用いるモデルは文献 [1] を参考にして記述した．帰無仮説 $\mathcal{H}_0:\Delta=0$ vs. 対立仮説 $\mathcal{H}_1:\Delta>0$ のノンパラメトリックな線形型検定統計量は，$R_{ij}$ を $n$ 個すべての観測値 $X_{11}$ , $\cdots$，$X_{kn_k}$ を小さいほうから並べたときの $X_{ij}$ の順位とする．このとき，文献 [6] より，帰無仮説 $\mathcal{H}_0$ : vs. 対立仮説 $\mathcal{H}_1$ に対する検定統計量は，

$$S_R\equiv\sum_{i=1}^{k}c_i\sum_{j=1}^{n_i}R_{ij} \tag{1}$$

で与えられる．次の (条件 1) を仮定する．

(条件 1)　　$\lim_{n\to\infty}\dfrac{n_i}{n}=\lambda_i>0\ (i=1,\cdots,k).$

$\mathcal{H}_0$ の下で，$S_R$ の平均を求める．文献 [1] より，

$$E_0(S_R)=\frac{n+1}{2}\sum_{i=1}^{k}c_in_i$$

ここで，文献 [2] より，

$$\hat{S}_{R_i}\equiv\sqrt{\frac{12}{n+1}}\left(\bar{R}_i-\frac{n+1}{2}\right)$$

とおくと，

$$S_R-E_0(S_R)=\sum_{i=1}^{k}d_i\hat{S}_{R_i} \tag{2}$$

となる．(2) を満たす $d_i$ を求めると，

$$\begin{aligned}((2)\text{ の左辺})&=\sum_{i=1}^{k}c_i\sum_{j=1}^{n_i}R_{ij}-\frac{n+1}{2}\sum_{i=1}^{k}n_ic_i\\&=\sum_{i=1}^{k}n_ic_i\left(\frac{1}{n_i}\sum_{j=1}^{n_i}R_{ij}-\frac{n+1}{2}\right)\\((2)\text{ の右辺})&=\sum_{i=1}^{k}d_i\sqrt{\frac{12}{n+1}}\left(\bar{R}_{ij}-\frac{n+1}{2}\right)\\&=\sum_{i=1}^{k}d_i\sqrt{\frac{12}{n+1}}\left(\frac{1}{n_i}\sum_{j=1}^{n}R_{ij}-\frac{n+1}{2}\right)\end{aligned}$$

である．したがって，

$$n_ic_i=d_i\sqrt{\frac{12}{n+1}}$$

になればよい．つまり，

$$d_i=\frac{n_ic_i\sqrt{12(n+1)}}{12}$$

である．ここで，

$$\lim_{n\to\infty}\frac{d_i}{n\sqrt{n}}=\frac{1}{\sqrt{12}}c_i\lambda_i$$

となる．$Z_i=\sqrt{\lambda_i}Y_i$ とおくと，

$$\begin{aligned}\frac{S_R-E_0(S_R)}{n\sqrt{n}}&=\frac{1}{n\sqrt{n}}\sum_{i=1}^{k}d_i\hat{S}_{R_i}\\&\xrightarrow{\mathcal{L}}\frac{1}{\sqrt{12}}\sum_{i=1}^{k}c_i\lambda_i\left(Y_i-\sum_{j=1}^{k}\lambda_j\frac{Z_j}{\sqrt{\lambda_j}}\right)\\&=\frac{1}{\sqrt{12}}\left(\sum_{i=1}^{k}c_i\sqrt{\lambda_i}Z_i-\sum_{i=1}^{k}c_i\lambda_i\sum_{j=1}^{k}\sqrt{\lambda_j}Z_j\right)\end{aligned} \tag{3}$$

が導かれる．

## 3　$k$ 標本比率モデルと漸近的性質の基礎

水準 $A_i$ における標本の観測値を $(X_{i1}, X_{i2}, \cdots, X_{in_i})$ とし，$X_{ij}$ は成功の確率が $p_i$ のベルヌーイ試行とする．すなわち，$X_{ij}$〜$B(1, p_i)$ 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．）を考える．$H_0$の下で，$p_i = p_0 \ \ (i = 1, \cdots, k)$ とする．$p_0$は未知である．

