C．研究内容・成果

（1）「統計的モデルと統計量」に関する研究報告

Bonferroniの不等式を用いた多変量多重比較についての一考察

東海大学理学部　　道家暎幸

1. はじめに

多変量観測値に基づいた多重比較法には、Hochberg and Tamhane(1987) のUnion-Intersection 法があるが、その方法は、変量の数や母集団の数が増えるにつれて最大固有値の分布を計算することが容易でないという非実用的な問題が存在する。ここでの多変量多重比較法は、Srivastava and Worsley(1986) の最大ベータ統計量の分布関数を用い、改良されたBonferroniの不等式を用いて二項目までで打ち切り、二項目のベータ統計量の同時分布を線形変換とテイラー展開を行う。この分布関数を用い同時棄却限界を求め検定を行う。

2 検定統計量

k 個の p 変量母集団があり、i 番目の母集団の反応ベクトル $\boldsymbol{x}_i(i=1,2,\cdots,k)$ は p-変量正規分布 $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$ に従う。分散共分散行列 $\boldsymbol{\Sigma}$ は未知とする。ここで $i,j(i<j,\ i=1,2,\cdots,k-1,\ j=2,3,\cdots,k)$ に対して

$$\text{帰無仮説 } H_{\{i,j\}}:\boldsymbol{\mu}_i=\boldsymbol{\mu}_j,\ \text{ 対立仮説 } K_{\{i,j\}}:\boldsymbol{\mu}_i\neq\boldsymbol{\mu}_j \tag{2.1}$$

を与える。全ての i,j の組み合わせに対する ${}_kC_2(=m)$ 個の同時仮説

$$\{H_{\{1,2\}},H_{\{1,3\}},\cdots,H_{\{k-1,k\}}\} \tag{2.2}$$

を考える。今、i 番目の母集団からサンプルサイズ $N_i(i=1,2,\cdots,k)$ の無作為標本ベクトルを $\boldsymbol{x}_1^{(i)},\boldsymbol{x}_2^{(i)},\cdots,\boldsymbol{x}_{N_i}^{(i)}$ とする。i 番目の母集団の標本平均ベクトルを $\bar{\boldsymbol{x}}_i$、標本分散共分散行列を $\boldsymbol{S}_i$ で表す。$\boldsymbol{S}_i$ をもとにプールした標本分散共分散行列を

$$\boldsymbol{S}_{i,j}=\frac{n_i\boldsymbol{S}_i+n_j\boldsymbol{S}_j}{n_i+n_j},\ \ i<j,\ i=1,2,\cdots,k-1,\ j=2,3,\cdots,k$$

とする。帰無仮説 $H_{\{i,j\}}:\boldsymbol{\mu}_i=\boldsymbol{\mu}_j$ の下では、$\boldsymbol{y}_{i,j}=\sqrt{\dfrac{N_iN_j}{N_i+N_j}}(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}}_j)\sim N_p(\boldsymbol{0},\boldsymbol{\Sigma})$ また統計量 $T_{i,j}^2=\boldsymbol{y}_{i,j}'\boldsymbol{S}_{i,j}^{-1}\boldsymbol{y}_{i,j}$ は、自由度 (p,ν) の Hotelling T^2 分布に従う。ここで、$\nu=N_i+N_j-p-1$ である。Srivastava and Worsley(1986) によって提案された次の

$$S_{i,j}=\frac{T_{i,j}^2}{N_i+N_j-2+T_{i,j}^2}=\boldsymbol{y}_{i,j}'\boldsymbol{V}^{-1}\boldsymbol{y}_{i,j} \tag{2.3}$$

を用いる。ここで $\boldsymbol{V}=\sum_{i=1}^{k}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'$ である。この時 $H_{\{i,j\}}$ の下で $S_{i,j}$ はパラメータ $(\frac{1}{2}p,\frac{1}{2}\nu)$ のベータ分布に従う。改良型Bonferroniの不等式は、次に示さ

れる。

$$\mathrm{P}(\max_{i<j} S_{i,j} > t^*) \leq \sum_{i<j}^{m} \mathrm{P}(S_{i,j} > t^*) - \sum_{i<j,\ i^*<j^*}^{m-1} \mathrm{P}\{(S_{i,j} > t^*) \cap (S_{i^*,j^*} > t^*)\}. \qquad (2.4)$$

3 $S_{i,j}$ と $S^*_{i,j}$ の同時分布

いま $\boldsymbol{X}(N \times p) = (\boldsymbol{x}_1^{(1)}, \cdots, \boldsymbol{x}_{N_1}^{(1)}, \boldsymbol{x}_1^{(2)}, \cdots, \boldsymbol{x}_{N_2}^{(2)}, \cdots, \boldsymbol{x}_1^{(k)}, \cdots, \boldsymbol{x}_{N_k}^{(k)})'$,

$$\boldsymbol{a}_{i,j}(N \times 1) = \sqrt{\frac{N_i N_j}{N_i + N_j}} \left(0, \cdots, 0, \frac{1}{N_i}, \cdots, \frac{1}{N_i}, 0, \cdots, 0, -\frac{1}{N_j}, \cdots, -\frac{1}{N_j}, 0, \cdots, 0\right)'$$

とおと $\boldsymbol{X}'\boldsymbol{a}_{i,j} = \boldsymbol{y}_{i,j}$ となる。また $\boldsymbol{A}(2 \times N) = (\boldsymbol{a}_{i,j}, \boldsymbol{a}_{i^*,j^*})'$ とおくと $\boldsymbol{y}'_{i,j} = \boldsymbol{a}'_{i,j}\boldsymbol{X}$, $\boldsymbol{y}'_{i^*,j^*} = \boldsymbol{a}'_{i^*,j^*}\boldsymbol{X}$ であるから

$$\boldsymbol{A}\boldsymbol{X} = \begin{pmatrix} \boldsymbol{a}'_{i,j} \\ \boldsymbol{a}'_{i^*,j^*} \end{pmatrix} \boldsymbol{X} = \begin{pmatrix} \boldsymbol{y}'_{i,j} \\ \boldsymbol{y}'_{i^*,j^*} \end{pmatrix} = \boldsymbol{Y}(2 \times p) \qquad (3.1)$$

と表せる。ここで、$\boldsymbol{S}(2 \times 2) = \boldsymbol{Y}\boldsymbol{V}^{-1}\boldsymbol{Y}'$ とおくと

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{y}'_{i,j}\boldsymbol{V}^{-1}\boldsymbol{y}_{i,j} & \boldsymbol{y}'_{i,j}\boldsymbol{V}^{-1}\boldsymbol{y}_{i^*,j^*} \\ \boldsymbol{y}'_{i^*,j^*}\boldsymbol{V}^{-1}\boldsymbol{y}_{i,j} & \boldsymbol{y}'_{i^*,j^*}\boldsymbol{V}^{-1}\boldsymbol{y}_{i^*,j^*} \end{pmatrix} = \begin{pmatrix} S_{i,j} & s \\ s & S_{i^*,j^*} \end{pmatrix} \qquad (3.2)$$

と表せる。Srivastava and Khatri(1979) より $\boldsymbol{S}$ の同時確率密度関数は

$$f(\boldsymbol{S}) = c_0(ab - \rho^2_{ij,i^*j^*})^{-\frac{p+\nu-3}{2}} (S_{i,j}S_{i^*,j^*} - s^2)_+^{\frac{p-3}{2}} \{(a - S_{i,j})(b - S_{i^*,j^*}) - (\rho_{ij,i^*j^*} - s)^2\}_+^{\frac{\nu-3}{2}} \qquad (3.3)$$

と表せる。ここで c_0 はガンマ関数を含む定数である。ここで $S_{i,j}$ と $S^*_{i,j}$ の周辺密度を求めた後、$S_{i,j}$ と $S^*_{i,j}$ の同時確率 を $\rho_{ij,i^*j^*} = 1$ についてテイラー展開をすると

$$\mathrm{P}\{(S_{i,j} > t^*) \cap (S_{i^*,j^*} > t^*)\} \approx 1 - G_{p,\nu}(t^*) - q_1 t_{ij,i^*j^*} + q_2 t^3_{ij,i^*j^*} \qquad (3.4)$$

となる。ここで、$t_{ij,i^*j^*} = \sqrt{1 - \rho_{ij,i^*j^*}}$ であり、q_1, q_2 は定数である。この (3.4) 式を (2.4) 式に代入し、改良型 Bonferroni の不等式を用い同時棄却限界を求め、同時検定を行う。

参考文献

[1] Hochberg, Y. and Tamhane, A. C.(1987) : *Multiple Comparison Procedures*, John Wiley & Sons.

[2] Srivastava, M. S. and Khatri, C, G.(1979) : *An Introduction to Multivariate Statistics*, New York : North-Holland.

[3] Srivastava, M. S. and and Worsley, K. J.(1986) : "Likelihood Ratio Tests for a Change in the Multivariate Normal Mean",
Journal of the American Statistical Association 81, pp199-204.

U-統計量の線形結合の分布の収束についての非 Berry-Esseen 型上限

鹿児島大理工研 D3　戸田 光一郎
鹿児島大理学部　大和 元

1. 序

$X_1,...,X_n$ を分布 F からの大きさ n の標本とする. 次数 k の対称な kernel $g(x_1,...,x_k)$ をもつ estimable parameter $\theta(F)$ の推定量として, U-統計量 U_n と V-統計量 V_n ;

$$U_n = \binom{n}{k}^{-1} \sum_{1\le j_1<\cdots<j_k\le n} g(X_{j_1},...,X_{j_k}),$$

$$V_n = \frac{1}{n^k}\sum_{j_1=1}^{n}\cdots\sum_{j_k=1}^{n} g(X_{j_1},...,X_{j_k})$$

が知られている (例えば, Lee [2]). また, Yamato [4] により, $\theta(F)$ の推定量として LB-統計量 B_n ;

$$B_n = \binom{n+k-1}{k}^{-1} \sum_{r_1+\cdots+r_n=k} g(\underbrace{X_1,...,X_1}_{r_1},...,\underbrace{X_n,...,X_n}_{r_n})$$

が得られた. 但し, $\sum_{r_1+\cdots+r_n=k}$ は $r_1+\cdots+r_n=k$ を満たすすべての非負整数 $r_1,...,r_n$ 上でとられる和を表す.

Toda and Yamato [3] は, V-統計量, LB-統計量を含む U-統計量の線形結合 Y_n を提案した. Y_n は以下のようにして与えられる.

$j=1,...,k$ に対して, $w(r_1,...,r_j;k)$ を $r_1+\cdots+r_j=k$ を満たす正整数 $r_1,...,r_j$ の対称な非負関数とする. 但し, k は kernel g の次数である. $j=1,...,k$ に対して, $w(r_1,...,r_j;k)$ の少なくとも 1 つは正であるとし,

$$d(k,j) = \sum\nolimits^{+}_{r_1+\cdots+r_j=k} w(r_1,...,r_j;k)$$

とおく. 但し, $\sum^{+}_{r_1+\cdots+r_j=k}$ は $r_1+\cdots+r_j=k$ を満たすすべての正整数 $r_1,...,r_j$ 上でとられる和を表す. $j=1,...,k$ に対して, $g_{(j)}(x_1,...,x_j)$ を

$$g_{(j)}(x_1,...,x_j) = \frac{1}{d(k,j)}\sum\nolimits^{+}_{r_1+\cdots+r_j=k} w(r_1,...,r_j;k)\, g(\underbrace{x_1,...,x_1}_{r_1},...,\underbrace{x_j,...,x_j}_{r_j})$$

により与えられる kernel とし, $g_{(j)}(x_1,...,x_j)$ に対応する U-統計量を $U_n^{(j)}$ とする. このとき, $\theta(F)$ の推定量として

$$Y_n = \frac{1}{D(n,k)}\sum_{j=1}^{k} d(k,j)\binom{n}{j} U_n^{(j)}$$

を提案した. 但し, $D(n,k)=\sum_{j=1}^{k} d(k,j)\binom{n}{j}$ である. 特に, $r_1+\cdots+r_j=k\,(j=1,...,k)$ となる正整数 $r_1,...,r_j$ に対して, $w(r_1,...,r_j;k)=k!/(r_1!\cdots r_j!)$ のとき Y_n は V-統計量 V_n となり, $w(r_1,...,r_j;k)=1$ のとき Y_n は LB-統計量 B_n となる.

以下では, C を分布 F に依存しない正定数とし, $\Phi(x)$ を標準正規分布の分布関数とおく. また, $\psi_1(x_1)=E\{g(X_1,...,X_k)\mid X_1=x_1\}$, $g^{(1)}(x_1)=\psi_1(x_1)-\theta$, $\sigma_1^2=E\{g^{(1)}(X_1)\}^2$ とおき, kernel g は非退化, 即ち, $\sigma_1^2>0$ とする.

2. U-統計量の線形結合 Y_n に対する非 Berry-Esseen 型上限

U-統計量の線形結合 Y_n に対する非 Berry-Esseen 型の上限について紹介する. $d(k,k)>0$ のとき,

$$\frac{d(k,k)}{D(n,k)}\binom{n}{k} = 1-\frac{\beta}{n}+O(\frac{1}{n^2}),\quad \sum_{j=1}^{k-1}\frac{d(k,j)}{D(n,k)}\binom{n}{j} = \frac{\beta}{n}+O(\frac{1}{n^2})$$

となる定数 $\beta\ (\geq 0)$ が存在する. U-統計量 U_n に対しては $\beta = 0$, V-統計量 V_n に対しては $\beta = k(k-1)/2$, LB-統計量 B_n に対しては $\beta = k(k-1)$ である. Zhao and Chen [5], Koroljuk and Borovskich [1] により, U-統計量に対する様々な形の非 Berry-Esseen 型の上限が得られている. 我々は, U-統計量に対する Berry-Esseen bound と Edgeworth 展開を用いて次の結果を得た.

