多標本ポアソンモデルにおける順序制約がある場合の線形型検定法

2013SE113 松島七海　2013SE116 光田卓矢

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，よく知られている線形型検定に興味を持ち，本研究を行うことを決めた．文献 [1]，[6] 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線形型検定について考察する．

2　線形型順位検定統計量

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 $P(Y_{ij} \leqq x) = F(x - c_i\Delta)$, $E(Y_{ij}) = c_i\Delta$とする．ただし,$\mathrm{F}(\cdot)$は連続分布関数とする．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	$Y_{11}, \cdots, Y_{1\mathrm{n}_1}$	$c_1\Delta$	$F(x - c_1\Delta)$
第 2 標本	n_2	$Y_{21}, \cdots, Y_{2\mathrm{n}_2}$	$c_2\Delta$	$F(x - c_2\Delta)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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)，$\Delta$ は未知パラメータとする．
帰無仮説 $\mathcal{H}_0 : \Delta = 0$　vs. 対立仮説 $\mathcal{H}_A : \Delta > 0$ のノンパラメトリックな線形順位検定統計量は，文献 [6] より

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

で与えられる．ただし，R_{ij} を n 個すべての観測値 X_{11},$\cdots$,X_{kn_k} を小さい方から並べたときの X_{ij} の順位とする．
帰無仮説 H_0 の下で統計量 S の平均を求めると，
文献 [1] の p.123 より

$$E_0(S) = \frac{n+1}{2} \sum_{i=1}^{k} c_i n_i$$

を得る．

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

$$\bar{R}_{i\cdot} \equiv \frac{1}{n_i} \sum_{j=1}^{n} R_{ij}$$

とすると，
方程式

$$S - E_0(S) = \sum_{i=1}^{k} d_i \hat{S}_i$$

が成り立つような d_i を求める．
上式は，

$$\sum_{i=1}^{k} c_i \sum_{j=1}^{n_i} R_{ij} - \frac{n+1}{2} \sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} d_i \sqrt{\frac{12}{n+1}} \left(\bar{R}_{i\cdot} - \frac{n+1}{2} \right) \tag{1}$$

と同等である．(1) の両辺を比較すると，

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

となるので，

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

となる．また，
(条件 1)　　$0 < \lim_{n\to\infty} \dfrac{n_i}{n} \equiv \lambda_i < 1$
を仮定する．
このとき，

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

が成り立つ．
確率変数 Y_i，Z_iと確率ベクトル$\hat{\mathbf{S}}$ を

$$Y_i \sim N\left(0, \frac{1}{\lambda_i}\right),$$

$$Z_i \equiv \sqrt{\lambda_i} Y_i \ ,$$

$$\hat{\mathbf{S}} \equiv {}^t(\hat{S}_1, \cdots, \hat{S}_k)$$

とおくと，文献 [2] の p.85 より

$$\hat{\mathbf{S}} \xrightarrow{\mathcal{L}} {}^t\left(Y_1 - \sum_{j=1}^{k} \lambda_j Y_j, \cdots, Y_k - \sum_{j=1}^{k} \lambda_j Y_j \right)$$

である．

この結果を使って，

$$\begin{aligned}\frac{S-E_0(S)}{n\sqrt{n}} &= \frac{1}{n\sqrt{n}}\sum_{i=1}^{k} d_i\hat{S}_i \\ &\xrightarrow{\mathcal{L}} \frac{1}{\sqrt{12}}\sum_{i=1}^{k}\lambda_i c_i\left(Y_i-\sum_{j=1}^{k}\lambda_j\frac{Z_j}{\sqrt{\lambda_j}}\right) \\ &= \frac{1}{\sqrt{12}}\sum_{i=1}^{k}\sqrt{\lambda_i}c_i Z_i-\sum_{i=1}^{k}\lambda_i c_i\sum_{j=1}^{k}\sqrt{\lambda_j}Z_j \qquad (2)\end{aligned}$$

である．

3 k 標本ポアソンモデルと基礎漸近理論

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 群と呼ばれる．表 2 の X_{ij} は平均 μ_i のポアソン分布 $\mathcal{P}_o(\mu_i)$ 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，