表 2　$k$ 標本比率モデル

| 標本 | サイズ | データ | $X_{i\cdot}$ の分布 |
|---|---|---|---|
| 第 1 標本 | $n_1$ | $X_{11}, \cdots, X_{1n_1}$ | $B(n_1, p_1)$ |
| 第 2 標本 | $n_2$ | $X_{21}, \cdots, X_{2n_2}$ | $B(n_2, p_2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 第 $k$ 標本 | $n_k$ | $X_{k1}, \cdots, X_{kn_k}$ | $B(n_k, p_k)$ |

$p_i$ の点推定量は

$$\hat{p}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \tag{4}$$

である．このとき，

$$\sqrt{n_i}\{\arcsin(\sqrt{\hat{p}_i}) - \arcsin(\sqrt{p_i})\} \xrightarrow{\mathcal{L}} W \sim N\left(0, \frac{1}{4}\right)$$

が成り立つ．さらに，

$$E\left(\hat{p}_i\right) = p_i,\ V\left(\hat{p}_i\right) = \frac{1}{n_i} \cdot p_i(1 - p_i)$$

が成り立つ．

## 4　提案する検定

$$\bar{c} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$$

とおく．また，(条件 1) の下で，$n \longrightarrow \infty$ として，

$$c' \equiv \sum_{i=1}^{k} \lambda_i c_i$$

となる．

$$S_p \equiv \sum_{i=1}^{k} (c_i - \bar{c}) n_i \arcsin(\sqrt{\hat{p}_i})$$

とおく，

**定理 1**　$H_0$ の下で，$p_i = p_0 (i = 1, \cdots k)$ であるので

$$T_p = \sum_{i=1}^{k} (c_i - \bar{c}) n_i \left\{\arcsin(\sqrt{\hat{p}_i}) - \arcsin(\sqrt{p_0})\right\}$$

とおくと，$T_p = S_p$ が成り立つ．

**証明**

$$\begin{aligned} T_p &= \sum_{i=1}^{k} (c_i - \bar{c}) n_i \left\{\arcsin(\sqrt{\hat{p}_i}) - \arcsin(\sqrt{p_0})\right\} \\ &= \sum_{i=1}^{k} (c_i - \bar{c}) n_i \arcsin(\sqrt{\hat{p}_i}) - \sum_{i=1}^{k} (c_i - \bar{c}) n_i \arcsin(\sqrt{p_0}) \end{aligned}$$

となる．ここで，

$$\begin{aligned} & S_p - T_p \\ &= \sum_{i=1}^{k} (c_i - \bar{c}) n_i \arcsin(\sqrt{p_0}) \\ &= \sum_{i=1}^{k} c_i n_i \arcsin(\sqrt{p_0}) - \sum_{i=1}^{k} n_i \arcsin(\sqrt{p_0}) \frac{1}{n} \sum_{i'=1}^{k} n_{i'} c_{i'} \\ &= \sum_{i=1}^{k} c_i n_i \arcsin(\sqrt{p_0}) - n \arcsin(\sqrt{p_0}) \frac{1}{n} \sum_{i'=1}^{k} n_{i'} c_{i'} \\ &= \sum_{i=1}^{k} c_i n_i \arcsin(\sqrt{p_0}) - \sum_{i'=1}^{k} n_{i'} c_{i'} \arcsin(\sqrt{p_0}) \\ &= 0 \end{aligned}$$

となり，$T_p = S_p$ が示された．　□

したがって，文献 [1] の系 3.6 より，

$$\begin{aligned} \frac{s_p}{\sqrt{n}} &= \frac{\displaystyle\sum_{i=1}^{k} (c_i - \bar{c}) n_i \left\{\arcsin(\sqrt{\hat{p}_i}) - \arcsin(\sqrt{p_0})\right\}}{\sqrt{n}} \\ &\xrightarrow{\mathcal{L}} \sum_{i=1}^{k} \sqrt{\lambda_i}(c_i - c') W_i \qquad (5) \\ &\sim N\left(0, \frac{1}{4} \sum_{i=1}^{k} \left\{(c_i - c')\sqrt{\lambda_i}\right\}^2\right) \end{aligned}$$

となるので，

$$B_n \equiv \frac{2}{\sqrt{\displaystyle\sum_{i=1}^{k} \left\{\sqrt{\frac{n_i}{n}}(c_i - \bar{c})\right\}^2}}$$

とおくと，

$$\frac{S_p \cdot B_n}{\sqrt{n}} \xrightarrow{\mathcal{L}} N(0, 1) \tag{6}$$

である．

したがって，$c_1, \cdots, c_k$ が既知のとき，検定統計量を

$$T_{\mathbf{c}} \equiv \frac{2\displaystyle\sum_{i=1}^{k} (c_i - \bar{c}) n_i \left\{\arcsin(\sqrt{\hat{p}_i})\right\}}{\sqrt{\displaystyle\sum_{i=1}^{k} n_i (c_i - \bar{c})^2}} \tag{7}$$