Proposition 1 $\sigma_1^2 > 0, E \mid g(X_1,...,X_k) \mid^3 < \infty$ とし, $\lim_{|t|\to\infty} \mid Ee^{it\psi_1(X_1)} \mid < 1$ と仮定する. このとき, 任意の $x \in \mathcal{R}$ と十分大きな n に対して

$$\mid \Pr\left[\frac{\sqrt{n}(U_n - \theta)}{k\sigma_1} \leq x\right] - \Phi(x) \mid \leq \frac{C}{\sqrt{n}(1+\mid x \mid)^3}.$$

以下では $\beta > 0$ と仮定する. U-統計量の線形結合 Y_n は, U-統計量 U_n を用いて

$$Y_n = U_n + R_n$$

と表わすことができる. 但し,

$$R_n = [\frac{d(k,k)}{D(n,k)}\binom{n}{k} - 1]U_n + \frac{1}{D(n,k)}\sum_{j=1}^{k-1} d(k,j)\binom{n}{j}U_n^{(j)}$$

である. U-統計量に対する非 Berry-Esseen 型上限 ([1], [5], Prop.1) を用いて, 以下の結果を得た.

Theorem 2 ある $\delta\ (0 \leq \delta \leq 1)$ と任意の整数 $j_1,...,j_k\ (1 \leq j_1 \leq \cdots \leq j_k \leq k)$ に対して, kernel g は $\sigma_1^2 > 0, E \mid g^{(1)}(X_1) \mid^{2+\delta} < \infty, E \mid g(X_1,...,X_k) \mid^{(4+\delta)/3} < \infty, E \mid g(X_{j_1},...,X_{j_k}) \mid^{(8+\delta)/6} < \infty$ を満たすとする. このとき, $n \to \infty$ に対して

$$\sup_x \mid \Pr\left[\frac{\sqrt{n}(Y_n - EY_n)}{k\sigma_1} \leq x\right] - \Phi(x) \mid = O(n^{-\frac{\delta}{2}}).$$

Theorem 3 $\sigma_1^2 > 0, E \mid g(X_1,...,X_k) \mid^3 < \infty$ とし, 任意の整数 $j_1,...,j_k\ (1 \leq j_1 \leq \cdots \leq j_k \leq k)$ に対して, $E \mid g(X_{j_1},...,X_{j_k}) \mid^2 < \infty$ とする. このとき, 任意の $x \in \mathcal{R}$ に対して

$$\mid \Pr\left[\frac{\sqrt{n}(Y_n - EY_n)}{k\sigma_1} \leq x\right] - \Phi(x) \mid \leq \frac{C}{\sqrt{n}(1+x^2)}.$$

Proposition 4 $\sigma_1^2 > 0$ とし, 任意の整数 $j_1,...,j_k\ (1 \leq j_1 \leq \cdots \leq j_k \leq k)$ に対して, $E \mid g(X_{j_1},...,X_{j_k}) \mid^3 < \infty$ とする. さらに, $\lim_{|t|\to\infty} \mid Ee^{it\psi_1(X_1)} \mid < 1$ と仮定する. このとき, 任意の $x \in \mathcal{R}$ と十分大きな $n\ (\geq 8)$ に対して

$$\mid \Pr\left[\frac{\sqrt{n}(Y_n - EY_n)}{k\sigma_1} \leq x\right] - \Phi(x) \mid \leq \frac{C}{\sqrt{n}(1+\mid x \mid)^3}.$$

Theorem 5 $\sigma_1^2 > 0, E \mid g(X_1,X_2) \mid^3 < \infty, E \mid g(X_1,X_1) \mid^3 < \infty$ とする. このとき, 任意の $x \in \mathcal{R}$ と $n\ (\geq 8)$ に対して

$$\mid \Pr\left[\frac{\sqrt{n}(Y_n - EY_n)}{2\sigma_1} \leq x\right] - \Phi(x) \mid \leq \frac{C}{\sqrt{n}(1+\mid x \mid)^3}.$$

参考文献

[1] Koroljuk, V.S. and Borovskich, Yu.V. (1994), *Theory of U-statistics.* Kluwer, Dordrecht.

[2] Lee, A.J. (1990), *U-statistics.* Marcel Dekker, New York.

[3] Toda, K. and Yamato, H. (2001), *J. Japan Statist. Soc.* **31**, No.2, 225–237.

[4] Yamato, H. (1977), *J. Japan Statist. Soc.* **7**, 57–66.

[5] Zhao, Lincheng. and Chen, Xiru. (1983), *Scientia Sinica.* Ser. A, **26**, No.8, 795–810.

RANDOMIZED PLAY-THE-WINNER RULE FOR ORDERED CATEGORICAL DATA: A FOLLOW UP MODEL.

Gopaldeb Chattopadhyay
Bureau of Applied Economics & Statistics
Government of West Bengal, India

*Aditya Chattopadhyay(Chatterjee)**
Department of Statistics
The University of Burdwan, India

**[Presently visiting the Department of Mathematics, Hiroshima University* and Speaker]

Comparison of a new drug with a standard or placebo drug through clinical trials is a very common problem of interest in the pharmaceutical industries. In many such trials the treatment responses are measured on an ordinal scale rather than on a continuous scale. A number of authors over the past decade have considered techniques for analyzing such type of ordered categorical data. A simple scoring system called *ridits* (relative to an identified distribution) that was first introduced by Bross(1958) may be used towards analyzing such data where cumulative probability scores instead of arbitrarily selected scores are considered. Brockett and Levine (1977) noticed that the ridit scores, estimated from the data, have the property that if we combine two adjacent categories and redefine the scores by the same method, then the scores for the remaining categories remain unchanged. The most common category scores or equal interval scores do not have this property. Ridit analysis has been successfully applied to the study of automobile accident (Bross(1960)), cancer (Wynder, Bross, Hirayama(1960)), schizophrenia (Spitzer et. al. (1965)), preference studies (Pouolard et. al.(1997)).

The technique of data-dependent allocation of treatments to the patients are of paramount interest with regard to clinical trials. For example, if the subjects enter into a system

sequentially, the problem of allocation of treatments among the entering subjects requires thorough scrutiny. Further, as the subjects are human beings, from ethical point of view, it is desirable to carry out a test procedure with smaller number of patients being treated by the inferior treatment in course of the decision making. With this idea in mind Zelen (1969) introduced the concept of play-the-winner rule for dichotomous treatment responses. Later Wei and Durham (1978) and Wei (1979) modified this idea and introduced randomized-play-the-winner rule. One of the major requirements in such sequential trials is that the outcomes are known relatively quickly and the treatment responses are dichotomous. In fact the method provided by Wei (1988), holds only if the treatment responses are dichotomous and instantaneous. Following Wei, Chattopadhyay (2002) proposed a test procedure for more than two treatment response categories. However, that procedure is not suitable when the treatment response of all previously treated patients are not readily available with the clinician before treating a particular patient, i.e. when the treatment responses are not instantaneous. In practice, treatment responses are not always instantaneous and often it is required to follow up the patients after certain time period since administering the drug. In the present article our aim is to provide a suitable test procedure for comparing two treatments (say treatment A and treatment B) when the treatment responses are ordered categorical in nature and each patient is followed up after certain time period (say, D days) from the date of administering the treatment. On an average this rule also allows more patients to be treated by better treatment in course of decision making, preserving the ethical aspect of clinical trail. At the same time the treatment responses are *not* required to be instantaneous. Various small sample and asymptotic empirical results of the test have been derived. Moreover power and ASN studies have been done by simulation to establish the claims.

空間スキャン統計量とechelon を利用した分割表データの解析

栗原考次（岡山大学・環境理工学部）

1．はじめに

分割表で与えられるデータを分析するためには、独立性の検定、対数線形モデルやロジットモデルなどを用いた接近法が行われる。カイ２乗検定などを利用し独立性の検定を行う場合、非独立性は表の全体を通して得られたのか、あるいは分割表内の特定のセルまたはセル群が独立性から逸脱しているという情報が得られない。こうした問題については、各セル毎に調整化された残差を用いた正規検定などが用いられる。しかし、順序カテゴリーを持つ分割表データの場合、隣接したセルの間には順序性があり空間的なつながりがあるのでセル毎に個別に検定を行うのは適当ではなく、隣接した空間的な位置情報も考慮したセル群を見つける検定を行うことが望まれる。本研究では、分割表データの解析に空間構造を取り入れ、有意に連関の源になっているセル群の検出を行う方法論について提唱する。すなわち、echelon 解析により分割表の空間的な階層構造を求め、その構造に基づき空間スキャン統計量を計算することにより、有意に独立性から逸脱しているセル群を見つけだす方法を提唱する。

2．空間スキャン統計量

空間スキャン統計量は、病気の発生率のように地域毎に得られるデータにおいて有意に高いまたは低い値を示す地域（ホットスポット）を見つけるために用いられる（Kulldorff (1997)）。

すべての領域を G とし、その部分集合の領域を Z とする。領域 Z の内部では個人はある属性を確率 p、領域 Z の外では確率 q で持つものとする。属性を持つ確率は互いに独立とする。このとき、仮説は以下の通りである。

帰無仮説　$H_0: p=q$　　*v.s.*　　対立仮説　$H_1: p>q$.

ここで、　$n(G)$をすべての領域Ｇでの母集団の数、$n(Z)$を領域Ｚ内での母集団の数、$c(G)$をすべての領域Ｇで属性を持つもの数、$c(Z)$を領域Ｚ内で属性を持つもの数とする。

このとき、ポアソンモデルに基づく尤度比 λは以下の式で示され、ホットスポットは全領域の部分集合の領域Ｚで最大のものとする。

$$\lambda = \frac{\left(\frac{c(Z)}{n(Z)}\right)^{C(Z)}\left(\frac{c(G)-c(Z)}{n(G)-n(Z)}\right)^{C(G)-C(Z)}}{\left(\frac{c(G)}{n(G)}\right)^{C(G)}} = \left(\frac{c(Z)}{e(G)}\right)^{C(Z)}\left(\frac{c(G)-c(Z)}{e(G)-e(G)}\right)^{C(G)-C(Z)}$$

ここに、$e(Z)$ は領域Z内で属性を持つ数の期待値である。ホットスポットの候補を見つける検定統計量として、対数尤度比統計量 logλ を計算する。Kulldorff はスキャンする円の中心は領域の中心とし、半径は母集団の半分になるまで変化させている。また、帰無仮説上の logλ をモンテカルロ法により計算し、同時にp-value も計算している。

3．ｒ×ｃの順序カテゴリーを持つ分割表におけるホットスポットの検出

２つの変量 ｘ と ｙ に関してｒ個とｃ個の順序カテゴリーによって分類されたｒ×ｃ分割表は、ｒ×ｃの２次元メッシュ上に空間構造をもつデータとみなすことができる。 p_{ij} を i

行 j 列カテゴリーにおける母集団確率($i=1,2,..,r\ j=1,2,...c$)とし 、$p_{i.}$ と $p_{.j}$ をそれぞれ行及び列の周辺確率を表すものとする。その時、帰無仮説として２つの変量が独立であるという仮説を考える。

$$H_0:\ \ p_{ij} = p_{i.}p_{.j} \qquad \text{for i} = 1,2,\ldots,\text{r and j} = 1,2,\ldots\text{c}$$

$n(ij)$ と $c(ij)$ をそれぞれ i 行 j 列カテゴリーに対する母集団の大きさ及び観測度数、$n(i.)$, $c(i.)$ と $n(.j)$, $c(.j)$ をそれぞれ行及び列の母集団及び観測度数に対する周辺度数とする。また、$n(..)$ と $c(..)$ をそれぞれ総母集団の大きさ及び総観測度数とする。帰無仮説の下で、i 行 j 列カテゴリーの母集団確率の最尤推定量及び期待度数は次式で与えられる。

$$\hat{p}_{ij} = \frac{c(i.)c(.j)}{c(..)^2}., \qquad \hat{e}(ij) = c(..)\hat{p}_{ij} = \frac{c(i.)c(.j)}{c(..)}$$

例として、２節で説明したポアソンモデルに基づく対数尤度比統計量を利用し、表１の両親の社会経済状態と精神的健康状態の分割表データに対して有意に独立性から逸脱しているセル群を見つける。ここでは、連関の源泉をはかるデータとして Kulldorff らが開発したソフトウェア SaTScan に従い、分割表のセルデータとして relative risk を用いる。

表１　両親の社会経済状態と精神的健康状態で分類された６×４分割表

		Mental Health Status				
		well	mild symptom formation	moderate symptom formation	impaired	sum
Parent's socio economic status	A(high)	64	94	58	46	262
	B	57	94	54	40	245
	C	57	105	65	60	287
	D	72	141	77	94	384
	E	36	97	54	78	265
	F(low)	21	71	54	71	217
	sum	307	602	362	389	1660

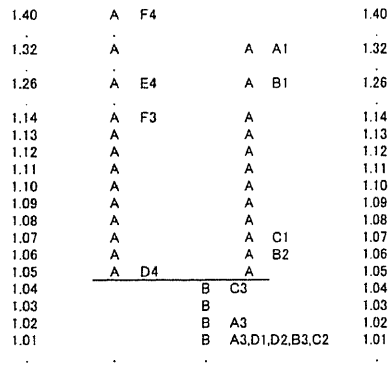

図１．　relative risk に基づく階層構造

echelon 解析により分割表の空間的な階層構造として図１が得られ、その構造に基づき空間スキャン統計量を計算し、有意に独立性から逸脱しているセル群を発見する。Echelon デンドログラムに基づき対数尤度比統計量を計算すると、第１ピークとして impaired-F(low)をトップとして、impaired-D,E,F と,moderate-F、第２ピークとして well-A(high)をトップとして、well-A,B,C と mild-B、それらのファウンデーションとして moderate-C、によって構成される、((F4, E4, F3, D4), (A1, B1,C1, B2), C3)が統計量９．７１でホットスポットの候補として選ばれる。

References

Agresti, A (1984). Analysis of ordinal categorical data. John Wiley & Sons.

Cressie, N. and Chan, N.H. (1989). Spatial modelling of regional variables. Journal of the American Statistical Association, 84, 393-401.

Kulldorff M. (1997). A spatial scan statistics, Communications in Statistics, Theory and Methods, 26, 1481-1496.

Kurihara, K. Myers, W. L., and Patil, G. P. (2000). Echelon analysis of the relationship between population and land cover patterns based on remote sensing data. Community Ecology, 1103-122.