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

である．

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

を考える．

表 2 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	$\mathcal{P}_o(\mu_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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．

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

とする．このとき，μ_i の点推定量は，

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

で与えられる．$n\equiv n_1+\cdots+n_k$ とおく．

命題 1 (条件 1) を仮定する．
$\hat{\sigma}_i\equiv\frac{1}{2}\left(\sqrt{\frac{W_i+1}{n_i}}+\sqrt{\frac{W_i}{n_i}}\right)$ とすると

$$2\sqrt{n_i}\,(\widehat{\sigma}_i-\sigma_i)\xrightarrow{\mathcal{L}}Z_i\sim N(0,1)\qquad(i=1,2,3)$$

が成り立つ．

証明 $\hat{\sigma}_i\equiv\frac{1}{2}\left(\sqrt{\frac{W_i+1}{n_i}}+\sqrt{\frac{W_i}{n_i}}\right)$ のとき

$$\begin{aligned}&\sqrt{n_i}\left(\sqrt{\frac{W_i+1}{n_i}}-\sigma_i\right)\\ &\geqq\sqrt{n_i}\frac{1}{2}\left(\sqrt{\frac{W_i+1}{n_i}}+\sqrt{\frac{W_i}{n_i}}\right)-\sigma_i\\ &\geqq\sqrt{n_i}\left(\sqrt{\frac{W_i}{n_i}}-\sigma_i\right)\qquad(3)\end{aligned}$$

の関係が成り立つ．(3) より

$$\begin{aligned}&P\left(2\sqrt{n_i}\left(\sqrt{\frac{W_i+1}{n_i}}-\sigma_i\right)\leqq x\right)\\ &\leqq P\left(2\sqrt{n_i}\left\{\frac{1}{2}\left(\sqrt{\frac{W_i+1}{n_i}}+\sqrt{\frac{W_i}{n_i}}\right)-\sigma_i\right\}\leqq x\right)\\ &\leqq P\left(2\sqrt{n_i}\left(\sqrt{\frac{W_i}{n_i}}-\sigma_i\right)\leqq x\right)\qquad(4)\end{aligned}$$

また，スラツキーの定理より，

$$2\sqrt{n_i}\left(\sqrt{\frac{W_i}{n_i}}-\sigma_i\right)\xrightarrow{\mathcal{L}}Z_i\qquad(5)$$

$$2\sqrt{n_i}\left(\sqrt{\frac{W_i+1}{n_i}}-\sigma_i\right)\xrightarrow{\mathcal{L}}Z_i\qquad(6)$$

を示すことができる．
(4)，(5)，(6) より

$$2\sqrt{n_i}\left\{\frac{1}{2}\left(\sqrt{\frac{W_i+1}{n_i}}+\sqrt{\frac{W_i}{n_i}}\right)-\sigma_i\right\}\xrightarrow{\mathcal{L}}Z_i$$

が成立する． □

4 提案する検定法

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

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

が成り立っている．

$$T_{\boldsymbol{c}}\equiv2\sum_{i=1}^{k}(c_i-\bar{c})n_i\hat{\sigma}_i$$

とおく．ただし，$\bar{c}=\dfrac{1}{n}\sum_{i=1}^{k}n_i c_i$ である．
このとき，$T_{\boldsymbol{c}}=2\sum_{i=1}^{k}(c_i-\bar{c})n_i(\hat{\sigma}_i-\sigma_0)$ と表現できる．
また，

$$\bar{c}_0\equiv\lim_{n\to\infty}\bar{c}=\sum_{j=1}^{k}\lambda_j c_j$$

とおくと，文献 [1] の系 3.6 より，

$$\frac{T_{\boldsymbol{c}}}{\sqrt{n}} = \frac{\sum_{i=1}^{k}(c_i-\bar{c})\sqrt{n_i}\cdot 2\sqrt{n_i}(\hat{\sigma}_i-\sigma_0)}{\sqrt{n}}$$
$$\xrightarrow{\mathcal{L}} \sum_{i=1}^{k}(c_i-\bar{c}_0)\sqrt{\lambda_i}Z_i \tag{7}$$
$$\sim N\left(0, \sum_{i=1}^{k}(c_i-\bar{c}_0)^2\sqrt{\lambda_i}^2\right)$$
$$= N\left(0, \sum_{i=1}^{k}\lambda_i(c_i-\bar{c}_0)^2\right) \tag{8}$$

