多標本ポアソンモデルにおける順序制約がある場合の Jonckheere-Terpstra 型検定法

2012SE067 石崎幸市　2012SE278 山田智己
指導教員：白石高章

1　はじめに

ポアソン分布に従う観測値からなるデータは，地震の回数，交通事故の件数など，身の回りに多く存在する．そこで，我々はノンパラメトリックにおける順序制約がある場合の検定法として，よく知られている Jonckheere-Terpstra 検定に興味を持ち，本研究を行うことを決めた．文献 [1] を参照しノンパラメトリックモデルをポアソンモデルに変えた Jonckheere-Terpstra 型検定について考察する．

2　正規標本モデルでの Jonckheere-Terpstra 型検定統計量

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，$Y_{ij}\sim N(\mu_i,\sigma^2)$ 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．

$$\begin{cases} \text{帰無仮説} \quad H_0:\mu_1=\cdots=\mu_k \\ \text{対立仮説} \quad H_1:\mu_1\leqq\cdots\leqq\mu_k \\ \qquad\quad (\text{少なくとも 1 つの}\leqq\text{は}\lneqq\text{である}) \end{cases} \tag{1}$$

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

とおく．ただし，

$$T_J=\sum_{i'=2}^{k}\sum_{i=1}^{i'-1}n_in_{i'}\left(\bar{Y}_{i'\cdot}-\bar{Y}_{i\cdot}\right), \tag{2}$$

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

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

とする．文献 [3] により，S_N が正規標本モデルでの Jonckheere-Terpstra 型検定統計量である．

3　k 標本ポアソンモデル

ある要因 A があり，k 個の水準 $A_1,\cdots,A_k$ を考える．水準は標本とも呼ばれる．水準 A_i における標本の観測値 $(X_{i1},\cdots,X_{in_i})$ は第 i 標本または第 i 群と呼ばれ，平均が μ_i である．表 1 の X_{ij} がポアソン分布 $\mathcal{P}_o(\mu_i)$ に従うものとする．さらに，全ての X_{ij} は互いに独立であるとする．すなわち，

$$X_{ij}\sim\mathcal{P}_o(\mu_i)\quad,\quad(j=1,\cdots,n_i\ ,\ i=1,\cdots,k)$$

表 1　k 標本ポアソンモデル

標本	サイズ	データ	平均	分布関数
第 1 標本	n_1	$X_{11},\cdots,X_{1n_1}$	μ_1	$\mathcal{P}_o(\mu_1)$
第 2 標本	n_2	$X_{21},\cdots,X_{2n_2}$	μ_2	$\mathcal{P}_o(\mu_2)$
⋮	⋮	⋮	⋮	⋮
第 k 標本	n_k	$X_{k1},\cdots,X_{kn_k}$	μ_k	$\mathcal{P}_o(\mu_k)$

総標本サイズ：$n\equiv n_1+\cdots+n_k$（すべての観測値の個数）

$\mu_1,\cdots,\mu_k$ はすべて未知パラメータとする．

ただし，

$$W_i=X_{i1}+X_{i2}+\cdots+X_{in_i}$$

とする．このとき，μ_i の点推定量は，

$$\widehat{\mu}_i\equiv\frac{W_i}{n_i}\,(i=1,\cdots,k)$$

で与えられる．$n\equiv n_1+\cdots+n_k$ とおき，

(C1)　$$0<\lim_{n\to\infty}\frac{n_i}{n}=\lambda_i<1$$

を仮定する．
$\widehat{\sigma}_i\equiv\sqrt{\widehat{\mu}_i}$ とすると，文献 [1] より，

$$2\sqrt{n_i}\,(\widehat{\sigma}_i-\sigma_i)\xrightarrow{\mathcal{L}}Z\sim N(0,1)$$