Joint distributions associated with patterns, successes and failures in a sequence of multi-state trials

統計数理研究所 井上 潔司

1 はじめに

$\{Z_t, t \geq 1\}$ は, 集合 $\{0, 1, ..., m\}$ に値をとる試行列とする (ここでは, "0" を失敗, 残りの値 "1",...,"m" を成功と呼ぶことにする). $\mathcal{E}$ を, ある (single または compound) パターンとする. このとき, 長さ n の 試行列 Z_1, Z_2,...,Z_n に現れるパターン $\mathcal{E}$ の数を X_n, 各成功 "i" の数を $S_{n,i}$ $(i = 1, 2, ..., m)$, 失敗 "0" の数を $F_{n,0}$ で表すものとする. また, パターン $\mathcal{E}$ が r 回現れるまでに必要な試行回数を T_r とする.

X_n, T_r の周辺分布は様々な応用分野で重要な役割を果たしてきた. その導出方法もいままでに多くの研究者達によって提案されてきた. 例えば, Koutras and Alexandrou (1995), Koutras (1997) によって提案された Markov chain imbedding method は, 大変強力な方法である.

しかしながら, $(X_n, F_{n,0}, S_{n,1}, ..., S_{n,m})$, または $(T_r, F_{T_r,0}, S_{T_r,1}, ..., S_{T_r,m})$ の同時分布を考察するためには従来の手法を改良して用いる必要がある. 本報告では, Markov chain imbedding method に基づき, これら同時分布の導出方法の提案を行う. ここでは, (i) X_n が a Markov chain imbeddable variable of binomial type (M.V.B.) であるとき, また, (ii) X_n が a Markov chain imbeddable variable of returnable type (M.V.R.) であるとき, という 2 種類に分けて導出方法を提案する (M.V.B., M.V.R. の定義については, それぞれ Koutras and Alexandrou (1995), Han and Aki (1999) を参照).

実際問題においては, X_n, T_r の周辺分布だけではなく, 同時に成功, 失敗の数の分布もあわせて考察することで, そこから有益な情報を得られる場合が多くある. そこで, 最後にいくつかの計算例, 数値計算結果を工学的な応用例とあわせて紹介する予定でいる.

2 Generating functions

まず, X_n が M.V.B. である場合に, $(X_n, S_{n,1}, ..., S_{n,m})$ の同時分布を考察する. probability generating function $\phi_n(u, \boldsymbol{v})$, および double generating function $\Phi(u, \boldsymbol{v}; w)$ を次のように定義する.

$$\begin{aligned}
\phi_n(u, \boldsymbol{v}) &= E(u^{X_n} v_1^{S_{n,1}} \cdots v_m^{S_{n,m}}), \\
&= \sum_{x=0}^{\infty} \sum_{y_1=0}^{\infty} \cdots \sum_{y_m=0}^{\infty} \Pr[X_n = x, S_{n,1} = y_1, ..., S_{n,m} = y_m] u^x v_1^{y_1} \cdots v_m^{y_m}, \\
\Phi(u, \boldsymbol{v}; w) &= \sum_{n=1}^{\infty} \phi_n(u, \boldsymbol{v}) w^n, \\
&= \sum_{n=0}^{\infty} \sum_{x=0}^{\infty} \sum_{y_1=0}^{\infty} \cdots \sum_{y_m=0}^{\infty} \Pr[X_n = x, S_{n,1} = y_1, ..., S_{n,m} = y_m] u^x v_1^{y_1} \cdots v_m^{y_m} w^n.
\end{aligned}$$

このとき $\phi_n(u, \boldsymbol{v})$ は, 適当な $s \times s$ 行列 $A_{t,j}$, $B_{t,j}$ $(j = 0, 1, ..., m)$ を用いて, それぞれ次のように表される (証明は, Inoue (2002) を参照).

$$\phi_n(u, \boldsymbol{v}) = \boldsymbol{a}(\boldsymbol{v}) \prod_{t=2}^{n} \left[A_{t,0} + uB_{t,0} + \sum_{j=1}^{m} v_j (A_{t,j} + uB_{t,j}) \right] \mathbf{1}',$$

ただし, $\mathbf{1} = (1, ..., 1) \in R^s$, $\boldsymbol{a}(\boldsymbol{v})$ は, 適当な初期条件とする.

いま, $F_{n,0} + S_{n,1} + \cdots + S_{n,m} = n$ という関係式に着目する. すると $(X_n, F_{n,0}, S_{n,1}, ..., S_{n,m})$ の probability generating function $\psi_n(u, v_0, \boldsymbol{v})$, double generating function $\Psi(u, v_0, \boldsymbol{v}; w)$ は, $\phi_n(u, \boldsymbol{v})$, $\Phi(u, \boldsymbol{v}; w)$ を通して求めることができる.

$$\psi_n(u, v_0, \boldsymbol{v}) = E\left(u^{X_n} v_0^{F_{n,0}} v_1^{S_{n,1}} \cdots v_m^{S_{n,m}}\right) = v_0^n \phi_n(u, v_1/v_0, ..., v_m/v_0),$$

$$\Psi(u, v_0, \boldsymbol{v}; w) = \sum_{n=1}^{\infty} \psi_n(u, v_0, \boldsymbol{v}) w^n = \Phi(u, \boldsymbol{v}/v_0; wv_0).$$

$\psi_n(u, v_0, \boldsymbol{v})$, $\Psi(u, v_0, \boldsymbol{v}; w)$ を利用すれば, 各確率変数 $X_n, F_{n,0}, S_{n,1}, ..., S_{n,m}$ の平均, 分散, またそれらの間の共分散についても論じることができる (時間があれば当日報告).

3 補足

当日は, $s \times s$ 行列 $A_{t,j}, B_{t,j}$ $(j = 0, 1, ..., m)$ の構成方法を述べたい. $(T_r, F_{T_r,0}, S_{T_r,1}, ..., S_{T_r,m})$ の同時分布も考察する予定である. また X_n が M.V.R. である場合も扱い, M.V.B. の場合との違いに触れながら $(X_n, F_{n,0}, S_{n,1}, ..., S_{n,m})$ の probability generating function, double generating function の導出方法を与える.

最後にいくつかの具体例を取り上げ, 工学的応用とともに紹介する予定である.

参考文献

Han, Q. and Aki, S. (1999). Joint distributions of runs in a sequence of multi-state trials, *Ann. Inst. Statist. Math.*, **51**, 419–447.

Inoue, K. (2002). Joint distributions associated with patterns, successes and failures in a sequence of multi-state trials, *Research Memorandum*, No.**839**, The Institute of Statistical Mathematics, Japan.

Koutras, M. V. (1997). Waiting time distributions associated with runs of fixed length in two-state Markov chain, *Ann. Inst. Statist. Math.*, **49**, 123–139.

Koutras, M. V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach, *Ann. Inst. Statist. Math.*, **47**, 743–766.

On estimating discrete multivariate probabilities by pooling incomplete samples

関東学院大経済学部　布能 英一郎

1. Introduction　用語　$m < n$　とする。

$\mathbf{X}_1 = (X_{1,0}, X_{1,1}, \cdots, X_{1,n}) \sim Multinomial(N_1, \theta_0, \theta_1, \cdots, \theta_{n-1}, \theta_n)$,

$\mathbf{X}_2 = (X_{2,0}, X_{2,1}, \cdots, X_{2,m}) \sim Multinomial(N_2, \eta_0, \eta_1, \cdots, \eta_m)$　にて

(1)　$\mathbf{X}_2$ が $\mathbf{X}_1$ の比例配分とは、$\eta_0 = \dfrac{\theta_0}{\sum_{j=0}^{m}\theta_j}$,　$\eta_1 = \dfrac{\theta_1}{\sum_{j=0}^{m}\theta_j}, \cdots, \eta_m = \dfrac{\theta_m}{\sum_{j=0}^{m}\theta_j}$

(2)　$\mathbf{X}_2$ が $\mathbf{X}_1$ の吸収合併とは、$\eta_0 = \theta_0$,　$\eta_1 = \theta_1, \cdots, \eta_{m-1} = \theta_{m-1}$,　$\eta_m = \sum_{j=m}^{n}\theta_j$

であることを言う。

定理 1.　$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_l$　は互いに独立な多項分布からのランダムサンプリングで

$\mathbf{X}_1 = (X_{1,0}, X_{1,1}, ..., X_{1,n-1}, X_{1,n}) \sim Multinomial(N_1, \theta_0, \theta_1, ..., \theta_{n-1}, \theta_n)$

$\mathbf{X}_2$ は $\mathbf{X}_1$ の吸収合併または比例配分

$\mathbf{X}_3$ は $\mathbf{X}_1$ あるいは $\mathbf{X}_2$ の吸収合併または比例配分

$\mathbf{X}_4$ は $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ のいずれかの吸収合併または比例配分

$\vdots$

$\mathbf{X}_l$ は $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{l-1}$ のいずれかの吸収合併または比例配分

更に、$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_l$　の確率構造を樹木構造で書くことができると仮定する。更に、比例配分の場合等で MLE 計算に支障が生じるときは、このようなランダムサンプリングを行わないとする。そうすると

(a)　θ_i　$(i = 0, 1, 2, \cdots, n)$ の MLE は exact に求められる。

(b)　更に、MLE は自乗損失の下で許容的。

上記の結果を、今回はターゲットとする推定量を MLE に限定しないで考察した。

2. MLE の不偏性

Asano(1965) は、$\mathbf{X}_1 = (X_{1,0}, X_{1,1}, \cdots, X_{1,m}, \cdots, X_{1,n}) \sim Multinomial(N_1, \theta_0, \theta_1, \cdots, \theta_m, \cdots, \theta_n)$,　$\mathbf{X}_2 = (X_{2,0}, X_{1,1}, \cdots, X_{1,m}) \sim Multinomial(N_2, \dfrac{\theta_0}{\sum_{i=0}^{m}\theta_i}, \dfrac{\theta_1}{\sum_{i=0}^{m}\theta_i}, \cdots, \dfrac{\theta_m}{\sum_{i=0}^{m}\theta_i},)$

にて、θ_i の MLE が不偏推定量であることを示した。このアイデアを用いると、次のことが示される：

定理 2.　定理 1 で求められた MLE は、不偏推定量である。

Underlying distribution が Poisson の場合でも、Asano のアイデアを用いて、吸収合併や、ある種の比例配分の場合に MLE の不偏性が示せる

定理 3. $X_{1,i} \sim Poisson(\lambda_i),\ i=1,2,\cdots,m,\quad X_{2,i} \sim Poisson(\lambda_i),\ i=1,2,\cdots,m-2,\ X_{2,m-1} \sim Poisson(\lambda_{m-1}+\lambda_m),\ \cdots\cdots,\ X_m \sim Poisson(\lambda_1+\cdots+\lambda_m)$ で、これらの確率変数はすべて独立とする。このとき、λ_i の MLE は不偏推定量。 **Note** λ_i の MLE は、improper prior $\prod_{i=1}^{m}(d\lambda_i/\lambda_i)$ に関する自乗損失下でのベイズ解でもある。

定理 4. 定理 3. における分布の仮定を、$X_{1,i} \sim Poisson(\lambda_i),\ i=1,2,\cdots,m,\quad X_{2,i} \sim Poisson((\sum_{j=1}^{m}\lambda_j)\times(\lambda_i/\sum_{l=1}^{m-1}\lambda_l)),\ i=1,2,\cdots,m-1,\ \cdots\cdots,\ X_{m-1,i} \sim Poisson((\sum_{j=1}^{m}\lambda_j)\times(\lambda_i/(\lambda_1+\lambda_2)))$, $i=1,2$ に変更しても、λ_i の MLE は不偏性（およびベイズ性）が示せる。

3. Negative Multinomial(以後、NM と略記) に対する不偏推定量の考察

良く知られているように、$\mathbf{X}=(X_1,X_2,\cdots,X_n) \sim NM(k;\theta_1,\theta_2,\cdots,\theta_n)$ にて、improper prior $d\theta_1 d\theta_2\cdots d\theta_n/(\theta_0^2\theta_1\cdots\theta_n)$ によって自乗損失下で $\tilde{\theta}_i = x_i/(k+\sum_{j=1}^{n}x_j-1),\ (i=1,2,\cdots,n),\ \tilde{\theta}_0=(k-1)/(k+\sum_{j=1}^{n}x_j-1)$ が得られる。

Example 1. $\mathbf{X}_1=(X_{1,1},X_{1,2},\cdots,X_{1,n-1},X_{1,n}) \sim NM(k_1;\theta_1,\theta_2,\cdots,\theta_{n-1},\theta_n),\ \mathbf{X}_2=(X_{1,1},X_{1,2},\cdots,X_{1,n-1}) \sim NM(k_2;\theta_1,\theta_2,\cdots,\theta_{n-1}+\theta_n),\ \cdots,\ \mathbf{X}_{n-1}=(X_{n-1,1},X_{n-1,2}) \sim NM(k_{n-1};\theta_1,\theta_2+\theta_3+\cdots+\theta_n),\ X_n \sim NB(k_n;\theta_1+\theta_2+\cdots+\theta_n)$ にて、improper prior $d\theta_1 d\theta_2\cdots d\theta_n/(\theta_0^2\theta_1\cdots\theta_n)$ による自乗損失下での θ_i の Bayes 推定量 $\tilde{\theta}_i$ は

$$\tilde{\theta}_0=\frac{k_+-1}{Total-1},\quad \tilde{\theta}_j=\frac{Total-k_+}{Total-1}\,\frac{x^{[1]}}{x^{[1]}+x[+:1]}\,\frac{x^{[2]}}{x^{[2]}+x[+:2]}\times\cdots\times\frac{x[+:j]}{x^{[j]}+x[+:j]},$$

$j=1,2.\cdots$ である。但し、$k_+=\sum_{i=1}^{n}k_i,\ x_{i+}=\sum_{j=1}^{n-i+1}x_{1,j},\ x[+:j]=\sum_{i=1}^{n-j}x_{i,j},\ x^{[a]}=\sum_i\sum_{j\geq a+1}x_{i,j},\ Total=k_++\sum\sum x_{i,j}$ そして、これらの推定量は不偏。

Example 2. $(X_{1,1},X_{1,2},X_{1,3}) \sim NM(k_1,\theta_1,\theta_2,\theta_3),\quad (X_{2,1},X_{2,2}) \sim NM(k_2,(\theta_1+\theta_2+\theta_3)\frac{\theta_1}{\theta_1+\theta_2},(\theta_1+\theta_2+\theta_3)\frac{\theta_2}{\theta_1+\theta_2})$, にて、improper prior $d\theta_1 d\theta_2\cdots d\theta_n/(\theta_0^2\theta_1\cdots\theta_n)$ による自乗損失下での θ_i の Bayes 推定量 $\tilde{\theta}_i$ は不偏。

しかしながら、Underlying distribution が Negative Multinomial の場合、Asano のアイデアを用いて推定量の不偏性を示せるのは、確率構造がかなり限定された場合である。

参考文献

[1] Asano, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. *Annals of the Institute of Statistical Mathematics* **17**, 1-13.