とし，未知のときは，

$$T_{\mathbf{0}} \equiv \frac{2\sum_{i=1}^{k}(i-\bar{c}_0)\,n_i\left\{\arcsin(\sqrt{\hat{p}_i})\right\}}{\sqrt{\sum_{i=1}^{k} n_i\,(i-\bar{c}_0)^2}} \tag{8}$$

ただし，$\bar{c}_0 = \frac{1}{n}\sum_{i=1}^{k} i n_i$ とする．
ここで，(7) より，

$$\begin{aligned}\frac{S_p}{\sqrt{n}} &\xrightarrow{\mathcal{L}} \sum_{i=1}^{k}(c_i-c')\sqrt{\lambda_i}W_i \\ &= \sum_{i=1}^{k} c_i\sqrt{\lambda_i}W_i - \sum_{i=1}^{k} c'\sqrt{\lambda_i}W_i \\ &= \sum_{i=1}^{k} c_i\sqrt{\lambda_i}W_i - \sum_{i=1}^{k}\lambda_i c_i \cdot \sum_{j=1}^{k}\sqrt{\lambda_j}W_j \end{aligned} \tag{9}$$

を得る．(3)，(9) より順位検定統計量と提案する検定統計量の漸近的表現が一致する．

## 5 検定方式

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

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，(6) より，(条件 1) の下で，

$$\lim_{n\to\infty} P_0(T_{\boldsymbol{c}} > z(\alpha)) = \alpha \tag{10}$$

であるので，
$(i)$　$c_1,\cdots,c_k$ が既知のとき，検定関数 $\phi(\cdot)$ を

$$\phi(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_k) = \begin{cases} 1 & (T_{\boldsymbol{c}} > z(\alpha) \text{ のとき}) \\ 0 & (T_{\boldsymbol{c}} < z(\alpha) \text{ のとき}) \end{cases} \tag{11}$$

で定義すれば，

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

が得られる．
$(ii)$　$c_1,\cdots,c_k$ が未知のとき，検定関数 $\phi(\cdot)$ を

$$\phi(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_k) = \begin{cases} 1 & (T_{\mathbf{0}} > z(\alpha) \text{ のとき}) \\ 0 & (T_{\mathbf{0}} < z(\alpha) \text{ のとき}) \end{cases} \tag{12}$$

で定義する．

## 6 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のある場合の線形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();
    keisan5();
    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てください\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(T>XN){
            printf("H0 を棄却する\n");
        }

