分割表の一致率検定への有限群論と組合せ論の応用

吉田 知行 (北大・理)
yoshidat@math.sci.hokudai.ac.jp

2007/08/17 神戸

1 はじめに—代数統計学の勃興

今世紀に入り,「代数統計学」あるいは「計算代数統計学」なる分野が統計学で急速に勢力を拡大しつつある．とくに，グレブナー基底による分割表の列挙問題の解決と分割表検定への実用化が画期となった．これについては，日比『グレブナー基底の現在』([日比 06]) の中の

第３章 統計学におけるグレブナー基底 (竹村彰通, 青木敏)
第４章 高次元配列データ解析とグレブナー基底 (坂田年男)

の章に詳しく説明されている．

他方，系統分類学や遺伝学などへの数学の応用も始まっている．そもそもこれら分野と数学の相性は悪くないと思うのだが，数学からの積極的関与は乏しかった．例えば，生物の系統樹 (葉にラベルの付いた木) 作成のアルゴリズムとして，現在もっともポピュラーな「近隣結合法」は，分子生物学の斎藤成也や根井正利が，基本原理を考え出し系統分類に応用してきた．方法は完全に組合せ論の言葉で書き表せるのだが，数学者の貢献はあまりなかった．しかし，今世紀に入って「代数的生物学」のでも言うべき分野が生まれつつあるように思う．Strumfels たちが書いた本 ([Sturmfels 05]) は計算生物学への代数統計などの応用を目指すもので，統計，計算，代数，生物，の４つをテーマとしている．

さらに時代をさかのぼると，比較言語学における数理的方法がある．Polya は,『発見的推論２』([Polya 59]) で，数詞を比較することによって，ヨーロッパの１０の言語の近さを測り，二項検定法でその近さが偶然で得られるかどうかを調べている．さらに Oswalt のシフト法，安本による検定法の改良や比較言語学への応用 ([安本 83, 95]) がある．統計学的には，並べ替え検定 (あるいはブートストラップ法) による分割表の一致率検定であるが，これらも群論・組合せ論の言葉で書くことが出来る．

これらの分野 (統計，生物，言語) は一見ばらばらで無関係に見える．しかし「有限群上のランダムウォーク」の観点から共通して扱えそうなテーマがいくつかある．この講演では，とくに，一致率検定を中心とする分割表の検定とその応用に，有限群とその表現論がどう使われるかを紹介したい．なお統計の予備知識は想定していない．

2 現代統計学の方法

2.1 分割表による検定

もっとも簡単な例として分割表による独立性の検定を取り上げる．今 43 名の学生に数学と国語の試験を行い，それぞれの科目の合否を決めた．結果は分割表にまとめる：

	数学合	数学否	計
国語合	14	8	22
国語否	4	17	21
計	18	25	43

この表は例えば，数学と国語両方の合格者が 14 人おり，数学合格だが国語不合格が 4 人いることを意味する．また周辺度数を見れば，数学合格者が 18 人，国語合格者が 22 人いることなども読み取れる．

数学の合否と国語の合否は関連があるのだろうか．このような独立性の検定のために，次のような**帰無仮説**を立てる．

(H_0) 数学の合否と国語の合否は関係がない．

この仮定のもとで，数学と国語両方の合格者人数は $22 \times 18/43 \fallingdotseq 9.2$ に近いはずである．実際の 14 人という数字は確かに 9.2 より大きいのだが，偶然 14 になる可能性もあり得る．そのような偶然の確率を計算するのがカイ二乗統計量である：

$$\chi^2 = \frac{43 \times (14 \times 17 - 8 \times 4)^2}{18 \times 25 \times 22 \times 21} \fallingdotseq 8.7770$$

帰無仮説のもとでは $\chi_0^2 = 0$ となるはずである．自由度 1 のカイ二乗分布の表で $\chi^2 \geq 8.7770$ なる確率を求めると，$P_\chi = 0.0007278 < 0.001$ となり，0.1% 水準の危険率で，帰無仮説は棄却される．この P_χ を**有意確率**とか ***p*-値**という．要するに数学の成績と国語の成績は独立でない，ただし千回に一度くらいはこの判断が間違っている可能性がある．

この方法は典型的な統計的仮説検定の例である．しかし問題もある．確かに，帰無仮説 (H_0) のもとで，χ^2 は「漸近的」にカイ二乗分布をする．しかし今の場合データ数が 43 と大きくない．おまけにセルのひとつが 4 とこれも小さい．このような場合カイ二乗分布で近似できる保証はない．実際，Yates の補正によると有意確率は $P_Y = 0.004954834$ となって，先ほどの P_χ とは一桁も違う．さらに後述の「Fisher の正確確率法」によれば，$P_e = 0.005089$ となるが，これは P_Y にかなり近い．結論としてどの方法でも有意確率 $P < 0.01$ なので，**危険率** 1% で，数学と国語の成績は関係ありとなる．