Slippage rank tests for k location parameters in the presence of gross errors

神戸大学工学部　垣内逸郎
南山大学数理科学部　木村美善

本報告では，k 個の位置母数のロバストなスリッペイジ検定問題を定式化し，漸近的に有意水準 α となるスリッペイジ順位検定を構成する．そのとき，漸近的最小検出力の下界を評価するとともに，構成した検定が漸近的不偏検定になることを示した．

1．k 標本漸近的ロバストスリッペイジ検定問題

$X_{i1},\dots,X_{in}$ $(i=1,\dots,k)$ は，第 i 群からの大きさ n の連続型確率変数とし，その確率分布は $G_{i1},\dots,G_{in}$ に従うとする．$G_{i1},\dots,G_{in}\in\mathcal{P}(\theta_i;\epsilon,\delta)$ であり，分布の同一性は仮定しない．ここで，中心が F_θ の近傍 $\mathcal{P}(\theta;\epsilon,\delta)$ を $\mathcal{P}(\theta;\epsilon,\delta)=\{G\in\mathcal{M}_c\,;\,G(B)\geq(1-\epsilon)F_\theta(B)-\delta \text{ for all } B\in\mathcal{B}\}$, $\epsilon,\ \delta\in[0,1)$; $\epsilon+\delta<1$ とする．また，$\mathcal{P}^{(N)}(\boldsymbol{\theta};\epsilon,\delta)=\{\boldsymbol{W}_N=\otimes_{i=1}^k\otimes_{j=1}^n G_{ij}\mid G_{ij}\in\mathcal{P}(\theta_i;\epsilon_n,\delta_n),\ i=1,\dots,k;j=1,\dots,n\}$, $\boldsymbol{\theta}=(\theta_1,\dots,\theta_k)$ とする．

次のような漸近的ロバストスリッペイジ検定問題を考える．

$$\begin{aligned}&H_0:\{(\boldsymbol{W}_N)\mid \boldsymbol{W}_N\in\mathcal{P}^{(N)}(\boldsymbol{\theta}_0;\epsilon_n,\delta_n)\quad \text{for } \forall n\in\mathbf{N}\}\\&H_i(\Delta):\{(\boldsymbol{W}_N)\mid \boldsymbol{W}_N\in\mathcal{P}^{(N)}(\boldsymbol{\theta}_i(\Delta_n);\epsilon_n,\delta_n)\quad \text{for } \forall n\in\mathbf{N}\},\quad i=1,\dots,k,\end{aligned}\tag{1.1}$$

ここで，$\boldsymbol{\theta}_0=(\underbrace{\theta_0,\dots,\theta_0}_{k})$, $\boldsymbol{\theta}_i(\Delta_n)=(\underbrace{\theta_0,\dots,\theta_0}_{i-1},\theta_0+\Delta_n,\underbrace{\theta_0,\dots,\theta_0}_{k-i})$, $i=1,\dots,k$, $\epsilon_n=n^{-1/2}\epsilon$, $\delta_n=n^{-1/2}\delta$, $\Delta_n=n^{-1/2}\Delta$, $\Delta\geq 2\tau$ とする．$\tau\in(0,+\infty)$ は，$0<(\epsilon+2\delta)/\tau<\int\Lambda^+\,dF_0$, $\Lambda(x)=\partial\log f_\theta(x)/\partial\theta\,|_{\theta=0}$ を満たす定数とする．

2．スリッペイジ順位検定

R_{ij} は，合併標本 $X_{11},X_{12},\dots,X_{kn}$ における X_{ij} の順位とする．k 標本線形順位統計量 $T_{Ni}(\boldsymbol{X}_N)=\frac{1}{n}\sum_{j=1}^n a(R_{ij}(\boldsymbol{X}_N)/(N+1))$, $i=1,\dots,k$ とする．このとき，スリッペイジ検定問題 (1.1) に対して，$\boldsymbol{T}_N=(T_{N1},\dots,T_{Nk})$ に基づき次のような順位検定 $\varphi_n=(\varphi_{n0},\varphi_{n1},\dots,\varphi_{nk})$ を考える．

$$\begin{aligned}\varphi_{n0}(\boldsymbol{x}_N)&=\begin{cases}1, & \text{if } \max\limits_{1\leq j\leq k}T_{Nj}(\boldsymbol{x}_N)\leq\lambda,\\ 0, & \text{if } \max\limits_{1\leq j\leq k}T_{Nj}(\boldsymbol{x}_N)>\lambda,\end{cases}\\
\varphi_{ni}(\boldsymbol{x}_N)&=\begin{cases}\dfrac{1}{m(\boldsymbol{x}_N)}, & \text{if } T_{Ni}(\boldsymbol{x}_N)=\max\limits_{1\leq j\leq k}T_{Nj}(\boldsymbol{x}_N)>\lambda_n,\\ 0, & \text{otherwise,}\end{cases}\\
&\qquad i=1,\dots,k\end{aligned}\tag{2.1}$$

ここで，$m(\boldsymbol{x}_N)$ は $\max_{1\leq j\leq k}T_{nj}(\boldsymbol{x}_N)$ が達成される個数，λ_n は棄却限界値である．

スリッペイジ順位検定 (2.1) に対し，検定列 φ_n の漸近的最大有意水準と漸近的最小検出力は，それぞれ次のように定義される．

$$1-\liminf_{n\to\infty}\alpha_n(\varphi_n),\qquad \liminf_{n\to\infty}\sum_{i=1}^k\beta_{ni}(\varphi_n)$$

ここで，$\alpha_n(\varphi_n) = \inf\{ E_{\boldsymbol{W}_{N0}}(\varphi_{n0}) ; \boldsymbol{W}_{N0} \in \mathcal{P}^{(N)}(\boldsymbol{\theta}_0; \epsilon_n, \delta_n)\}$，$\beta_{ni}(\varphi_n) = \inf\{E_{\boldsymbol{W}_{Ni}}(\varphi_{ni}) ; \boldsymbol{W}_{Ni} \in \mathcal{P}^{(N)}(\boldsymbol{\theta}_i(\Delta_n); \epsilon_n, \delta_n)\}$，$i = 1, \ldots, k$，である．

3．スリッペイジ順位検定の漸近的結果

$\boldsymbol{\mu}^M = (\mu_1^M, \ldots, \mu_k^M)$，$\boldsymbol{D}_0(\lambda)$ を次のように定義する．

$$\mu_i^M = (1/k)(k - 2i + 1)(\epsilon + 2\delta)\big(a(1) - a(0)\big), \quad i = 1, \ldots, k,$$

$$\boldsymbol{D}_0(\lambda) = \left\{ (x_1, \ldots, x_k) \mid \max_{1 \le i \le k} x_i \le \lambda \right\}$$

このとき，λ_α を次式で決定される定数とする．

$$P\left(\boldsymbol{Z} + \frac{\boldsymbol{\mu}^M}{A} \in \boldsymbol{D}_0\left(\frac{\lambda_\alpha}{A}\right)\right) = 1 - \alpha,$$

ここで，$\boldsymbol{Z}$ は $N(\boldsymbol{0}, \Sigma)$ に従う確率ベクトルで，$\boldsymbol{\Sigma} = (\sigma_{ij}) ; \sigma_{ij} = 1 - 1/k\,(i = j)$，$-1/k\,(i \ne j)$ また，$A^2 = \int_0^1 \big(a(t) - \bar{a}\big)^2\, dt$，$\bar{a} = \int_0^1 a(t)\, dt$ である．

定理 1．$\lambda_n = \bar{a} + n^{-1/2}\lambda_\alpha$ とする順位検定 (2.1) の列 (φ_n) は，漸近的有意水準 α の検定である．すなわち，

$$\liminf_{n \to \infty} \alpha_n(\varphi_n) \ge 1 - \alpha.$$

$\boldsymbol{\nu}(\Delta) = (\nu_1(\Delta), \ldots, \nu_{k-1}(\Delta))$，$\boldsymbol{D}_k(\lambda)$ を次のように定義するとき，漸近的最小検出力の下界が次の定理で与えられる．

$$\nu_i(\Delta) = -\Delta \int_0^1 \Lambda(F_0^{-1}(t)) a(t) dt + \frac{2(k-i)}{k}(\epsilon + 2\delta)(a(1) - a(0)), \quad i = 1, \ldots, k-1.$$

$$\boldsymbol{D}_k(\lambda) = \left\{ (x_1, \ldots, x_{k-1}) \mid \max_{1 \le i \le k-1} x_i \le 0,\ \sum_{i=1}^{k-1} x_i \le -k\lambda \right\}.$$

定理 2．定理 1 で与えられた検定列 (φ_n) に対し，

$$\liminf_{n \to \infty} \sum_{i=1}^{k} \beta_{ni}(\varphi_n) \ge kP\left(\boldsymbol{U} + \frac{\boldsymbol{\nu}(\Delta)}{A} \in \boldsymbol{D}_k\left(\frac{\lambda_\alpha}{A}\right)\right),$$

ここで，$\boldsymbol{U} = (U_1, \ldots, U_{k-1})$ は，$N(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}})$ に従う確率ベクトルで $\tilde{\boldsymbol{\Sigma}} = (\tilde{\sigma}_{ij}); \tilde{\sigma}_{ij} = 2\,(i = j)$，$1\,(i \ne j)$，である．

次の条件は，漸近的不偏性の条件である．

$$\tau \int_0^1 \Lambda(F_0^{-1}(t)) a(t)\, dt \ge \frac{k}{2}(\epsilon + 2\delta)(a(1) - a(0)). \tag{A.1}$$

定理 3．条件 (A.1) の下で，定理 1 で与えられた検定列 (φ_n) は，漸近的不偏検定である．すなわち，

$$\liminf_{n \to \infty} \sum_{i=1}^{k} \beta_{ni}(\varphi_n) \ge \alpha.$$

続 確率生成母関数の利用

統計数理研究所 平野 勝臣

「 確率生成母関数の利用」の表題で，確率生成母関数 (pgf) の離散分布への有効な利用法について報告した (シンポジウム (2000.10.5-6) 於鹿児島大)．そこでは，系列が *i.i.d.* からマルコフに一般化されたとき，連に関する分布論は pgf を用いて解析することができることを具体例で報告した．また，連に関する問題だけでなく他の問題へ利用できることを示すために，k-match prblems への利用を述べた．ここでも系列が *i.i.d.* からマルコフに一般化して解析できることを報告した (Hirano and Aki (2002))．本報告ではその後の研究について述べた．

0 か 1 の値をとる確率変数の系列 $\{X_i, i = 1, 2, \ldots\}$ を考える．慣例に従って適宜上 X_i を i 番目の試行，X_i が値 1 をとることを成功，値 0 をとることを失敗と言うことにする．系列が *i.i.d.*，$P(X_1 = 1) = p$ であるとき，長さ k の成功連がはじめて起こるまでの試行数の分布をオーダー k の幾何分布といい $G_k(p)$ とかく．この分布の台を $\{a, a+1, \cdots\}$ にシフトしたときの分布を $G_k(p, a)$ とかく．長さ k の成功連がはじめて起こるまでに長さ ℓ $(\ell < k)$ の成功連 (overlapping count) の起こる回数 $G_{k-\ell}(p, k-\ell+1)$ に従う．これはオーダー間の関係として興味ある結果である．これを "negative version" とすると，この "positive version" を考察する．すなわち，オーダー k の一般 2 項分布を与え，この分布がこの性質を持っている (Aki and Hirano (2000)) ことを報告した．またパターンの待ち時間分布を求め，そこから得られる性質について報告した (Han and Hirano (2001))．

本報告は上記の論文に基づいている．詳細についてはこれらを参照されたい．

a, k, n を任意の固定した正整数とする．{0,1}-値 *iid* 確率変数列 $X_1, X_2, \cdots$ で，$P(X_i = 1) = p = 1 - q$ とする．X_i を i 番目の試行とよぶ．スコア $s(i)$ を，i 番目の試行で長さ k の 1 の連を観測したとき値 a, そうでないとき 1 をとると定義する．

定義 スコアの和が n 以下であるとき，長さ k の 1 の連 (non-overlapping) の数の分布をオーダー k の一般 2 項分布といい，$B_k(n, p, a)$ とかく．($a = 1$ のときオーダー k の 2 項分布である)．

τ を $X_1, X_2, \cdots$ において n 番目の 1 が起こるまでの試行数とし，k と ℓ は $\ell < k$ で $k > 2$ を満たす正整数で，τ_ℓ を長さ ℓ の成功連 (overlapping count) の n 番目が起こるまでの試行数とする．そのとき，次が成り立つ．

命題 1. τ までに長さ k の 1 の連 (non-overlapping) の起こる回数は $B_{k-1}(n, p, 2)$ に従う．

定理 2. τ_ℓ までに長さ k の 1 の連 ($(\ell-1)$-overlapping) の起こる回数は $B_{k-\ell}(n, p, 2)$ に従う．

m, ℓ, k は $m \leq \ell < k$ をみたす正整数とすると，上の結果は m 次マルコフ系列に拡張される．

次にパターンとその逆パターンの待ち時間問題への利用を述べる．$X_1, X_2, \ldots$ を $\Omega = \{\omega_1, \ldots, \omega_N\}$-値マルコフ系列とする．この系列において複数パターンの sooner and later waiting time problems と パターンとその逆パターンがはじめて起こるまでの待ち時間

