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

2012SE188 野澤　慎

指導教員：白石高章

1　はじめに

統計学の基礎を学び，それをもとに，ウィルコクソンの順位検定や順位に基づく区間推定などのノンパラメトリック法について学んだ．そして，ノンパラメトリック法における平均に順序制約がある場合の検定法として知られる 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，その検定法について紹介し，さらに多標本正規分布モデルにおける Jonckheere-Terpstra 型検定法について考察する．

2　Jonckheere-Terpstra 検定の紹介

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 $(X_{i1}, X_{i2}, \cdots, X_{in_i})$ を第 i 標本とし，$P(X_{ij} \leqq x) = F(x-\mu_i)$, $E(X_{ij}) = \mu_i$ とする．ただし，$F(x-\mu_i)$ は連続型の分布関数とし，$\mu_1, \cdots, \mu_k$ はすべて未知パラメータとする．また，総標本サイズを $n \equiv n_1 + \cdots + n_k$ とおき，すべての X_{ij} は互いに独立であると仮定する．なお，本研究に用いるモデルは文献 [2] を参考にした．帰無仮説 $H_0 : \mu_1 = \cdots = \mu_k$ vs. 対立仮説 $H_A : \mu_1 \leqq \cdots \leqq \mu_k$(少なくとも 1 つの $\leqq$ は $\lneqq$ である) の Jonckheere-Terpstra 検定統計量 (文献 [3]) は，

$$JT \equiv \sum_{i<i'} \sum W_{ii'} = \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} W_{ii'}$$

で与えられる．ただし，

$$W_{ii'} = \#\left\{(j, j') \mid X_{i'j'} > X_{ij},\ 1 \leqq j \leqq n_i,\ 1 \leqq j' \leqq n_{i'}\right\}$$

である．また，

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

と仮定する．

帰無仮説 H_0 の下での，JT の平均と分散を求めると，

$$E_0(JT) = \frac{1}{4}\left(n^2 - \sum_{i'=1}^{k} n_{i'}^2\right),$$

$$V_0(JT) = \frac{1}{12} \sum_{i'=2}^{k} \left\{N_{i'-1} n_{i'} (N_{i'-1} + n_{i'} + 1)\right\}$$

となる．ただし，$E_0(\cdot), V_0(\cdot)$ は，それぞれ帰無仮説 H_0 の下での平均と分散を表す．また，$N_i \equiv \sum_{j=1}^{i} n_j$ である．中心極限定理と同様に，文献 [3] より，

$$S_{JT} \equiv \frac{JT - E_0(JT)}{\sqrt{V_0(JT)}} \xrightarrow{\mathcal{L}} N(0,1) \qquad (1)$$

が成り立つ．ただし，$D_n \xrightarrow{\mathcal{L}} D$ は，D_n が D に分布収束することを表す．

よって，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，水準 α の検定は，$S_{JT} \geqq z(\alpha)$ のとき帰無仮説 H_0 を棄却することである．

3　Jonckheere-Terpstra 型検定法

2 節のモデルで，$F(x-\mu_i) = \Phi\left((x-\mu_i)/\sigma\right)$ とする．ただし，$\Phi\left((x-\mu_i)/\sigma\right)$ は $N(\mu_i, \sigma^2)$ の分布関数であり，$\mu_1, \cdots, \mu_k, \sigma^2$ はすべて未知パラメータとする．帰無仮説 $H_0 : \mu_1 = \cdots = \mu_k$ vs. 対立仮説 $H_A : \mu_1 \leqq \cdots \leqq \mu_k$(少なくとも 1 つの $\leqq$ は $\lneqq$ である) として，帰無仮説 H_0 の下で $\mu_i = \mu_0$ $(i = 1, \cdots, k)$ とする．

以下，帰無仮説 H_0 の下で考える．

$$\widehat{T}_{ii'} = W_{ii'} - \frac{n_i n_{i'}}{2}$$

より，

$$JT - E_0(JT) = \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} \widehat{T}_{ii'}$$

となる．さらに，

$$\mathcal{V}_p(i,i') \equiv n_i \sum_{j'=1}^{n_{i'}} \left\{F\left(X_{i'j'} - \mu_0\right) - \frac{1}{2}\right\} - n_{i'} \sum_{j=1}^{n_i} \left\{F\left(X_{ij} - \mu_0\right) - \frac{1}{2}\right\} \qquad (2)$$

とおくと，文献 [2] の定理 3.3 より，

$$\frac{\sqrt{n_{i'} + n_i}\widehat{T}_{ii'}}{n_{i'} n_i} - \frac{\sqrt{n_{i'} + n_i}\mathcal{V}_p(i,i')}{n_{i'} n_i} \xrightarrow{P} 0$$

が示される．ただし，$D_n \xrightarrow{P} d$ は，D_n が d に確率収束することを表す．したがって，$JT - E_0(JT)$ は，(2) の $F(X_{ij} - \mu_0) - 1/2 \sim U(-1/2, 1/2)$ を $X_{ij} \sim N(\mu_0, \sigma^2)$ に置き換えることにより，

$$T_J \equiv \sum_{i'=2}^{k} \sum_{i=1}^{i'-1} n_i n_{i'} \left(\bar{X}_{i'\cdot} - \bar{X}_{i\cdot}\right)$$