ここで，(7) を変形する．

$$\sum_{i=1}^{k}(c_i-\bar{c}_0)\sqrt{\lambda_i}Z_i$$
$$= \sum_{i=1}^{k}\sqrt{\lambda_i}c_iZ_i - \sum_{j=1}^{k}\lambda_j c_j \sum_{i=1}^{k}\sqrt{\lambda_i}Z_i \tag{9}$$

(2)，(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．
また，

$$\widehat{T}_{\boldsymbol{c}} = \frac{T_{\boldsymbol{c}}}{\sqrt{\sum_{i=1}^{k} n_i(c_i-\bar{c})^2}}$$

とおくと，(8) より，

$$\widehat{T}_{\boldsymbol{c}} \xrightarrow{\mathcal{L}} N(0,1)$$

が成り立つ．
ここで，$c_i = i \ \ (i = 1, \cdots, k)$ とした $\widehat{T}_{\boldsymbol{c}}$ を $\widehat{T}_k$ とすると，$\widehat{T}_k$ を帰無仮説 H_0vs 対立仮説 H_A に対する検定統計量とすることができる．
すなわち，

$$\widehat{T}_k = \frac{2\sum_{i=1}^{k}\left(i - \frac{1}{n}\sum_{j=1}^{k} jn_j\right)n_i\hat{\sigma}_i}{\sqrt{\sum_{i=1}^{k} n_i\left(i - \frac{1}{n}\sum_{j=1}^{k} jn_j\right)^2}}$$

とする．
ここで検定統計量 $\widehat{T}_k$ について，検定方式を考える．
帰無仮説 H_0 vs 対立仮説 H_A に対する水準 α の検定は，

$$\widehat{T}_k \geqq z(\alpha) \text{ のとき } H_0\text{を棄却する}$$
$$\widehat{T}_k < z(\alpha) \text{ のとき } H_0\text{を棄却しない}$$

で与えられる．ただし，標準正規分布の上側 $100\alpha\%$ 点 を $\mathrm{z}(\alpha)$ とする．
検定関数 ϕ(x) を使って表すと，

$$\phi(x) = \begin{cases} 1 & (\widehat{T}_k \geqq \mathrm{z}(\alpha)) \\ 0 & (\widehat{T}_k < \mathrm{z}(\alpha)) \end{cases}$$

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，線形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 [4] 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();
    keisan6();
    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(Th>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，$\hat{\sigma}$ の値を計算．
3. keisan2 関数の中で，$\bar{c}$ の値を計算．
4. keisan3 関数の中で，$\widehat{T}_k$ の分子の値を計算．
5. keisan4 関数の中で，$\widehat{T}_k$ の分母の値を計算．(段階 1)
6. keisan5 関数の中で，$\widehat{T}_k$ の分母の値を計算．(段階 2)
7. keisan6 関数の中で，$\widehat{T}_k$ の値を計算．
8. 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，線形型検定の結果を表示する．

6　東北大震災のデータとその解析結果

6.1　東北大震災以前の震度 1 以上のデータ

文献 [1] より，地震の回数はポアソン分布に従う．そこで東北大震災のデータをもとに，検定を行った．マグニチュード 9.0 の東北大震災が 2011 年 3 月 11 日 14 時 46

分に発生した．その直前と，起きる 3 ヶ月前の東北地方で発生した震度 1 以上の回数を表 3 に示す．尚，データは文献 [5] より引用している．

表 3　東北大震災以前の震度 1 以上のデータ

日付	震度 1	2	3 以上	合計
2010/12/1〜2010/12/31	6	6	2	14
2011/1/1〜2011/1/31	9	6	5	20
2011/2/1〜2011/2/28	14	13	3	30
2011/3/1〜2011/3/11	31	13	7	51