分布について調べる．長さ k と ℓ の任意のパターンをそれぞれ $S_0 = a_1a_2\cdots a_k$ と $S_1 = b_1b_2\cdots b_\ell$, $(a_i, b_j \in \Omega, 1 \le i \le k$ で $1 \le j \le \ell,\ k \le \ell)$ とする．W_0 (resp. W_1) をはじめて S_0 (resp. S_1) が起こるまでの待ち時間，W_S をはじめて S_0 か S_1 のどちらかが起こるまでの待ち時間 (*i.e.* $W_S = \min\{W_0, W_1\}$, sooner waiting time)，W_L をはじめて S_0 と S_1 のどちらも起こるまでの待ち時間 (*i.e.* $W_L = \max\{W_0, W_1\}$, later waiting time) とする．S_i と S_j $(i, j = 0, 1)$ 間の overlapping indicator $\varepsilon_{i,j}(r)$ を，S_i の最後の r 文字が S_j のはじめの r 文字に等しいとき 1，そうでないとき 0 と定義する．

確率変数 $W_S^{(0)}$ (resp. $W_S^{(1)}$) を S_0 (resp. S_1) が S_1 (resp. S_0) より先に起こり，それまでの試行数とする．また，$p_S^{(0)}(t) = Pr(W_S^{(0)} = t)$，$p_S^{(1)}(t) = Pr(W_S^{(1)} = t)$ とおき，W_S, $W_S^{(0)}$, $W_S^{(1)}$, W_L, W_0, W_1 の確率生成母関数 (*i.e.* $\phi_S(x) = \sum_{t=1}^{\infty} p_S(t)x^t$) をそれぞれ $\phi_S(x)$, $\phi_S^{(0)}(x)$, $\phi_S^{(1)}(x)$, $\phi_L(x)$, $\phi_0(x)$, $\phi_1(x)$ とする．このとき $\phi_S^{(0)}(x)$ と $\phi_S^{(1)}(x)$ の関係式を得る．この結果から $\phi_S(x)$ は $\phi_S(x) = \phi_S^{(0)}(x) + \phi_S^{(1)}(x)$ で与えられる．$\{W_S = t\} \cup \{W_L = t\} = \{W_0 = t\} \cup \{W_1 = t\}$ の関係から W_L の確率関数と確率生成母関数を得ることができる．以上はマルコフ系列で議論できる．

ここで W_0 の確率生成母関数を求める．$\ell \to \infty$ のとき，$p_S^{(1)}(t) = 0$, $(t = 1, 2, \ldots)$ であり $\phi_S^{(1)}(x) = 0$ である．したがって W_0 の確率生成母関数 $\phi_0(x)$ が求まる．このことから $X_1, X_2, \ldots$ を *i.i.d.* 系列とし，$P(X_1 = a_i) = p_{a_i}, i = 1, 2, \ldots, k)$ とすると，この系列においてパターン S_0 がはじめて起こるまでの待ち時間分布の確率生成母関数 $\phi_0(x)$ は

$$\phi_0(x) = \frac{p_{a_1}\cdots p_{a_k}x^k}{1 - x + (1-x)\sum_{r=1}^{k-1}\varepsilon_{0,0}(r)(p_{a_{r+1}}\cdots p_{a_k}x^{k-r}) + p_{a_1}\cdots p_{a_k}x^k}$$

で与えられることがわかる．これを用いれば $X_1, X_2, \ldots$ を *i.i.d.* 系列とし，パターン $\bar{S}_0$ を パターン S_0 の逆パターンとする．$\bar{S}_0$ がはじめて起こるまでの待ち時間分布の確率生成母関数を求めると，これは $\phi_0(x)$ に等しいことがわかる．

しかしこの事実を，例えば exchangeable のような *i.i.d.* より一般的な系列の場合に確率生成母関数を用いて示すことは困難を伴う．Aki and Hirano (2002) は typical sequence を考察することによってこの事実を示している．

参考文献

Aki, S. and Hirano, K. (2000). Numbers of success-runs of specified length until certain stopping time rules and generalized binomial distributions of order k, *Ann. Inst. Statist. Math.* **52**, 767-777.

Aki, S. and Hirano, K. (2002). On waiting time for reversed patterns in random sequences. To appear in *Ann. Inst. Statist. Math.*

Han, Q. and Hirano, K. (2001). Sooner and later waiting time problems for patterns in Markov dependent trials. *Research Memorandum*, The Institute of Statistical Mathematics, No. 816.

Hirano, K. and Aki, S. (2002). On k-match problems. To apperar in *Journal of Statistical Planning and Inference* .

巨大数・微小数を処理する計算アルゴリズムとその離散型確率計算への応用

岡山理科大学・情報科学科：中村　忠
岡山大学・教育学部：平井　安久

1．はじめに

2項分布 B(n,p) の確率関数 $b(x;n,p)$ や分布関数 $B(x;n,p)$ は四則演算だけで計算可能な量である．n が大きいとき，パソコンの市販ソフトでこれらの計算ができるものにはデータ処理ソフト(SPSS，S-Plus 等)や数式処理ソフト(マセマティカ，メイプル等)がある．個人が所有できるパソコンの処理速度などの機能向上により，高価なソフトウェアや近似法を使わず直接計算できる環境が整っていると思われる．このことが動機となって，我々は通常の計算機の算術システムでは処理できない巨大数・微小数が処理できる算術システム（モグラ算術システムという）を開発した．この算術システムを離散型確率分布の確率関数や分布関数の値の計算に応用した．これまでに知られている方法で計算した値とモグラたたき法で計算した値の精度の比較をし，モグラたたき法と近似法との役割を論ずる．結果として，標本の大きさが1億より小さいときは，計算精度や計算時間から判断して，2項分布の確率関数や分布関数の値をモグラたたき法で計算する方法は有効であることが示せた．正確な2項分布関数の値と Peizer-Pratt の近似式を用いて，従来とは逆に2項分布関数の値から正規分布関数の値を計算するという問題点を提起する．

2．モグラたたき法

実数 s と整数 i の組からなる集合 $\mathcal{M}=\{<s,i>;10^{-7}\le|s|\le10^{7},-L\le i\le L\}$ を考える．ここに $L=9007199254740991$（約 9007 兆 1992 億）．$\mathcal{M}$ の元 $<s,i>$ をモグラたたき表現という．モグラたたき表現 $<s,i>$ は i が小さいときは実数 $s\times10^{7i}$ の別表現と思ってよい．コンピュータに標準で装備されている算術システムでは，巨大な i に対しては，実数 $s\times10^{7i}$ を直接に表現できないが，モグラたたき表現では可能である．$\mathcal{M}$ に四則演算

乗算 $<s,i>\times<t,j>$，除算 $<s,i>/<t,j>$，
加算 $<s,i>\oplus<t,j>$，減算 $<s,i>-<t,j>$

を導入する．この演算は処理する数が 10^{7} より大きくなったら 10^{-7} で小さくなるようにたたき，10^{-7} より小さくなったら 10^{7} でたたくといった特徴を持つ．たたかれた数を補正するための10のベキ乗はモグラ表現の第2成分で表す．また，等号，不等号，組み込み関数なども導入される．

3．2項確率関数の計算アルゴリズム

2項確率分布関数の値をモグラたたき法で計算する準備として，確率関数の計算をする必要がある．2項確率関数 $b(x;n,p)$ の値をモグラたたき法で計算するアルゴリズムを述べる．$b(0;n,p)$ および $b(x;n,p)$ $(0<x<n)$ についていろいろな計算方法を比較・検討する．精度や計算時間の観点から良い方法を選択する．実際に，標本の大きさが２０億までの数値実験を行った．あらかじめ，何種類かの階乗 $k!$ の値を計算しておき，これを利用して $b(x;n,p)=nCx\,p^{x}q^{n-x}$ をモグラたたき法で計算する方法がよいことがわかった．

4．2項分布関数の計算アルゴリズム

正規近似による方法としては，Stirling の公式を使用する方法，連続補正をした正規近

似式，正規 Gram-Chalier 近似式(竹内(1975),竹内啓編(1997))，Camp-Paulson の近似式(Camp and Paulson,1951)，Peizer-Pratt の近似式(Peizer and Pratt,1968)などが知られている．これまでの数値結果によると，Peizer-Pratt の近似式Ⅱが一番精度がよいことが報告されている．ここでは，我々が提案するモグラたたき法と Peizer-Pratt の近似式Ⅱとの精度の比較をしながら，提案する方法の良さ具合を調べる．

前節で選択された確率関数の値を計算する方法を利用して，2 項分布関数 $B(x;n,p)=\sum_{i=0}^{x} b(i;n,p)$，$x=0,\cdots,n$ の値を求める．大きな n に対しては膨大な項数の和の計算に時間がかかることが弱点となる．各計算を倍精度で行うので，その精度は高々 15 ないし 16 桁であることの 2 点を考慮すれば，モード m (または平均 np)からかなり離れた x に対する $b(x;n,p)$ の値の小さい項の計算は省略可能となる．このような x の範囲 (x_L,x_R) を見つけるために，以下のような改良された Uspensky の不等式(Kambo & Kotz,1966)と Bahadur の不等式(Bahadur, 1960)を用いる．これにより，計算精度を保持しながら計算時間を短縮することが可能になる．結局，和

$$B(x;n,p)=b(x_L+1;n,p)+b(x_L+2;n,p)+\cdots+b(x;n,p)$$

のような形をどのような方法によって求めるかということになる．ここでも精度を落とさないためにいろいろな計算方法を導入し，精度や計算時間の観点から良い計算方法を選択する．

5．今後の課題

2 項分布関数 $B(x;n,p)$ の値を標準正規分布関数 $\Phi(t)$ と Peizer-Pratt の近似式 $T(x;n,p)$ で近似する方法はかなり正確であるということはよく知られている．すなわち，

$$B(x;n,p)\doteqdot\Phi(T(x;n,p)).$$

ここでは従来のやり方とは逆の方法を考える．$\Phi(-8)<10^{-15}$ であるから，$-8<t<8$ なる t が与えられたとき，$\Phi(t)$ の値を 2 項分布関数 $B(x;n,p)$ の近似値 $B(x;n,p)$ を使って求めることを考える．集合 $S=\{x;0<x<n,\ -8<T(x;n,p)<8\}$ の要素がたくさんあるという条件の下でよい近似が得られると思われる．実際，標本の大きさが 1 千万で $p=0.5$ のとき，S の要素の個数は約 25315 である．区間 $(-8,8)$ を 25315 個に等分すると 1 区間の長さは 0.000632036 となり，かなり小さく分割される．言い換えれば，$T(x;n,p)$ が t に十分近いような整数 x が選べることになる．このとき，$B(x;n,p)$ を $\Phi(t)$ の近似値として採用できるであろう．他の近似式を利用した方法も考えられる．標本の大きさといろいろな近似式を組み合わせてどれがよいかを調べ，どれが実用に耐えうる近似精度を持つかを調べることを今後の課題とする．

一標本モデルにおける分布探索による統計的推測論

白石高章 横浜市大総合理学研究科

1 序

$(X_1,\cdots,X_n)$ を連続分布関数 $F(\frac{x-\mu}{\sigma})$ をもつ母集団からの大きさ n の無作為標本とする. さらに, $F(x)$ の密度関数 $f(x)\equiv F'(x)$ は $f(-x)=f(x)$ を満たす 0 について対称な関数とし, 一般性を失うことなく $\int_{-\infty}^{\infty}x^2dF(x)=1$ と仮定する. すなわち $X_1,\cdots,X_n$ は互いに独立で各 X_i は μ について対称な同一の連続分布関数 $F(\frac{x-\mu}{\sigma})$ をもつ. μ と σ^2 は, それぞれ X_i の平均と分散であるが未知パラメータとする.

帰無仮説 $H_0:\mu=\mu_0$ v.s. 対立仮説 $H_1:\mu\neq\mu_0$ (μ_0 は定数)

の場合について水準 α の検定, 点推定と区間推定の手法として, $F(x)$ が標準正規分布か, 未知, または正規分布の近傍に入る場合により, それぞれパラメトリック法, ノンパラメトリック法, セミパラメトリック法を使うことができる. これら 3 つの手法のシミュレーション比較を行い, 特長を述べる. この特長と正規性の検定法, 分布の探索法によりこれら 3 つの手法の 1 つを選択する方法を解説する. 最後に分布の探索法による解析手法とパラメトリック法, ノンパラメトリック法, セミパラメトリック法の比較を行う.

2 手法の比較

以下の結論を得る.

○ 観測値が正規分布に従っている場合は, 頑健な手法は最良な手法に比べてほんの少し劣り, ノンパラメトリック法が最もよくない.

○ 観測値が混合正規分布 $0.95N(0,1)+0.05N(0,9)$, 異常値をもつ混合正規分布 $0.98N(0,1)+0.02I_5$, ロジスティック分布 $LG(0,\frac{\sqrt{3}}{\pi})$ などの正規分布に近い分布に従っている場合は, 頑健な手法が最も良く, 正規母集団での最良手法は劣る.

○ 観測値が両側指数分布 $DE(0,\frac{1}{\sqrt{2}})$ などの正規分布からかなり離れた分布に従っている場合は, ノンパラメトリック法が最も良く, 正規母集団での最良手法は非常に劣る. 頑健な手法はノンパラメトリック法にくらべれば劣るが, 正規母集団での最良手法よりも非常に良い.

3 分布の探索による手法

前節の手法の特長からつぎの解析チャートに沿ってデータ解析することが考えられる.

解析チャート

< 1 > 正規性の検定,　< 2 > 分布の探索, < 3 >, < 4 > 経験分布関数と分布関数の重ねかきグラフ

↓

① < 1 > で正規性が棄却されず, < 2 > で正規分布が選択され, < 3 > により正規性が妥当と認められればパラメトリック法を選択 ② < 2 > で両側指数分布が選択されればノンパラメトリック法を選択 ③ これら以外であればセミパラメトリック法を選択

上記の解析チャートに沿った解析は1つの統計手法とみなせる. この解析手法による推定量を μ^* とし, 標本平均 $\tilde{\mu}$, 順位推定 $\hat{\mu}$, M 推定 $\check{\mu}$ との相対効率により比較したものを, つぎの表に載せている.