分割表による独立性の検定で本当に求めたいのは，正確な有意確率である．とくに医学や薬学といった人の生死に関わる分野では，たとえ正確な有意確率が求まらなくてもその精度・信頼性がつねに重要である．Agresti の本や [Agresti 92], [富澤 06] 参照．

2.2 分割表の列挙問題

一般に，$I \times J$ 型分割表とは非負整数行列 $\boldsymbol{x} = (x_{ij})$ のことである．

$$x_{+j} := \sum_i x_{ij}, x_{i+} := \sum_j x_{ij}$$

を**周辺和**という．

$$n := \sum_{ij} x_{ij} = \sum_i x_{i+} = \sum_j x_{+j}$$

をサイズという．与えられた周辺和 $\boldsymbol{a} = (a_i), \boldsymbol{b} = (b_j)$ を持つ分割表の集合を $\mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ で表す．$\mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ の中で，$\boldsymbol{x} = (x_{ij})$ の生起確率は多項超幾何分布

$$H(\boldsymbol{x}) := \frac{\boldsymbol{a}!\boldsymbol{b}!}{n!\boldsymbol{x}!} = \frac{\prod_i a_i! \prod_j b_j!}{n! \prod_{ij} x_{ij}!}$$

で与えられるとするのが自然である．実際，(A_i) と (B_j) を $N = \{1, \cdots, n\}$ の $\boldsymbol{a}$ 型と $\boldsymbol{b}$ 型の分割としたとき，$H(\boldsymbol{x})$ は $|A_i \cap B_j| = x_{ij} (\forall i, j)$ となる確率である．

さて独立性の検定で，P-値 (有意確率) の計算式は次で与えられる：

$$P(\chi_0^2) := \mathrm{Prob}(\chi^2(\boldsymbol{x}) \geq \chi_0^2) = \sum_{\chi^2(\boldsymbol{x}) \geq \chi_0^2} H(\boldsymbol{x})$$

ここで χ_0^2 は観測から得られた分割表 $\boldsymbol{x}_0$ のカイ二乗統計量である．したがって，$\mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ に属するすべての分割表が列挙できれば，正確な P-値を計算できる (**Fisher の正確確率法**)．しかしこの方法もちょっと $|I|, |J|$ が大きくなると破綻する．一般に $\mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ は巨大な集合であり，列挙問題は NP 問題である．

2.3 リサンプリング法

計算機時代の統計学を象徴するのがいわゆるモンテカルロ法とリサンプリング法である．

(1) **ブートストラップ法** ([Good 94], [Hinkley 97], [汪 03])．例として，母集団の平均とその値の信頼

度を求めることを考える．母集団から取った標本 $x_1, x_2, \cdots, x_n$ が与えられているとし，そこから繰り返しを許してランダムに n 個取る：$x_1^*, x_2^*, \cdots, x_n^*$. それらの平均値を求める：$m^* = \frac{1}{n}(x_1^* + \cdots + x_n^*)$. この操作を多数回行えば，多数の「平均値」が得られるそれら「平均値」の平均を m とする．この m を母集団平均の推定値とする．$\epsilon = 0.01$ などとして，

$$P = (|m - m^*| \geq \epsilon \text{ となった「平均値」} m^* \text{ の割合})$$

が信頼度 (0 に近いほど信頼性が高い) である．このように与えられた標本から新たな標本を作り出す操作がリサンプリング法である．とくに非復元抽出を使うのをブートストラップ法という．

(2) **マルコフ鎖モンテカルロ法 (MCMC 法)** ([伊庭 05])．先ほど述べたように，分割表すべての列挙が不可能でフィッシャーの正確確率法が使えない場合には，モンテカルロ法が有効である．先ほどのカイ二乗検定の場合は，多項超幾何分布にしたがう分割表 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m \in \mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ を大量に発生させ

$$P = (\chi^2(\boldsymbol{x}_k) \geq \chi_0^2 \text{ となる割合})$$

を有意確率の近似値とする．

分割表大量発生のためには，このような静的モンテカルロ法よりも，MCMC 法が有効である．$\mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ に属するサイズ n の分割表を大量に発生させるために，ランダムに $i < j$ と $k < l$ の対を選び，i, j 行と k, l 列に $\begin{array}{|cc|}\hline +1 & -1 \\ -1 & +1 \\ \hline \end{array}$ を加える．ただし (i, l) 成分か (j, k) 成分どちらかが 0 なら何もしない．この操作をくり返して多数の分轄表が得られる．また収束の様子を見たり，理論的に取り扱うのにも MCMC 法が適している．