     else{printf("H0 を棄却しない\n");
        }
    return(0);
}
```

### プログラムの流れ

1. input 関数により，データをテキストファイルから読み込み, 標本数, サイズを計算かつ表示する．
2. keisan 関数により，$\arcsin(\sqrt{\hat{p}_i})$ の値を計算する．
3. keisan2 関数により，$\bar{c}$ の値を計算する．
4. keisan3 関数により，$\sqrt{\sum_{i=1}^{k} n_i(c_i-\bar{c})^2}$ 値を計算する．
5. keisan4 関数により，$2\sum_{i=1}^{k} n_i(c_i-\bar{c})n_i\{\arcsin(\sqrt{\hat{p}_i})\}$ の値を計算する．
6. keisan5 関数により，検定統計量の値を計算する．
7. main 関数により，検定結果を出力する．

## 7　糖尿病患者のデータに関する検定内容とその考察

糖尿病患者が年齢が上がるにつれて発症する人が多くなるのか調べるにあたり, 男女合わせて 3022 人分のデータ（文献 [5]）を使用した．このとき, 男女ともに『40 代』,『50 代』,『60 代』,『70 代』の 4 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 3 で表す．

表 3　男女における年代別糖尿病患者の人数

| 糖尿病患者 | 40 代 | 50 代 | 60 代 | 70 代 |
|---|---|---|---|---|
| 男 | 160 | 217 | 418 | 512 |
| （疾患あり） | (6) | (23) | (84) | (114) |
| 女 | 277 | 339 | 505 | 594 |
| （疾患あり） | (5) | (22) | (67) | (101) |
| 合計 | 437 | 556 | 923 | 1106 |
| （疾患あり） | (11) | (45) | (151) | (215) |

$p_1$ を『40 代』の場合の糖尿病患者の母比率，同様に，$p_2$，$p_3$，$p_4$ を糖尿病患者の母比率であらわす．上記の検定の結果，有意水準 $\alpha$=0.05，0.01 のどちらの場合でも帰無仮説 $H_0$ は棄却された．ゆえに，年齢は，糖尿病患者に対して，増加関係があることがわかった．糖尿病は年齢を重ねるにつれて，糖尿病に発病する人が多くなると考察できる．

## 8　同時信頼区間

文献 [3] の 66 ページより，

$$\frac{L_i}{K_i \cdot F_{L_i}^{K_i}\left(\frac{1-(1-\alpha)^{\frac{1}{k}}}{2}\right)+L_i} < p_i < \frac{K_i^* \cdot F_{L_i^*}^{K_i^*}\left(\frac{1-(1-\alpha)^{\frac{1}{k}}}{2}\right)}{K_i^* \cdot F_{L_i^*}^{K_i^*}\left(\frac{1-(1-\alpha)^{\frac{1}{k}}}{2}\right)+L_i^*} \quad (i=1,\cdots,k). \quad (13)$$

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．

### 糖尿病患者についての解析

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，$0.010 < p_1 < 0.092$
$i=2$ の場合は，$0.060 < p_2 < 0.189$
$i=3$ の場合は，$0.155 < p_3 < 0.270$
$i=4$ の場合は，$0.195 < p_4 < 0.272$

となる．このとき $p_1$ と $p_3$ は信頼区間が交わっていないため，$p_1$ と $p_3$ は異なることがわかる．同様に，$p_1$ と $p_4$ も異なる．
また，$\alpha = 0.01$ のときは $0.007 < p_1 < 0.106$，$0.053 < p_2 < 0.183$，$0.146 < p_3 < 0.266$，$0.170 < p_4 < 0.282$ となる．$p_1$ と $p_2$ は，信頼区間が交わっているので，あまり差がないことが分かる．しかし，p1 と p3，p4 では信頼区間が交わっていないため，年齢が離れるほど異なることが分かる．$\alpha = 0.05$（女）のときを考える．

$i=1$ の場合は，$0.004 < p_1 < 0.049$
$i=2$ の場合は，$0.069 < p_2 < 0.195$
$i=3$ の場合は，$0.098 < p_3 < 0.174$
$i=4$ の場合は，$0.134 < p_4 < 0.211$

となる．このとき $p_1$ と $p_2$ は異なることがわかる．同様に，$p_1$ と $p_3$，$p_1$ と $p_4$ も異なる．しかし，$p_2$ と $p_3$，$p_3$ と $p_4$ は信頼区間が交わっているため，50 代と 60 代，60 代と 70 代では差があまりないことが分かった．
また，$\alpha = 0.01$ のときも $0.003 < p_1 < 0.057$，$0.060 < p_2 < 0.212$，$0.091 < p_3 < 0.183$，$0.126 < p_4 < 0.220$ なので，$\alpha$=0.05 のときと結果は同じである．
$\alpha = 0.05$（男女）のときを考える．

$i=1$ の場合は，$0.010 < p_1 < 0.050$
$i=2$ の場合は，$0.055 < p_2 < 0.114$
$i=3$ の場合は，$0.135 < p_3 < 0.196$
$i=4$ の場合は，$0.166 < p_4 < 0.225$

となる．このとき $p_1$ と $p_2$ は異なることがわかる．同様に，$p_1$ と $p_3$，$p_1$ と $p_4$ も異なる．しかし，$p_3$ と $p_4$ は交わっているため，あまり差がないことが分かった．
また，$\alpha = 0.01$ のときも $0.008 < p_1 < 0.057$，$0.050 < p_2 < 0.122$，$0.128 < p_3 < 0.203$，$0.159 < p_4 < 0.232$ なので，$\alpha$=0.05 のときと結果は同じである．

## 9　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線形型検定法について考察した．プログラムを作成し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] 白石高章:『多群連続モデルにおける多重比較法』
共立出版, 東京, 2011.

[3] 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.

[4] 早川由宏，白石高章:『Fortran と C 言語による統計プログラミングの基礎 Mathematica の使い方』
研究ノート,2015 年 2 月.

[5] 糖尿病ネットワーク
http://www.dm-net.co.jp/calendar/2015/024529/php

[6] Hájek，J.，Šidák，Z.and Sen，P.K.
*Theory of Rank Tests*，2nd Edition
Academic Press，1999.