表 8　$n = 15$ のときの分布探索での推定量の相対効率

$F(x)$	$e(\mu^*, \tilde{\mu})$	$e(\mu^*, \hat{\mu})$	$e(\mu^*, \check{\mu})$
N(0,1)	0.98	1.02	1.00
CN	1.16	0.99	0.98
CO	1.49	0.99	0.98
LG$(0, \frac{\sqrt{3}}{\pi})$	1.01	1.01	1.01
DE$(0, \frac{1}{\sqrt{2}})$	1.36	0.98	0.98

表　$n = 30$ のときの分布探索での推定量の相対効率

$F(x)$	$e(\mu^*, \tilde{\mu})$	$e(\mu^*, \hat{\mu})$	$e(\mu^*, \check{\mu})$
N(0,1)	0.98	1.02	1.00
CN	1.16	0.99	0.98
CO	1.32	1.02	1.00
LG$(0, \frac{\sqrt{3}}{\pi})$	1.08	1.01	1.00
DE$(0, \frac{1}{\sqrt{2}})$	1.31	0.98	0.99

4　参考文献

1. Davison, A. C. and Hinkley, D. V. (1997). Bootstrap Methods and their Application. Cambridge University Press.
2. Shiraishi, T. (1990). R-estimators and confidence regions in one-way MANOVA. J. Statist. Plan. Infer., 24, p203-214.
3. Shiraishi, T. (1996). On scale-invariant M-statistics in multivariate k samples. J. Japan Statist. Soc., 26, p241-253.
4. Shiraishi, T. (1998). Studentized robust statistics in multivariate randomized block design. J. Nonparametric Statist., 10, p95-110.
5. 白石高章 (2002). 1標本, 2標本モデルにおける頑健な信頼区間. 京大数理解析研シンポジウム予稿集.

離散複合確率分布の漸化式

北野　昌志 (慶応大・理工・院)
清水　邦夫 (慶應大・理工)
Ong, S.H. (Univ. of Malaya)

1　はじめに

K は非負整数上の確率変数, $Z_1, Z_2, \ldots$ は独立同分布に従い, K とは独立な確率変数列を表すとする. この時,

$$X = \begin{cases} \sum_{i=1}^{K} Z_i, & K > 0 \\ 0, & K = 0 \end{cases}$$

を複合変数という. 保険請求では, K は請求数, Z_i は請求額, X は総請求額を表す. 本稿では, Z_i は非負整数上の確率変数であることを仮定し、Z_i の確率関数を $s(z)$, $z = 0, 1, 2, \ldots$ で表す.

Panjer (1981) は, K の確率関数 p_k が 2 項漸化式

$$p_k = p_{k-1}\left(a + \frac{b}{k}\right), \quad k = 1, 2, 3, \ldots$$

を満たす時, 複合変数 X の確率関数が満たす漸化式を導いた. これを拡張して, 本稿では K の確率関数 p_k が 3 項漸化式

$$p_k = p_{k-1}\left(a + \frac{b}{k}\right) + p_{k-2}\left(c + \frac{d}{k} + \frac{e}{k-1}\right), \quad k = 2, 3, 4, \ldots \tag{1}$$

を満たす時, X の確率関数が満たす漸化式を導く.

2　複合確率関数の漸化式

K が (1) を満たす時, X の確率関数 $f(x)$ は $x = 0$ の時,

$$f(0) = \sum_{k=0}^{\infty} p_k \{s(0)\}^k = G_K(s(0))$$

となり, $x > 0$ の時, 漸化式

$$\begin{aligned} f(x) \;=\; & \frac{1}{1 - as(0) - c(s(0))^2}\left(\left\{p_1 - (a+b)p_0 + eH_K(s(0))\right\}s(x)\right. \\ & + \sum_{j=1}^{x}\left[\left(a + \frac{bj}{x}\right)s(j) + \left(c + \frac{dj}{2x}\right)s^{2*}(j)\right]f(x-j) \\ & \left. + e\sum_{i=1}^{x}\sum_{j=1}^{i}\frac{j}{i}s(x-i)s(j)f(i-j)\right) \end{aligned}$$

を満たす．ただし，$s^{2*}(z) = P(Z_1 + Z_2 = z)$ を表し，G_K は K の確率母関数，H_K は

$$H_K(t) = \int_0^t G_K(u)\mathrm{d}u = \sum_{k=0}^{\infty} \frac{1}{k+1} p_k t^{k+1}$$

である．特に，$c = d = e = 0$ の時，Panjer (1981) の結果に帰着する．

3　3項漸化式を満たす分布

(1) を満たす分布はいくつか紹介されている．例えば，$e = 0$ では非心負の 2 項分布 (Ong and Lee, 1979) などがある．

(1) を満たす分布として新しい分布を紹介する．合流型超幾何関数のテイラー展開 (Erdélyi et al., 1953, p. 283, eq. (2)) を $\Lambda^{-1} = q = 1 - p$ $(0 < p < 1)$, $\Lambda x = \lambda$ $(\lambda > 0)$, $a = s$ $(s \geq 0)$, $c = N + 1$ (N は正の整数) と置き換えると確率関数

$$p_k = \binom{N}{k} p^k q^{N-k} \frac{{}_1F_1(s, N-k+1; \lambda q)}{{}_1F_1(s, N+1; \lambda)}, \quad k = 0, 1, 2, \ldots \tag{2}$$

を得る．ただし、(2) は $k \geq N$ の時は

$$p_k = \frac{N!\lambda^{k-N} p^k (s)_{k-N}}{k!(k-N)!} \frac{{}_1F_1(s+k-N, k-N+1; \lambda q)}{{}_1F_1(s, N+1; \lambda)}$$

と解釈する．この分布は，特に $s = 0$ もしくは $\lambda \to 0$ の時、2 項分布 B (N,p) となり，もしも $s = N + 1$ の時は Charlier 級数分布 (Ong, 1988) の確率関数である．よって，確率関数 (2) を持つ分布を 一般 Charlier 級数分布と呼ぶ．漸化式

$$p_k = \frac{p}{q}\left(-1 + \frac{N + \lambda q + 1}{k}\right) p_{k-1} + \frac{\lambda p^2}{q}\left(\frac{N + 2 - s}{k} - \frac{N + 1 - s}{k-1}\right) p_{k-2}, \quad k = 2, 3, 4, \ldots$$

を満足する．

また，同様に $\Lambda = p = 1 - q$ $(0 < p < 1)$, $\Lambda x = \lambda$ $(\lambda > 0)$, $1 - c = n$ ($n \geq 2$ である整数) , $a = 1 - n$ と変換すると負の二項タイプの分布が得られ，この分布もまた(1) を満たした確率関数を持つ．

参考文献

Erdélyi, A. et al., 1953. Higher Transcendental Functions. McGraw-Hill, New York.

Ong, S.H., Lee, P.A., 1979. The non-central negative binomial distribution. Biom. J. 21, 611-627.

Ong, S.H., 1988. A discrete charlier series distribution, Binom.J., 30, 1003-1009

Panjer, H.H., 1981. Recursive evaluation of a family of compound distributions. Astin Bulletin 12, 22-26.

Minimax Empirical Bayes Ridge-Principal Component Regression Estimators

久保川達也 (東京大学)　M.S. Srivastava (University of Toronto)

線形回帰モデル $\boldsymbol{y} = \boldsymbol{A}\beta + \epsilon$ において説明変数の間に強い相関関係，すなわち多重共線性が存在する場合に回帰係数ベクトル $\beta \in \boldsymbol{R}^p$ を推定する問題を考える。ここで $\boldsymbol{y}$ は N 次元の観測値ベクトル，$\boldsymbol{A}$ は $N \times p$ 計画行列，N 次元誤差ベクトル ϵ は $\mathcal{N}_N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ に従っているとする。

β の最小2乗推定量 (LS) $\widehat{\beta} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\boldsymbol{y}$ の共分散は $\mathbf{Cov}\,(\widehat{\beta}) = \sigma^2(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ で与えられる。多重共線性が存在する場合は，$(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ が ill-conditined になるため LS は不安定になる。問題を単純にするために，$(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ を対角化する直交行列 $\boldsymbol{H}$ をとる。すなわち，$\boldsymbol{H}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{H}^t = \boldsymbol{D} = \mathrm{diag}\,(d_1, \ldots, d_p)$, $d_1 \geq \cdots \geq d_p$, とし，$\boldsymbol{x} = (x_1, \ldots, x_p)^t = \boldsymbol{H}\widehat{\beta}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t = \boldsymbol{H}\beta$ とおく。多重共線性が存在するときには少なくとも d_1 が非常に大きいことを意味する。

最小2乗推定量の不安定性を回避するための方法として主成分回帰法，リッジ回帰法，部分最小2乗法などが知られているが，ここでは Hoerl and Kennard (1970) によって導出されたリッジ回帰推定量

$$\widehat{\beta}^R(\lambda) = [\boldsymbol{A}^t\boldsymbol{A} + k\boldsymbol{I}]^{-1}\boldsymbol{A}^t\boldsymbol{y} = \widehat{\beta} - [\boldsymbol{I} + \lambda\boldsymbol{A}^t\boldsymbol{A}]^{-1}\widehat{\beta} \quad \text{for} \quad \lambda = 1/k,\ k > 0$$

の枠組みを取り上げることにする。これを標準系で表すと成分毎に

$$\hat{\theta}_i^R(\lambda) = \frac{\lambda}{d_i + \lambda}x_i = x_i - \frac{d_i}{d_i + \lambda}x_i$$

と書かれるので，x_i のバラツキの程度 d_i によって調整されていることがわかる。ここで λ は x_i を縮小する程度を調整するパラメタであり，交差検証法を用いて予測誤差の推定値を小さくするように決めるのが実際の場面ではよくなされる。しかしその導出はそれほど容易ではなく，LS を改良しているか否かなと解析的な性質を調べることは困難である。そこで経験ベイズ法の枠組みで λ を推定して推定量 $\widehat{\beta}^R(\hat{\lambda})$ の性質を調べてみることにする。

この研究では，ランク q の $p \times q$ 行列 $\boldsymbol{C}$ に対して仮説 H_0 : $\beta = \boldsymbol{C}\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \boldsymbol{R}^q$, が予想される場合を取り扱う。このとき，$\beta$ の経験ベイズ推定量は次のように求められる。事前分布 $\beta \sim \mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}, \sigma^2\lambda\boldsymbol{I}_p)$ を想定すると事後分布は $\beta\,|\,\widehat{\beta} \sim \mathcal{N}_p\left(\widehat{\beta}^B(\lambda, \boldsymbol{\alpha}), \sigma^2(\boldsymbol{A}^t\boldsymbol{A} + \lambda^{-1}\boldsymbol{I})^{-1}\right)$ となり，β のベイズ推定量は

$$\widehat{\beta}^B(\lambda, \boldsymbol{\alpha}) = (\boldsymbol{A}^t\boldsymbol{A} + \lambda^{-1}\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{A}(\widehat{\beta} - \boldsymbol{C}\boldsymbol{\alpha}) + \boldsymbol{C}\boldsymbol{\alpha} = \widehat{\beta} - \left(\boldsymbol{I} + \lambda\boldsymbol{A}^t\boldsymbol{A}\right)^{-1}(\widehat{\beta} - \boldsymbol{C}\boldsymbol{\alpha}).$$

と書かれる。未知母数 λ, $\boldsymbol{\alpha}$ は $\widehat{\beta}$ の周辺分布 $\mathcal{N}_p\left(\boldsymbol{C}\boldsymbol{\alpha}, \sigma^2\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}\}\right)$ に基づいて推定される。1つの方法は $\boldsymbol{\alpha}$, λ を　$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{C}^t\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{A}^t\boldsymbol{A}\widehat{\beta}$, $\hat{\lambda}_{EB} = \max(\lambda^*, \lambda_0)$ で推定することであり，λ^*, λ_0 は方程式

$$(\widehat{\beta} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})^t\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda^*\boldsymbol{I}\right\}^{-1}(\widehat{\beta} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}) = \frac{p - q - 2}{n + 2}S$$

$$\sum_{i=1}^{p}(1 - b_{ii})\frac{d_i - d_p}{d_i + \lambda_0} = (p - q - 2)/2$$

の解として与えられる。このとき，経験ベイズ推定量 (EB)

$$\widehat{\beta}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}}) = \widehat{\beta} - (\boldsymbol{I} + \hat{\lambda}_{EB}\boldsymbol{A}^t\boldsymbol{A})^{-1}(\widehat{\beta} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})$$

が Strawderman の2乗損失関数に関して最小2乗推定量 $\widehat{\beta}$ を改良すること，すなわちミニマクスになることが示される。

経験ベイズ推定量 (EB) の欠点は，$\boldsymbol{\alpha}$ のランク q が大きいときには改良の程度が小さくなってしまうことにある。そこで階層的ベイズ推定量 (HB) を考える。まず，階層的事前分布

$$\begin{aligned}\beta\,|\,\boldsymbol{\alpha} &\sim \mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}, \sigma^2\lambda\boldsymbol{I}_p),\\ \boldsymbol{\alpha} &\sim \mathcal{N}_q(\boldsymbol{\alpha}_0, \sigma^2\tau\boldsymbol{I}_q),\end{aligned}$$

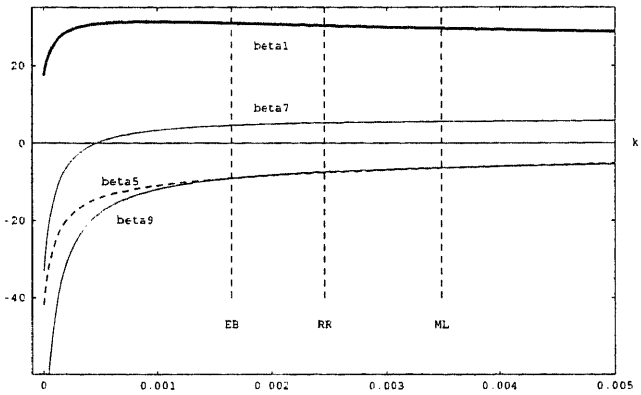

Figure 1: $\beta_1, \beta_5, \beta_7, \beta_9$ のリッジ回帰推定値の軌跡　(横軸は $k = 1/\lambda$ の値で，縦軸に平行な波線 EB, RR, ML は $1/\hat{\lambda}_{EB}, 1/\hat{\lambda}_{RR}, 1/\hat{\lambda}_{ML}$ の値を示している。)