(3) **フィッシャーの並べ替え検定** ([Mielke 01], [Good 94])．観測データ $x_1, x_2, \cdots, x_n$ を並べ替えて新たな分割表を作る．このアイデアは相当古い (1935). 当時の貧弱な計算機では実用にならなかったが，この 10 年ほどで大きな発展を見ている．並べ替え検定の例 (比較言語学) はあとで紹介する．

2.4 統計学ににどのような代数が使われてきたか

両者はかなり離れた分野であり，交流も少なかったと思う．統計関係の分野から，代数的組合せ論の分野に入ってきたものとして，ブロックデザインとアソシエーションスキームはよく知られている．

代数系の研究者はかえって知らないようだが，統計学の最先端では，いろいろな代数が使われている．

(1) 多変量解析への線形代数や連続群の表現などの応用 ([Muirhead 82])．多変量解析とは，ベクトル値のデータの統計的解析法で，理論的には線形代数 (と多変数微分積分) の塊である．Zonal 多項式の応用もある ([竹村 84], [Muirhead 82], [水川 04]).

(2) 有限群上の確率過程への有限群の表現論の応用 ([Diaconis 88], [Saloff 03])．線形群上の確率測度 ([Heyer 94]).

(3) グレブナー基底，代数幾何の応用 ([伊庭 05], [日比 06], [青木 07]).

とくに近年のグレブナー基底の応用は著しい．もちろん統計でよく使われる多項式計算のアルゴリズムの中にグレブナー基底が使われているのだが，近年の応用はグレブナー基底が，統計やその関連分野 (分割表の列挙問題，実験計画法，線形計画法など) に生の形で使われている．現代の統計学でもっとも熱い分野である．

3 分割表と対称群

ここからは筆者の関わった分野を述べる．与えられた周辺和を持つ分割表を大量かつランダムに構成するために，対称群上のランダムウォークを利用する．この考えは，Fisher の並べ替え検定の流儀をひくが，きっかけになったのは，後で述べるように比較言語学のシフト検定法である．まず分割表の元になるデータセットの概念から始める．

3.1 データセット

以下，$N = \{1, 2, \cdots, n\}$ としておく．統計的には，N は実験・観測や個人の集合を表す．S_n は対称群とする．

I 型の 1 次元**データセット** $[f]$ とは，写像 $f : N \to I$ のことである．$I \times J$ 型の 2 次元データセットとは，ふたつの写像 $f : N \to I$ と $g : N \to J$ の対 $[f, g]$ のことである．単なる写像と区別するために，$[f]$ とか $[f, g]$ と表す．3 次元以上のデータセットの定義も同様になされる．

データセット $[f : N \to I]$ の**度数分布表** $\mathsf{tab}[f]$ とは，ベクトル $(|f^{-1}(i)|)_{i\in I}$ のことである．同じ度数分布表を持つデータセット $[f]$ と $[f']$ は同型であるといい，$[f] \cong [f']$ と書く．$I \times J$ 型の 2 次元データセット $[f, g]$ の**分割表** $\mathsf{tab}[f, g]$ は

$$\mathsf{tab}[f, g] = \left(|f^{-1}(i) \cap g^{-1}(j)|\right)_{i\in I, j\in J}$$

で定義される．この分割表の**周辺和** (または**周辺分布**) は，ふたつの度数分布表 $\mathsf{tab}[f]$ と $\mathsf{tab}[g]$ で与えられる．

$\mathsf{DS}(\boldsymbol{a})$ を，与えられた度数分布表 $\boldsymbol{a} = (a_i)$ を持つ I 型のデータセットの集合とする．$\mathsf{DS}(\boldsymbol{a}, \boldsymbol{b})$ を与えられた周辺和 $\boldsymbol{a} = (a_i), \boldsymbol{b} = (b_j)$ を持つ $I \times J$ 型のデータセットの集合とする:

$$\mathsf{DS}(\boldsymbol{a}) := \{[f] \mid \mathsf{tab}[f] = \boldsymbol{a}\},$$
$$\mathsf{DS}(\boldsymbol{a}, \boldsymbol{b}) := \{[f, g] \mid \mathsf{tab}[f] = \boldsymbol{a}, \mathsf{tab}[g] = \boldsymbol{b}\}.$$