が成り立つ．ここで，$\widehat{\sigma}_i$ を $\sigma_i\equiv\sqrt{\mu_i}$ の推定量である．

4　提案する検定法

(1) の，H_0の下で$\mu_i=\mu_0\,(i=1,\cdots,k)$ とすると，$\sigma_0=\sqrt{\mu_0}$ となる．

$$\widehat{Z}_i\equiv2\sqrt{n_i}\,(\widehat{\sigma}_i-\sigma_0)\xrightarrow{\mathcal{L}}Z_i\sim N(0,1)$$

である．上式は，

$$\widehat{\sigma}_i-\sigma_0=\frac{\widehat{Z}_i}{2\sqrt{n_i}} \tag{3}$$

のような関係があり，$Z_1,\cdots,Z_k$ は互いに独立である．また，

$$V\left(\bar{Y}_{i\cdot}\right)=V\left(\frac{1}{n_i}\sum_{j=1}^{n_i}Y_{ij}\right)=\frac{\sigma^2}{n_i}$$

で与えられる．ここで，

$$\frac{1}{\sqrt{n_i}} \cdot \widetilde{Z}_i \equiv \left(\bar{Y}_i \cdot -\mu_0\right) \sim N\left(0, \frac{\sigma^2}{n_i}\right)$$

とおく．さらに，式 (2) を $\widetilde{Z}_i$ を用いて以下のように表現できる．

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

ここで，$i = 1, \cdots, k$ に対して$\widetilde{Z}_i$の代わりに, $\widehat{Z}_i$を代入したものを$\widehat{T}_J$とする．(3) より，

$$\begin{aligned}
\widehat{T}_J &= \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} n_i \cdot n_{i'} \left(\frac{\widehat{Z}_{i'}}{\sqrt{n_{i'}}} - \frac{\widehat{Z}_i}{\sqrt{n_i}} \right) \\
&= \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} n_i \cdot n_{i'} \left\{ 2\left(\widehat{\sigma}_{i'} - \sigma_0\right) - 2\left(\widehat{\sigma}_i - \sigma_0\right) \right\} \\
&= 2 \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} n_i \cdot n_{i'} \left(\widehat{\sigma}_{i'} - \widehat{\sigma}_i\right)
\end{aligned}$$

となる．また，$\widehat{Z}_i \xrightarrow{\mathcal{L}} Z_i$ ，(3) より，

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

とおく．$\widetilde{T}_J$ の平均は，$E\left(Z_{i'}\right) = 0$, $E\left(Z_i\right) = 0$ より

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

である．ここで，$T_{i'} \equiv n_{i'} \sum_{i=1}^{i'-1} n_i \left(\frac{Z_{i'}}{\sqrt{n_{i'}}} - \frac{Z_i}{\sqrt{n_i}} \right)$ とおくと，

$$\begin{aligned}
T_{i'} &= \left(\frac{n_{i'}}{\sqrt{n_{i'}}} \cdot \sum_{i=1}^{i'-1} n_i \right) Z_{i'} - \left(n_{i'} \sum_{i=1}^{i'-1} \frac{n_i}{\sqrt{n_i}} Z_i \right) \\
&= \left(\sqrt{n_{i'}} \sum_{i=1}^{i'-1} n_i \right) Z_{i'} - \left(n_{i'} \sum_{i=1}^{i'-1} \sqrt{n_i} \cdot Z_i \right)
\end{aligned}$$

となる．T_ℓ の平均は，

$$\begin{aligned}
E\left(T_\ell\right) &= \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) \cdot E\left(Z_\ell\right) \\
&\quad - \left(n_\ell \sum_{i=1}^{\ell-1} \sqrt{n_i} \cdot E\left(Z_i\right) \right) \\
&= 0
\end{aligned}$$