を想定すると，周辺的事前分布は $\boldsymbol{\Psi} = \lambda \boldsymbol{I}_p + \tau \boldsymbol{C}\boldsymbol{C}^t$ に対して $\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}_0, \sigma^2\boldsymbol{\Psi})$ と表される。このとき事後分布は $\boldsymbol{\beta}\,|\,\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau), (\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{\Psi}^{-1})^{-1})$ となり，$\boldsymbol{\beta}$ の階層的ベイズ推定量は

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau) &= (\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{\Psi}^{-1})^{-1}(\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{\alpha}_0) \\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\left\{\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)\right\}\end{aligned}$$

で与えられる。ただし

$$\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau) = \widehat{\boldsymbol{\alpha}}(\lambda) - \left[\boldsymbol{I}_q + \tau\boldsymbol{C}^t\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\boldsymbol{C}\right]^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0)$$

であり，重み付き最小2乗推定量 $\widehat{\boldsymbol{\alpha}}(\lambda) = (\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{G}^{-1}\widehat{\boldsymbol{\beta}}$ を $\boldsymbol{\alpha}_0$ の方向へ縮小した形をしている。

超母数 λ, τ は $\widehat{\boldsymbol{\beta}}$ の周辺分布 $\mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}_0, \boldsymbol{\Psi} + (\boldsymbol{A}^t\boldsymbol{A})^{-1})$ における最尤推定量 $\hat{\lambda}_{HB} = \max(\lambda^{**}, 0)$, $\hat{\tau}_{HB} = \max(\tau^{**}, 0)$ によって与えられる。ただし λ^{**}, τ^{**} は $\boldsymbol{G} = \boldsymbol{G}(\lambda) = (\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p$ に対して，方程式

$$\begin{aligned} n(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0)^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) &= S \cdot \mathrm{tr}\,(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}, \\ n(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0)^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}\boldsymbol{C}\boldsymbol{C}^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) &= S \cdot \mathrm{tr}\,\boldsymbol{C}\boldsymbol{C}^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}\end{aligned}$$

の解である。以上より $\boldsymbol{\beta}$ の階層的経験ベイズ推定量は

$$\widehat{\boldsymbol{\beta}}^{HB} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{HB} + \hat{\tau}_{HB}\boldsymbol{C}\boldsymbol{C}^t\right\}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0)$$

で表される。しかし残念ながらこの推定量のミニマクス性の証明は現段階ではできていない。

モンテカルロ・シミュレーションによって2乗誤差損失に関するリスクの挙動を通して推定量の良さを比較してみると，$H_0: \boldsymbol{\beta} = \boldsymbol{0}$ の方向へ縮小した経験ベイズ推定量 $\widehat{\boldsymbol{\beta}}^R(\hat{\lambda}_{EB}, \boldsymbol{0})$ や階層的経験ベイズ推定量 $\widehat{\boldsymbol{\beta}}^{HB}$ のリスクの挙動が優れており，d_1 が大きいときには最小2乗推定量に対する改良度が極めて大きくなっていることが示される。Marquardt and Snee (1975) によって扱われたデータに対してリッジ軌跡を描いてみると，Figure 1 となり，最小2乗推定値 ($k = 0$ のときの値) が不安定になっていることがわかる。$\hat{\lambda}_{ML}$ は $H_0: \boldsymbol{\beta} = \boldsymbol{0}$ の方向へ縮小したときの階層的経験ベイズ推定量の λ の推定値を表しており，$\hat{\lambda}_{EB}$, $\hat{\lambda}_{ML}$ は安定した推定値を与えていることがわかる。

試験問題の母集団とその構築

統計数理研究所　柳本 武美

1．動機

今日大学での高等教育を含めて、新しい試みが多くなされている。医学・歯学生に対する臨床実習前試験について考察する機会に、より広い視野で試験を考えた。特に、実際の試験問題は何らかの母集団からの無作為標本であると考え、試験の構築を論じると、新しい統計学、あるいはその周辺、の問題が見えてくると考えた。

2．議論の前提

共用試験では、1) コア・カリキュラムを作成する、 2) コア・カリキュラムに対応する項目（すべて５肢択一問題）のプールを作成する、3) 各受験者には分野別の層別無作為抽出により試験問題を選ぶ。この手続きの試験を CBT (Computer-Based testing) と呼んでいる。しかし、その進歩はコンピューターを使うことから生じているのではなく、無作為抽出の導入から生じている。そこで、項目プールの作成と試験問題の無作為抽出からなる枠組を無作為化項目試験 (Randomized item testing) と呼ぶ。

3．無作為化項目試験

無作為化項目試験では、次の連鎖を通して推定する：1) 求められる知識内容、2) 明示された試験内容、3) 項目プール、4) 出題される項目。この中で、最も大きな概念的なギャップが明示された試験内容と項目プールの間にある。一方、項目プールと出題される項目とのギャップは、本当は大きいけれども、無作為抽出により回避できる。

科学的な推論において、常識的科学観である論理実証主義を越えて、より洗練された命題が多くある。Linn 編（1989, chapter 2）には、試験の妥当性の議論で論理実証主義の影響が指摘されている。本稿での議論と直接に関連する命題として、「観察によって仮説は否定できない」がある。実際の観察には理論負荷性と呼ばれる多くの仮定を必要とすることを強調されている。別の見方をすると、観察という標本から、真の母集団でない、仮定された母集団への推論が困難であることを示している。

項目プールが作成されると、項目プールから出題項目を適切に選ぶ。無作為化抽出が最初に考えられるが、単純な無作為抽出が良いとは限らない。項目プールを層別して、出題される項目に対応する分野の偏りを減らすことも可能だし、予め推定された困難度に基づいて、出題項目の困難度を平均化させることもできる。前者は層別抽出と呼ばれ、後者は割り付けにおいて最小化法と呼ばれている

4．試験問題の母集団

前節では試験問題を標本と見なした。そして項目プールを代理 (surrogate) 母集団と見なした。この節では、もし標本と見なさない場合の帰結を議論する。

4-1. 標本でない試験問題

試験問題が明示された試験内容を代表していないとすれば、一体その試験は何を測定するのかが不明になる。だから試験問題は、試している内容を代表させるしかない。それでは、無作為抽出、あるいはその変形、以外に試験問題が明示された試験内容を代表できるのだろうか。この問題は、標本調査における標本抽出あるいは薬効評価における2群への割り付けと同じ問題である。一見愚直な無作為化が最も正統な方法である。

4-2. 項目プールの構築

項目プールの作成には、膨大で高度に知的な作業を必要とする。試験は明示化された試験内容の知識量を測定することが目的であるが、実際に出題されるのは項目プールの部分集合である。項目プールには 10,000 のオーダーの項目が望まれる。

5．項目プールの維持・管理

項目プールを作成しておくことの長所は、その項目の困難度、識別力を評価して悪問を排除できることにある。評価は受験者あるいは協力者による実際の解答結果を解析して得られる。項目反応理論では、多肢選択問題 j を受験者 i が解いた場合に正答する確率は

$$P_j(\theta_i) = \frac{\exp\{a_j(\theta_i - b_j)\}}{1+\exp\{a_j(\theta_i - b_j)\}} + \frac{1}{1+\exp\{a_j(\theta_i - b_j)\}} \cdot c_j$$

とされる。右辺第1項は正答肢が特定できる確率であり、第2項は特定できなくても運良く正答する確率が c_j であることを想定している。第2項を無視すると、これは線型共役連結回帰モデル（Linear canonical link regression model）でもあるロジットモデルである。各項目の特徴を計量的に把握しておくことにより、項目の実用上の価値を高める。一つの例として、悪問を排除するための有益な情報を与えることがある。

6．終わりに

従来の試験は、個人の努力・才能に依存し、山勘・体験・意欲に満ちていた。これに対して無作為化項目試験では、組織的で体系だった知的な営みにより、経験の蓄積、絶えまない点検・改善を行い、判定の変動を減少させる。

従属確率変数列に対するU-統計量の極限定理について

金川 秀也（武蔵工業大学工学部）

$\{\xi_j, j \geq 1\}$ を分布をμとする実数値確率変数列とする．また$u(x_1, x_2, \cdots, x_k)$を実対称関数とする．$u(x_1, x_2, \cdots, x_k)$を核関数とする統計量を対称統計量と呼ぶ．

Example 1. U-統計量: $U_n := \dfrac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} u(\xi_{i_1}, \xi_{i_2}, \cdots, \xi_{i_k})$

Sample mean : $u(x) = x \quad (k=1)$

Sample variance : $u(x_1, x_2) = \dfrac{1}{2}(x_1 - x_2)^2 \quad (k=2)$

Cramér-von Mises-Smirnov statistic :

$$u(x_1, x_2) = \int_0^1 w(u)\left(I_{\{x_1 \leq u\}} - u\right)\left(I_{\{x_2 \leq u\}} - u\right)du \quad (k=2)$$

このとき

$$U_n = \frac{1}{n-1}\int_0^1 w(x)\left(F_n(x) - 1\right)^2 dx,$$

ただし $w(x)$ は重み関数、$F_n(x)$ は$\{\xi_j, 1 \leq j \leq n\}$の経験分布関数.

Example 2. V-統計量: $V_n := n^{-k} \sum_{1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq n} u(\xi_{i_1}, \xi_{i_2}, \cdots, \xi_{i_k})$

任意の $x_2, \cdots, x_k$ に対して

$$\int_{-\infty}^{\infty} u(y, x_2, \cdots, x_k)\mu(dy) = 0$$

のとき$u(x_1, x_2, \cdots, x_k)$は退化しているという.

$\{\xi_j, j \geq 1\}$がi.i.d.の場合はHoeffding(1948)によるH-分解によって退化していない$u(x_1, x_2, \cdots, x_k)$ に関するU-統計量、V-統計量について中心極限定理（漸近正規性）、Donsker型不偏原理、概収束型不偏原理、漸近展開、大偏差原理など極めて精密に漸近的な性質が調べられている。またYoshihara(1976)によって$\{\xi_j, j \geq 1\}$がϕ-mixing性やabsolutely regular性 程度の弱従属性を持つ場合は$\{\xi_j, j \geq 1\}$をi.i.d.確率変数列で近似することによってH-分解を用いて退化していない$u(x_1, x_2, \cdots, x_k)$に関

するU-統計量、V-統計量について中心極限定理が成り立つことが示されている。

一方、核関数$u(x_1,x_2,\cdots,x_k)$が退化している場合や$\{\xi_j, j\geq 1\}$が強い従属性を持つ場合は直接H-分解が使えないために何らかの代わりの方法を見つける必要がある。その一つとして$u(x_1,x_2,\cdots,x_k)$をフーリエ級数展開し、各種の漸近的な性質を調べることについて考察する.

$T=[0,1]$、$C(T)$をT上の連続関数全体、$C(T)$上の内積を
$\langle f,g\rangle=\int_T f(t)\overline{g(t)}dt$, $f,g\in C(T)$、$e_1,e_2,\cdots$ を$C(T)$上の正規直行系とする。
$u:T^m\to\mathbf{R}$に関するフーリエ級数展開は次のように表される. フーリエ係数は、
$k=(k_1,k_2,\cdots,k_m)\in Z^m$ （Zは整数全体）として

$$\hat{u}(k)=\int_{T^m}u(t)\overline{e_k(t)}dt$$

で与えられる. ただし、$t=(t_1,t_2,\cdots,t_m)\in T^m$、

$$e_k(t)=e_{k_1}(t_1)e_{k_2}(t_2)\cdots e_{k_m}(t_m)=\exp\left(2\pi i\sum_{j=1}^m k_jt_j\right)$$

$u\in C^m(T^m)$に対して各点一様収束が成り立つので従属確率変数列$\{\xi_j, j\geq 1\}$に対し、$t=(\xi_{i_1},\xi_{i_2},\cdots,\xi_{i_m})$と置いたとき、確率１で

$$S_N(u,t)=\sum_{k_1=-N}^{N}\cdots\sum_{k_m=-N}^{N}\hat{u}(k)e_{k_1}(\xi_{i_1})e_{k_2}(\xi_{i_2})\cdots e_{k_m}(\xi_{i_m})\to u(\xi_{i_1},\xi_{i_2},\cdots,\xi_{i_m})\quad(N\to\infty)$$

が成り立つ. $\mathbf{x}\in R^\infty$に対して$\varphi(\mathbf{x}):=\sum_{-\infty\leq k_1,\cdots,k_m<\infty}\hat{u}(k)x_{k_1}\cdots x_{k_m}$とおく. また$R^\infty$値確率変数$\mathbf{G}_k:=(e_1(\xi_k),e_2(\xi_k),\cdots)$を用いて

$$n^mV_n:=\sum_{1\leq i_1\leq i_2\leq\cdots\leq i_m\leq n}u(\xi_{i_1},\xi_{i_2},\cdots,\xi_{i_m})$$

$$=\sum_{k_1=-\infty}^{\infty}\cdots\sum_{k_m=-\infty}^{\infty}\hat{u}(k)\left\{\sum_{j=1}^n e_{k_1}(\xi_j)\sum_{j=1}^n e_{k_2}(\xi_j)\cdots\sum_{j=1}^n e_{k_m}(\xi_j)\right\}=\varphi\left(\sum_{k=1}^n\mathbf{G}_k\right).$$

ゆえにn^mV_nが$\{\mathbf{G}_k, k\geq 1\}$の部分和の汎関数として表すことが出来るので、$\{\mathbf{G}_k, k\geq 1\}$に関する極限定理を応用することでV-統計量$V_n$の漸近的な性質を調べることが出来る。さらに$|U_n-V_n|$を評価することから$U_n$の漸近的な性質を調べる.