と表せる．ただし，$U(-1/2, 1/2)$ は区間 $(-1/2, 1/2)$ 上の一様分布を表し，$X \sim D$ は，X が D に従うことを表す．また，$\bar{X}_{i\cdot} \equiv \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ である．ここで，T_J の平均と分散を求めると，

$$E_0(T_J) = 0,$$

$$V_0(T_J) = \sum_{i'=2}^{k} n_{i'} \left(\sum_{i=1}^{i'-1} n_i\right)\left(\sum_{i=1}^{i'} n_i\right)\sigma^2 \qquad (3)$$

となる．さらに，

$$E_0\left(\frac{T_J - E_0(T_J)}{\sqrt{V_0(T_J)}}\right) = 0,\ V_0\left(\frac{T_J - E_0(T_J)}{\sqrt{V_0(T_J)}}\right) = 1$$

より，

$$Y \equiv \frac{T_J}{\sqrt{V_0(T_J)}} \sim N(0,1)$$

が成り立つ．また，

$$Z \equiv \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})^2 \sim \chi^2_{n-k}$$

である．ただし，χ^2_n は，自由度 n のカイ二乗分布を表す．Y, Z は互いに独立なので，文献 [1] の定理 3.21 より，

$$\frac{Y}{\sqrt{\frac{Z}{n-k}}} \sim t_{n-k}$$

が成り立つ．ただし，t_n は，自由度 n の t 分布を表す．また，

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

とする．ただし，$\hat{\sigma}^2$ は σ^2 の一様最小分散不偏推定量である．ここで，$V_0(T_J)$ は (3) より，未知パラメータ σ^2 を含むのでそれを消去し，代わりに $\hat{\sigma}^2$ を用いることにする．したがって，(1) の $JT - E_0(JT)$ を T_J に，$V_0(JT)$ を $V_0(T_J)$ に置き換えると，

$$\frac{T_J \cdot \sigma}{\sqrt{V_0(T_J)\hat{\sigma}^2}} = \frac{Y}{\sqrt{\frac{Z}{n-k}}}$$

が成り立つ．これを検定統計量とすると，

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

である．よって，自由度 $n-k$ の t 分布の上側 $100\alpha\%$ 点を $t(n-k;\alpha)$ とすると，水準 α の検定は，$S_N > t(n-k;\alpha)$ のとき帰無仮説 H_0 を棄却することである．

4 C 言語によるプログラム解説

Jonckheere-Terpstra 検定 (2 節)，Jonckheere-Terpstra 型検定法 (3 節) による検定結果を出力するプログラムを C 言語によりそれぞれ作成した．ただし，上側確率を求めるために，文献 [4] を参考にした．ページの都合上，Jonckheere-Terpstra 検定の main プログラムのみを以下に記載する．なお，Jonckheere-Terpstra 検定，Jonckheere-Terpstra 型検定法の詳細なプログラムについては，それぞれ本稿に記載した．

```
int main(void){
    yomikomi(); keisan1();  keisan2();
    keisan3();  keisan4(); output();
    return (0);
}
```

1. yomikomi 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. keisan1 関数により，JT の値を計算する．
3. keisan2 関数により，$E_0(JT)$ の値を計算する．
4. keisan3 関数により，$V_0(JT)$ の値を計算する．
5. keisan4 関数により，S_{JT} の値を計算する．
6. output 関数により，検定結果を出力する．

4.1 日本の年平均気温データ

日本の年平均気温が上昇しているかどうかを調べるにあたり，1981 年から 2015 年のデータ (文献 [5]) を使用した．このとき，μ_i を $(1980+i)$ 年の日本の年平均気温 $(i = 1, \cdots, 35)$ とする．北海道を除く 46 都府県については，各都府県庁所在地における気温のデータを集めた．北海道は，面積が大きいため 15 ヶ所の気温データを集めた．そして，全 61 地点を北海道，東北地方，関東地方，中部地方，近畿地方，中国地方，四国地方，九州地方の 8 つの地域に分けた．各地域の年平均気温は，そこに属する地点の平均を計算し，用いた．

4.2 解析結果

Jonckheere-Terpstra 検定では，$S_{JT} = 3.19, p$ 値 $= 0.000703$ となり，$\alpha = 0.05, 0.01$ のどちらの場合も帰無仮説 H_0 は棄却された．Jonckheere-Terpstra 型検定法では，$S_N = 1.85, p$ 値 $= 0.032311$ となり，$\alpha = 0.05$ では帰無仮説 H_0 は棄却され，$\alpha = 0.01$ では棄却されなかった．

4.3 考察

2 種類の検定により，1981 年から 2015 年の 35 年間で日本の年平均気温が上昇していることが確認できた．検定統計量 S_{JT} は S_N とは違い，データの順序と各標本サイズにのみ依存している．それに対して S_N は，データの値を使用している．このため，$\alpha = 0.01$ のときの結果に違いが出たと推察できる．

5 おわりに

本論では，Jonckheere-Terpstra 検定の検定方式を紹介し，さらに多標本正規分布モデルにおける検定方式を提案した．そして，それぞれの検定を行うための C 言語プログラムを作成し，実際のデータを用いることによってより理解を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』. 共立出版株式会社，東京，2011.

[3] Thomas P. Hettmansperger : *Statistical Inference Based on Ranks (Wiley Series in Probability and Statistics)*. Wiley, New York, 1984.

[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（2013 年 1 月）.

[5] 国土交通省:気象庁,http://www.jma.go.jp/jma/index.html.