である．ここで，上式の $E(T_\ell) = 0$ を用いると，$\ell < \ell'$ に対して，

$$\begin{aligned}
&\mathrm{Cov}\left(T_\ell, T_{\ell'}\right) \\
&\quad = E\left[\left\{ T_\ell - E\left(T_\ell\right) \right\} \left\{ T_{\ell'} - E\left(T_{\ell'}\right) \right\} \right] \\
&\quad = E\left(T_\ell \cdot T_{\ell'}\right) \\
&\quad = E\left\{ \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) Z_\ell \cdot Z_{\ell'} \right\} \\
&\qquad - E\left\{ \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) \left(n_{\ell'} \sum_{i=1}^{\ell'-1} \sqrt{n_i} \cdot Z_i \right) Z_\ell \right\} \\
&\qquad - E\left\{ \left(n_\ell \sum_{i=1}^{\ell-1} \sqrt{n_i} \cdot Z_i \right) \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) Z_{\ell'} \right\} \\
&\qquad + E\left\{ \left(n_\ell \sum_{i=1}^{\ell-1} \sqrt{n_i} \cdot Z_i \right) \left(n_{\ell'} \sum_{i=1}^{\ell'-1} \sqrt{n_i} \cdot Z_i \right) \right\}
\end{aligned}$$

である．ここで，$\ell < \ell'$ より，Z_ℓと $Z_{\ell'}$ は互いに独立なので，

$$\begin{aligned}
&E\left\{ \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) Z_\ell \cdot Z_{\ell'} \right\} \\
&= \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) E\left(Z_\ell \cdot Z_{\ell'}\right) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
&E\left\{ \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) \left(n_{\ell'} \sum_{i=1}^{\ell'-1} \sqrt{n_i} \cdot Z_i \right) Z_\ell \right\} \\
&= \left(\sqrt{n_\ell} \sum_{i=1}^{\ell-1} n_i \right) n_{\ell'} \cdot \sum_{i=1}^{\ell'-1} \left\{ \sqrt{n_i} \cdot E\left(Z_i \cdot Z_\ell\right) \right\} \\
&= n_\ell \cdot n_{\ell'} \left(\sum_{i=1}^{\ell-1} n_i \right)
\end{aligned}$$

$$\begin{aligned}
&E\left\{ \left(n_\ell \sum_{i=1}^{\ell-1} \sqrt{n_i} \cdot Z_i \right) \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) Z_{\ell'} \right\} \\
&= n_\ell \left(\sqrt{n_{\ell'}} \sum_{i=1}^{\ell'-1} n_i \right) \sum_{i=1}^{\ell-1} \left\{ \sqrt{n_i} \cdot E\left(Z_i\right) \cdot E\left(Z_\ell\right) \right\} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
&E\left\{ \left(n_\ell \sum_{i=1}^{\ell-1} \sqrt{n_i} \cdot Z_i \right) \left(n_{\ell'} \sum_{i'=1}^{\ell'-1} \sqrt{n_i} \cdot Z_i \right) \right\} \\
&= n_\ell \cdot n_{\ell'} \sum_{i=1}^{\ell-1} \sum_{i'=1}^{\ell'-1} E\left(\sqrt{n_i} \cdot \sqrt{n_{i'}} \cdot Z_i \cdot Z_{i'} \right) \\
&= n_\ell \cdot n_{\ell'} \sum_{i=1}^{\ell-1} n_i
\end{aligned}$$

である．したがって，

$$\begin{aligned}\mathrm{Cov}\left(T_\ell, T_{\ell'}\right) &= 0 - n_\ell \cdot n_{\ell'} \sum_{i=1}^{\ell-1} n_i - 0 + n_\ell \cdot n_{\ell'} \sum_{i=1}^{\ell-1} n_i \\ &= 0\end{aligned}$$

となる．$\ell > \ell'$ の場合も同様に，$\mathrm{Cov}\left(T_\ell, T_{\ell'}\right) = 0$ である．ここで，$V(T_{i'})$ を計算すると，

$$\begin{aligned}V\left(T_{i'}\right) &= V\left\{\left(\sqrt{n_{i'}} \sum_{i=1}^{i'-1} n_i\right) Z_{i'} - \left(n_{i'} \sum_{i=1}^{i'-1} \sqrt{n_i} \cdot Z_i\right)\right\} \\ &= n_{i'} \left(\sum_{i=1}^{i'-1} n_i\right)^2 + n_{i'}^2 \cdot \sum_{i=1}^{i'-1} n_i \\ &= \left(\sqrt{n_{i'}} \sum_{i=1}^{i'-1} n_i\right)^2 + n_{i'}^2 \cdot \sum_{i=1}^{i'-1} n_i V\left(Z_i\right) \\ &= n_{i'} \left(\sum_{i=1}^{i'-1} n_i\right) \left(\sum_{i=1}^{i'} n_i\right)\end{aligned}$$