6.2　実行結果の例

μ_1 を 2010 年 12 月の一日あたりの震度 1 以上の地震の平均回数，μ_2 を 2011 年 1 月の一日あたりの震度 1 以上の地震の平均回数，μ_3 を 2011 年 2 月の一日あたりの震度 1 以上の地震の平均回数，μ_4 を 2011 年 3 月 1 日から 3 月 11 日の一日あたりの震度 1 以上の地震の平均回数とする．東北大震災のデータについて，α=0.01 で検定した場合以下のようになる．

```
Th=7.405608
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 )
多標本ポアソン分布における順序制約がある場合の線形型検定では H_0 を棄却する.
```

6.3　解析結果とその考察

解析を行った結果，東北大震災のデータについて有意水準 $\alpha = 0.01$ について H_0 を棄却した．
この結果より，東北大震災が起こる前は増加傾向にあることが示された．

7　多重比較法

7.1　同時信頼区間

文献 [3] の p.73 より，$i = 1,\ 2,\ 3,\ 4$ に対して，
第 i 群の第 j 日目におきた震度 1 以上の地震回数を X_{ij} とする．このとき X_{ij}はポアソン分布に従い，

$$P(X_{ij} = x) = \frac{(\mu_i)^x}{x!}e^{-\mu},\ E(X_{ij}) = \mu_i$$

である．

$$G_i \equiv \left\{ \frac{\chi^2_{2w_i}(\{1+(1-\alpha)^{\frac{1}{4}}\}/2)}{2n_i} < \mu_i < \frac{\chi^2_{2(w_i+1)}(\{1-(1-\alpha)^{\frac{1}{4}}\}/2)}{2n_i} \right\} \quad (i = 1,\ 2,\ 3,\ 4)$$

とする．このとき，(条件 2)

$$e^{-n_i\hat{\mu}_i} \leqq 1-(1-\alpha)^{\frac{1}{4}} \qquad (i = 1,\ 2,\ 3,\ 4)$$

の下で G_1，G_2，G_3，G_4 は

$$P(\mu_1 \in G_1, \mu_2 \in G_2, \mu_3 \in G_3, \mu_4 \in G_4) \geqq 1-\alpha$$

を満たし，G_1，G_2，G_3，G_4，μ_1，μ_2，μ_3，μ_4 に関する信頼区間 $1-\alpha$ の信頼区間である．この 4 つの区間が交わらなければ μ_1，μ_2，μ_3，μ_4 が異なると判定する．χ^2_n は自由度 n のカイ二乗分布を表す．

7.2　東北大震災のデータに関する解析結果

信頼区間について $\alpha = 0.05$ として考える．
$n_1 = 31, n_2 = 31, n_3 = 28, n_4 = 11,$
$w_1 = 14, w_2 = 20, w_3 = 30, w_4 = 51$ を当てはめる．

$$\max\left\{e^{-n_1\hat{\mu}_1},\ e^{-n_2\hat{\mu}_2},\ e^{-n_3\hat{\mu}_3},\ e^{-n_4\hat{\mu}_4}\right\} = 9.12\times10^{-4} < 1.274\times10^{-2}$$

となり，信頼区間を与える（条件 2）が満たされる．

$i = 1$ の場合は $0.207 < \mu_1 < 0.850$
$i = 2$ の場合は $0.342 < \mu_2 < 1.101$
$i = 3$ の場合は $0.646 < \mu_3 < 1.664$
$i = 4$ の場合は $3.177 < \mu_4 < 6.517$

となる．このとき，μ_1とμ_4 の間 ，μ_2とμ_4 の間 ，μ_3とμ_4 に信頼区間の交わりはなく, 大震災の直前は震度 1 以上の地震が異常な回数起こっていると考察できる．

8　おわりに

本論では，ポアソンモデルにおける線形型検定を提案した．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ポアソンモデルにおける線形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] 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 :「多項の 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ーの多重比較法」

[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] 国土交通省 : 気象庁,
http://www.data.jma.go.jp/svd/eqdb/data/shindo/index.php

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