U-統計量の線形結合の Edgeworth 展開

鹿児島大学理学部　大和　元
都城工業高等専門学校　野町俊文
鹿児島大学理工学研究科　戸田光一郎

1　序

次数 $k \geq 2$ の対称なカーネル $g(x_1,\ldots,x_k)$ を持つ分布 F の推定可能な母数 $\theta = \theta(F)$ の、分布 F からの任意標本 $X_1,\cdots,X_n$ に基づく推定量として、U-統計量の線形結合 Y_n(Toda and Yamato (2001)) を考える：$w(r_1,\ldots,r_j;k)$ を整数 $j = 1,\ldots,k$ に対して $r_1+\cdots+r_j = k$ を満たす任意の正の整数の組 $(r_1,\ldots,r_j)$ について非負の値を取を取る対称な関数とする。$j = 1,\ldots,k$ に対して、$d(k,j) \neq 0$ のとき、

$$g_{(j)}(x_1,\ldots,x_j) = \frac{1}{d(k,j)}\sum\nolimits^{+}_{r_1+\cdots+r_j=k} w(r_1,\ldots,r_j;k)g(\underbrace{x_1,\ldots,x_1}_{r_1個},\cdots,\underbrace{x_j,\ldots,x_j}_{r_j個})$$

とする。ただし、 $j = 1,\ldots,k$ に対して、$d(k,j) = \sum^{+}_{r_1+\cdots+r_j=k} w(r_1,\ldots,r_j;k)$, とする。また、記号 $\sum^{+}_{r_1+\cdots+r_j=k}$ は、$r_1+\cdots+r_j = k$ を満たす全ての正の整数の組に対する和を表す。$U_n^{(j)}$ は、カーネル $g_{(j)}$ に対応する U-統計量とする。

$$Y_n = \frac{1}{D(n,k)}\sum_{j=1}^{k} d(k,j)\begin{pmatrix} n \\ j \end{pmatrix} U_n^{(j)}$$

とする。ただし、 $D(n,k) = \sum_{j=1}^{k} d(k,j)\begin{pmatrix} n \\ j \end{pmatrix}$ とする。

ところで、カーネル $g_{(j)}(x_1,\ldots,x_j)$ は非退化とする。$j = 1,\ldots,k$ に対して、$\theta_j = E[g_{(j)}(X_1,\ldots,X_j)]$, $c = 1,\ldots,j$ に対して

$$\psi_{(j),c}(x_1,\ldots,x_c) = E\Big[g_{(j)}(X_1,\ldots,X_j)\Big|X_1 = x_1,\ldots,X_c = x_c\Big]$$

と置く。さらに、$j = 1,\ldots,k$ に対して、$g_{(j)}^{(1)}(x_1) = \psi_{(j),1}(x_1) - \theta_j$ と置き、$c = 2,3,\ldots,k$ に対して、

$$g_{(j)}^{(c)}(x_1,\ldots,x_c) = \psi_{(j),c}(x_1,\ldots,x_c) - \sum_{i=1}^{c-1}\sum_{1\leq l_1<\cdots<l_i\leq c} g_{(j)}^{(i)}(x_{l_1},\ldots,x_{l_i}) - \theta_j$$

と置く。 カーネル $g_{(j)}^{(c)}$ に対応する U-統計量を $H_{(j),n}^{(c)}$ で表す。$g_{(k)} = g$ より、 $\theta_k = \theta$ である。また、$\psi_{(k),c}$ は ψ_c と略記し、$H_{(k),n}^{(c)}$ も $H_n^{(c)}$ と略記する。

2　エッジワース展開

標準化 Y-統計量の $1/n$ の項までのエッジワース展開が得られる。

定理 1 $d(k,k) > 0$, $\sigma_1^2 > 0$, $1 \leq j_1 \leq \cdots \leq j_k \leq k$ に対して

$$E[|g(X_{j_1},\ldots,X_{j_k})|^2] < \infty, E|\psi_1(X_1)|^4 < \infty, E|\psi_2(X_1,X_1)|^3 < \infty,$$

$$E|\psi_2(X_1,X_2)|^4<\infty, E|\psi_3(X_1,X_2,X_2)|^3<\infty, E|\psi_3(X_1,X_2,X_3)|^4<\infty$$

$$\lim_{|t|\to\infty}\sup\left|E\left[e^{it\psi_1(X_1)}\right]\right|<1$$

を仮定する。さらに、Lai and Wang(1993) の Condition (C) または (D) を仮定する。そのとき、

$$\sup_{-\infty<z<\infty}\left|P\left(\frac{\sqrt{n}}{k\sigma_1}(Y_n-\theta)\le z\right)-G_n^*(z)\right|=o\left(\frac{1}{n}\right)$$

が成り立つ。ただし、

$$\begin{aligned}
G_n^*(z) &= \Phi(z)-\frac{1}{\sqrt{n}}\phi(z)P_1^*(z)-\frac{1}{n}\phi(z)P_2^*(z),\\
P_1^*(z) &= \frac{\mu_k}{\sigma}+P_1(z)=\frac{\mu_k}{\sigma}+\frac{\kappa_3}{6\sigma^3}(z^2-1),\\
P_2^*(z) &= P_{21}^*(z)+P_{22}(z),\\
P_{21}^*(z) &= \frac{\mu_k^2}{2\sigma^2}z+\frac{\mu_k\kappa_3}{6\sigma^4}(z^3-3z)+P_{21}(z),\\
P_2^*(z) &= \left\{\frac{\mu_k^2}{2}+a'+\frac{1}{4}k^2(k-1)^2[\sigma_2^2-2\sigma_1^2]\right\}\frac{1}{\sigma^2}z\\
&\quad+\left[\frac{\mu_k\kappa_3}{6}+\frac{\kappa_4}{24}\right]\frac{1}{\sigma^4}(z^3-3z)+\frac{\kappa_3^2}{72\sigma^6}(z^5-10z^3+15z),
\end{aligned}$$

$\Phi(z)$ は標準正規分布の分布関数を表し、$\phi(z)$ は標準正規分布の密度関数を表す。

ところで、$\sqrt{n}Y_n$ のジャックナイフ分散推定量を $\hat{\sigma}_n^2=(n-1)\sum_{i=1}^n(Y_n^{(i)}-Y_n)^2$ と置く。ただし、$Y_n^{(i)}$ は $i=1,2,\ldots,n$ にたいして、X_i を取り除いた大きさ $n-1$ の標本から計算した Y-統計量とする。ステューデント化 Y-統計量の $1/\sqrt{n}$ の項までのエッジワース展開が得られる。

命題 2　$\epsilon>0$, $\sigma_1^2>0$ に対して、

$$E|g(X_1,\ldots,X_k)|^{4+\epsilon}<\infty,\ E|g_{(k-1)}(X_1,\ldots,X_{k-1})|^{4+\epsilon}<\infty,$$

を仮定し、$j=1,2,\ldots,k-2$ に対して、$E|g_{(j)}(X_1,\ldots,X_j)|^2<\infty$　さらに、

$$\lim_{|t|\to\infty}\sup\left|E\left[\exp\{itg^{(1)}(X_1)\}\right]\right|<1$$

を仮定する。そのとき、

$$\sup_{-\infty<z<\infty}\left|P\left(\frac{\sqrt{n}}{\hat{\sigma}_1}(Y_n-\theta)\le z\right)-H_n^*(z)\right|=o\left(n^{-\frac{1}{2}}\right)$$

を得る。ただし、

$$H_n^*(z)=H_n(z)-\frac{\mu_k}{\sqrt{n}\sigma}\phi(z)=\Phi(z)+\frac{1}{\sqrt{n}}\left(-\frac{\mu_k}{\sigma}+\frac{1}{6\sigma^3}[a_3z^2+\kappa_3(z^2+1)]\right)\phi(z).$$

参考文献

Lai, T. L. and Wang, J. Q. (1993). Edgeworth Expansions for symmetric statistics with applications to bootstrap methods, *Statistica Sinica*, **3**, 517–542.

Toda, K. and Yamato, H. (2001). Berry-Esseen bounds for some statistics including LB-statistic and V-statistic. *J. Japan Statist. Soc.* **31**, No. 2, 225–237.

Finite exchangeability and a simple covariance structure

Keio University and Takachiho University
Kunihiro Baba, Ritei Shibata and Masaaki Sibuya

1. Introduction

Multinomial distributions have the variance-covariance matrix

$$\Lambda = n(\mathrm{diag}(\boldsymbol{\xi}) - \boldsymbol{\xi}\boldsymbol{\xi}^T) = \mathrm{diag}(\boldsymbol{\mu}) - n^{-1}\,\boldsymbol{\mu}\boldsymbol{\mu}^T, \tag{1}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ are the probability and mean vectors of the components, respectively. There are some other distributions with a similar variance-covariance matrix. Hence, reasons leading to this form, and the generality of the reasons are questioned. This paper is to answer to these questions based on the notion of finite exchangeability. Because of the various aspects of this notion, the answer is not complete, and examples are provided to show situations.

Baba et al. (2002) found a sufficient condition for the partial correlation to be equal to the conditional correlation, and found that some distributions with a simple covariance structure, including the multinomial, satisfy the condition. Conclusions of this paper are (i) there are many distributions with a simple covariance structure, and (ii) most of them satisfy the condition of Proposition 1, but some distributions do not.

The equality of the partial correlation of $X_L = (X_1, X_2)$ and the conditional correlation of X_L given $X_{L^c} = (X_3, \ldots, X_m)$ holds under a general condition on the conditional moments.

Proposition 1 (Baba et al., 2002). *If*

$$\mathrm{E}(X_L|X_{L^c}) = \boldsymbol{a} + BX_{L^c}$$

for a constant vector $\boldsymbol{a}$ *and a constant matrix* B, *and if the conditional correlation* $\mathrm{Cor}(X_L|X_{L^c})$ *is independent of* X_{L^c}, *then the partial correlation* $\mathrm{pCor}(X_L; X_{L^c})$ *is equal to* $\mathrm{Cor}(X_L|X_{L^c})$.

2. Main results

For a fixed integer $n > 1$, an n-variate rv $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said finite exchangeable if its distribution is invariant with respect to the permutation of the components.

Proposition 2. **(a)** *Assume that an* n*-variate rv* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *is exchangeable, and*

$$\mu := \mathrm{E}(X_i), \quad \nu := \mathrm{Cov}(X_i, X_j) \quad \textit{and} \quad \sigma^2 := \mathrm{V}(X_i),$$

are finite. Partition the index set $\{1, \ldots, n\}$ *into* m *parts* L_j, $|L_j| = l_j$, $\sum_{j=1}^m l_j = n$, *and define* $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$, $Y_j = \sum_{j=1}^n \mathrm{I}[j \in L_j]\, X_j$, *and*

$$\mathrm{V}(\boldsymbol{Y}) = (\sigma^2 - \nu)\mathrm{diag}(l) + \nu\, ll^T. \tag{2}$$

(b) *If* $\mu \neq 0$,

$$\mathrm{V}(\boldsymbol{Y}) = \frac{\sigma^2 - \nu}{\mu}\mathrm{diag}(\boldsymbol{\mu}) + \frac{\nu}{\mu^2}\,\boldsymbol{\mu}\boldsymbol{\mu}^T, \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_m),\ \mu_j := \mu l_j. \tag{3}$$

(c) *If* $\sum_{i=1}^n X_i$ *is a constant, say* t,

$$\mathrm{V}(\boldsymbol{Y}) = (n-1)^{-1} n^2 \sigma^2 (\mathrm{diag}(\boldsymbol{\xi}) - \boldsymbol{\xi}\boldsymbol{\xi}^T), \quad \boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)^T,\ \xi_j = l_j/n = \mu_j/t, \tag{4}$$

where the last expression assumes $t \neq 0$. *Further, the condition of Proposition 1 is satisfied.*

Table 1: iid variables and partially exchangeable multivariate distributions

X iid	Y partially exchangeable
Poisson	Multinomial
Binomial, Bernoulli	Multivariate hypergeometric
Negative binomial, Geometric	Multivariate negative hypergeometric
Gamma, Exponential	Dirichlet

Table 2: Exchangeability of the mixture distributions

The mixed distributions	The mixing distributuins	The mixture distiributions
Poisson (intensity parameter)	Gamma	Negative multinomial
Binomial, Bernoulli (probabilities)	Beta	Multivariate beta-binomial
Multivariate hypergeometric (size parameter)	Negative binomial, geometric	equivalent to the above
Negative binomial geometric (probabilities)	Beta	Multivariate beta-negative binomial
Multivariate negative hypergeometric (size parameter)	GHgB3	equivalent to the above
Gamma, Exponential (scale parameter)	Gamma	Multivariate beta type 2 (Multivariate F), Multivariate Pareto

A condition of the finite exchangeability is the partial exchangeability. If $(X_1, \ldots, X_n)$ is a random sample from a univariate natural exponential family $T = \sum_{j=1}^{n} X_j$ is a sufficient statistics. Then the conditional distribution given $T = t$ is exchangeable for any $n > 1$, and satisfies the condition (c) of Proposition 2. Examples are shown in Table 1.

Another condition of the finite exchangeability is a mixture of i.i.d. random variables by mixing some parameter. Some examples are shown in Table 2, and these satisfy the condition of Proposition 1. However, a unified justification is an open problem.

3. Multivariate power series distributions

Generalizing a univariate power series distribution

$$p(x)\,\delta(\{x\}) = g(\theta)^{-1}\,a_x\,\theta^x;\ a_x = (d/d\theta)^x g(0)/x!;\ x = 0, 1, \ldots;\ \theta \in \Theta, \tag{5}$$

its multivariate version is defined

$$p(\boldsymbol{x};\boldsymbol{\theta}) \prod_{j=1}^{m} \delta(\{x_j\}) = g(\theta)^{-1}\,a_x\,x! \prod_{j=1}^{m} \binom{\theta_j}{x_j}, \quad \boldsymbol{x} \in \mathcal{N}_0^m,\ \theta,\ \theta_j \in \Theta,\ j = 1, \ldots, m,\ m = 2, 3, \ldots \tag{6}$$

Proposition 3. *For the distributions* (6) *with* $a_0 = 0$, *the conditional moments* $\mathrm{E}(x_j | x_3, \ldots, x_m)$, $j = 1, 2$, *and* $\mathrm{V}((x_1, x_2) | x_3, \ldots, x_m)$ *have usually gaps between the conditioning value* $x_* = \sum_{j=3}^{m} x_j = 0$ *and* $x_* > 0$. *Hence the conditions of Proposition 1 are not satisfied.*

References

Baba, K, Shibata, R and Sibuya, M. (2002). Implications of zero partial covariance of non-normal distributions, *Proc. Joint Meeting of Statist. Soc.*, Hino, (in Japanese).