である．これらの結果と，$\widetilde{T}_J = \sum_{i'=2}^{k} T_{i'}$ より，

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

が得られる．さらに，文献 [1] の命題 2.16，定理 2.19 より，

$$\begin{aligned}E\left(\frac{\widetilde{T}_J - E(\widetilde{T}_J)}{\sqrt{V(\widetilde{T}_J)}}\right) &= E\left(\frac{\widetilde{T}_J}{\sqrt{V(\widetilde{T}_J)}}\right) = 0 \\ V\left(\frac{\widetilde{T}_J - E(\widetilde{T}_J)}{\sqrt{V(\widetilde{T}_J)}}\right) &= V\left(\frac{\widetilde{T}_J}{\sqrt{V(\widetilde{T}_J)}}\right) = 1\end{aligned}$$

である．(1) の H_0 の下で，文献 [1] の系 3.6 より，

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

である．よって，H_0 の下で，

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

が成り立つ．そこで，検定統計量を

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

とすると，H_0 の下で，$S_K \xrightarrow{\mathcal{L}} N(0,1)$ である．
検定統計量 S_K について，検定方式を考える．
$\boldsymbol{X}_i \equiv (X_{i1}, \cdots, X_{in_i})(i = 1, \cdots, k)$ とおく．
標準正規分布の上側 $100\alpha\%$ 点を $z(\alpha)$ とすると，条件 (C1) の下で，$\lim_{n\to\infty} P_0(S_K > z(\alpha)) = \alpha$ である．これにより，検定関数 $\phi(\cdot)$ を

$$\phi(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k) = \begin{cases} 1 & (S_K > z(\alpha) \text{ のとき}) \\ 0 & (S_K < z(\alpha) \text{ のとき}) \end{cases} \tag{4}$$

で定義すれば，

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

より，(4) による検定方式は水準 α の漸近的な検定である．

5 C 言語によるプログラム解説

5.1 プログラムの解説

C 言語により，Jonckheere-Terpstra 型検定による検定結果及び，ポアソン分布に従う変数を用いた検定結果を作成した．ただし，上側 $100\alpha\%$ 点を求めるために文献 [5] を引用した．本研究で作成したプログラムの main プログラムは，

```
int main(void){
    input();
    keisan1();
    keisan2();
    keisan3();
    keisan4();
    keisan5();
    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("H_0:μ_1 ＝・・・＝μ_k\n");
    printf("H_1:μ_1 ＜＝・・・＜＝μ_k(少なくとも
1 つの＜＝は＜≠である )\n");
    printf("ポアソン分布における
            Jonckheere-Terpstra 型検定\n");
    if(SK>XN){
        printf("H_0 を棄却する\n");
    }
    else{
        printf("H_0 を棄却しない\n");
    }
    return(0);
}
```

である．

5.2 プログラムの流れ

1. input 関数の中で，標本数，標本サイズ，データ，有意水準を入力する．
2. keisan1 関数の中で，μ_i の値を計算する．
3. keisan2 関数の中で，$\hat{\sigma}$ の値を計算する．
4. keisan3 関数の中で，$\widehat{T_J}$ の値を計算する．
5. keisan4 関数の中で，S_K の分母の値を計算する．
6. keisan5 関数の中で，S_K の値を計算する．
7. main 関数にて以上のプログラムを実行し，有意水準 α を入力，Jonckheere-Terpstra 型検定の結果を表示する．