したがって，写像

$$\mathsf{tab} : \mathsf{DS}(\boldsymbol{a}, \boldsymbol{b}) \longrightarrow \mathsf{TAB}\ (\boldsymbol{a}, \boldsymbol{b})$$

を得る．

対称群 S_n の $\mathsf{DS}(\boldsymbol{a})$ への (右からの) 作用を $[f] \cdot \pi := [f\pi]$ で定義する．同様に $S_n \times S_n$ の $\mathsf{DS}(\boldsymbol{a}, \boldsymbol{b})$ への作用を $[f, g](\sigma, \tau) := [f\sigma, g\tau]$ で定義する．

補題. (1) S_n の $\mathsf{DS}(\boldsymbol{a})$ への作用は可移である．一点 $[f]$ の固定部分群 G_f は分割 $N = \sum_{i\in I} f^{-1}(i)$ に対応する Young 部分群である．したがって

$$|\mathsf{DS}(\boldsymbol{a})| = n!\,/\boldsymbol{a}! = n!\,\Big/\prod_i a_i!$$

(2) $S_n \times S_n$ の $\mathsf{DS}(\boldsymbol{a}, \boldsymbol{b})$ への作用は可移である．$\mathsf{tab}[f, g] = \mathsf{tab}[f', g']$ であるための必要十分条件は，ある $\pi \in S_n$ があって，$f' = f\pi, g' = g\pi$ となることである．

(3) tab は全射である．

系. $[f_0, g_0] \in \mathsf{DS}(\boldsymbol{a}, \boldsymbol{b})$ とする．

$$\mathsf{tab}' : S_n \longrightarrow \mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b}); \pi \longmapsto \mathsf{tab}[f_0, g_0\pi]$$

は全射で，各 $\boldsymbol{x} \in \mathsf{TAB}\ (\boldsymbol{a}, \boldsymbol{b})$ に対し

$$\frac{\sharp\{\pi \in S_n \mid \mathsf{tab}'(\pi) = \boldsymbol{x}\}}{n!} = \frac{\boldsymbol{a}!\boldsymbol{b}!}{n!\boldsymbol{x}!} = H(\boldsymbol{x})$$

とくに，tab' は，一様分布に収束する S_n 上のマルコフ鎖を，多項超幾何分布に収束する $\mathsf{TAB}\ (\boldsymbol{a}, \boldsymbol{b})$ 上のマルコフ鎖に写す.

3.2 有限群上のランダムウォーク

有限群 G 上のランダムウォークを考える．[Diaconis 88], [Gluck 97], [Heyer 94], 共役に関して不変な (つまり類関数であるような) G 上の確率測度 μ を取る：任意の $a, x \in G$ に対し，

$$0 \le \mu(x) \le 1, \sum_{x\in G} \mu(x) = 1, \mu(axa^{-1}) = \mu(x).$$

G の単位元から始め，G の元を次々と右から乗じて行くことで G 上のランダムウォーク (以下 RW) $1 = x_0 \to x_1 \to \cdots \to x_n \to \cdots$ が得られる．ただし G の元は μ に従った確率で選ぶ．遷移確率 (x が 1 回で y に移る確率) $P(x, y) = \mu(x^{-1}y)$ である．任意の $a, x, y \in G$ に対し，

$$0 \le P(x, y) \le 1, \sum_{y\in G} P(x, y) = 1,$$
$$P(ax, ay) = P(x, y), P(xa, ya) = P(x, y).$$

元 x が k ステップで y に移る確率は行列のベキ P^k で与えられる．確率測度 μ を用いて表すことも出来る．

$$M := \sum_{x\in G} \mu(x)x \in \mathbb{C}G$$

と置く ($\mathbb{C}G$ は複素数係数群環)．このとき $P^k(x,y)$ は M^k における $x^{-1}y$ の係数である．

μ が類関数なので，M は中心 $Z(\mathbb{C}G) \cong \mathbb{C}^r$ の元である．M をフーリエ展開する：

$$M = \sum_{\chi} \omega_\chi(M) e_\chi.$$

ここで χ は G の既約指標を動き，さらに

$$e_\chi := \frac{\chi(1)}{|G|} \sum_{g\in G} \chi(g^{-1})g \quad (\text{ベキ等元})$$

$$\omega_\chi(M) := \sum_{x\in G} \mu(x) \frac{\chi(x)}{\chi(1)}.$$

である．$\omega_\chi : Z(\mathbb{C}G) \to \mathbb{C}$ は多元環準同型である．これより

$$M^k = \sum_{\chi} \omega_\chi(M)^k e_\chi.$$

したがって $\{P^k\}$ の収束は $\omega_\chi(M)$ に依存する．

$$|\omega_\chi(M)| \le \sum_{x\in G} \mu(x) \left|\frac{\chi(x)}{\chi(1)}\right| \le \sum_{x\in G} \mu(x) = 1$$