6 交通事故死亡者数のデータとその解析結果

6.1 交通事故年間死亡者数のデータと一日あたりの平均交通事故死亡者数

文献 [2] より，交通事故における事象はポアソン分布に従う．そこで交通事故死亡者数のデータをもとに，検定を行った．まず，各都道府県ごとに平成２３年から２６年の年度別交通事故死亡者数を調べて，データ入力を行った．２３年から２６年にかけて，埼玉，新潟，愛知，滋賀，大阪の一日あたりの平均交通事故死亡者数が単調減少であったため、これらの県のデータ入力をする．尚，交通事故死亡者数のデータは文献 [4] より引用している．

表 2　交通事故死亡者数

府県名	平成 23 年	平成 24 年	平成 25 年	平成 26 年
埼玉県	207	200	180	173
新潟県	133	107	107	103
愛知県	276	235	219	204
滋賀県	85	79	74	63
大阪府	197	182	179	143

表 3　一日あたりの平均交通事故死亡者数

府県名	平成 23 年	平成 24 年	平成 25 年	平成 26 年
埼玉県	0.58	0.55	0.50	0.47
新潟県	0.36	0.29	0.29	0.28
愛知県	0.76	0.64	0.60	0.56
滋賀県	0.23	0.22	0.20	0.17
大阪府	0.54	0.50	0.49	0.39

平成 24 年は 366 日で計算されている．

6.2 実行結果の例

μ_1 を平成 26 年の一日あたりの平均交通事故死亡者数，μ_2 を平成 25 年の一日あたりの平均交通事故死亡者数，μ_3 を平成 24 年の一日あたりの平均交通事故死亡者数，μ_4 を平成 23 年の一日あたりの平均交通事故死亡者数とする．愛知県のデータについて，α=0.05，α=0.01 で検定した場合それぞれ以下のようになる．

```
標本数の入力 4
SK=3.354707
誤差 0.000010 の標準正規分布の上側 5.000000 パーセント点は 1.644859
H_0:μ_1 ＝・・・＝μ_k
H_1:μ_1 ＜＝・・・＜＝μ_k(少なくとも 1 つの＜＝は＜ ≠ である )
ポアソン分布における Jonckheere-Terpstra 型検定
H_0 を棄却する
標本数の入力 4
SK=3.354707
誤差 0.000010 の標準正規分布の上側 1.000000 パーセント点は 2.326347
H_0:μ_1 ＝・・・＝μ_k
H_1:μ_1 ＜＝・・・＜＝μ_k(少なくとも 1 つの＜＝は＜ ≠ である )
ポアソン分布における Jonckheere-Terpstra 型検定
H_0 を棄却する
```

6.3 解析結果とその考察

6.2 節で行った愛知県の解析と同様に，埼玉，新潟，滋賀，大阪のデータについて有意水準 $\alpha = 0.05$ を用いてそれぞれ Jonckheere-Terpstra 型検定を行った．その結果すべての府県で H_0 を棄却することがわかった．次に有意水準 $\alpha = 0.01$ で検定を行ったところ，愛知，大阪は H_0 を棄却した．一方，埼玉，新潟，滋賀は H_0 を棄却しなかった．この結果より，交通事故による死亡者数は減少傾向にあることが示された．特に愛知，大阪のような大都市は，交通の取り締まりや，飲酒運転の減少，自動制御プログラムを搭載した自動車の普及に伴い，減少傾向にあると考察できる．

7 おわりに

本論では, ポアソンモデルにおける Jonckheere-Terpstra 型検定を提案した．また，C 言語によって作成したプログラムによって，同様の結果を得られた．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 Jonckheere-Terpstra 検定とそれに基づくポアソンモデルにおける Jonckheere-Terpstra 型検定に対する理解をより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] 杉山高一ら :「統計データ科学辞典」. 朝倉書店，東京，2010

[3] 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，(2016 年 1 月)

[4] シンク出版 : http://www.think-sp.com/

[5] 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（2013 年 1 月）