に注意しておく．E を μ のサポートとする．このとき $k \to \infty$ での M^k の挙動は次のようになる：

$$\begin{aligned} M^k &= (\text{一様分布成分}) + (\text{振動成分の和}) \\ &\qquad +(\text{消滅成分の和}) \\ \text{一様分布成分} &= e_1 = \frac{1}{|G|}\sum_{g\in G} g, \\ \text{振動成分} &= (|\omega_\chi(M)| = 1, \chi \ne 1_G \text{なる成分}), \\ \text{消滅成分} &= (|\omega_\chi(M)| < 1 \text{なる成分}). \end{aligned}$$

一様分布成分に寄与するのは単位指標 1_G だけであり，既約指標 $\chi \ne 1_G$ が振動成分に寄与するための必要十分条件は，$E^{-1}E \subseteq \mathrm{Ker}(\chi)$ である．さらに

$$\rho := \mathrm{Max}\left\{\left|\frac{\chi(x)}{\chi(1)}\right| : x \in E, E^{-1}E \not\subseteq \mathrm{Ker}\chi\right\}$$

で**収束率**を定義すれば，(消滅部分) $\sim c\rho^k$ と評価される．とくに次を得る：

定理．確率過程 P^k が一様分布に収束するための必要十分条件は，$\langle E^{-1}E\rangle = G$．とくに E がひとつの共役類の場合，この条件は，$[t,G] = G$ (t は C の代表元) となる．

この種の定理は Diaconis と M. Shahshahani の 1981 年の論文にあるが，他にも何人かの人が独立に類似した結果に到達していたようだ．

3.3 計算例．

$G = S_n$ で，確率測度 μ の台 C_l が長さ l の巡回置換からなるとする．S_n は完全でないので，μ から得られる確率過程は一様分布に収束しない．l が 偶数の場合，この確率過程は既約 (C_l が G を生成する) だが周期が 2(偶置換と奇置換が入れ替わる) である．l が奇数の場合は，既約でない (奇置換には到達しない) が周期は 1 である．t_l を長さ l の巡回置換とする．このとき定常部分と振動部分の和は，

$$e_1 + (-1)^{l-1} e_{\mathrm{sgn}}$$

であり，収束部分の収束率は

$$\rho = \mathrm{Max}\,\{|\chi_\lambda(t_l)/\chi_\lambda(1)| : \chi_\lambda \ne 1, \mathrm{sgn}\}$$

で与えられる．

(1) $l = 2$ (互換) の場合，$\rho = (n-3)/(n-1)$ である．$\lambda = 1^{\mu_1}2^{\mu_2}\cdots n^{\mu_n}$ その共役 $\lambda' = 1^{\mu'_1}2^{\mu'_2}\cdots n^{\mu'_n}$ に対し

$$\frac{\chi_\lambda(t_2)}{\chi_\lambda(1)} = \frac{1}{n(n-1)} \sum_i i^2(\mu_i - \mu'_i)$$

であることを用いた．この ρ は次数 $n-1$ の既約指標 ($[n-1,1]$ と $[2\,1^{n-2}]$) によって実現される．これについては [Diaconis 81].

(2) $l = n$ の場合，収束率 $\rho = (n-1)^{-1}$ である．$\chi_\lambda(t_n)$ の値は，$\lambda = [a, 1^{b-1}](a+b = n, 1 \le a \le n-1)$ のとき $(-1)^{b-1}$ で，その他のときは 0 だからである．この収束率は $\lambda = [n-1,1], [2,1^{n-2}]$ のときに実現している．

(3) $l = n-1$ の場合，収束率 $\rho = 2(n(n-3))^{-1}$ である．$\chi_\lambda(t_{n-1})$ の値は，$\lambda = [a,2,1^{b-2}](a+b = n, 2 \le$

$a \leq n-2$) のとき $(-1)^{b-1}$ で，その他のときは 0 だからである．この収束率は $\lambda = [n-2, 2], [2^2, 1^{n-4}]$ のときに実現している．実際，対称群の共役類でもっとも収束率の小さなものは，長さ $n-1$ の巡回置換の類によって得られる．

μ の台 E として，$C \cup \{1\}$ (C は共役類で奇置換からなる) とすることによって，振動部分のない確率過程が得られる．これについては [Diaconis 81], [Saloff 03].

対称群以外の群についても問題は考えられる．群として n 次の可移置換群 $G \leq S_n$ を取る．問題は G 上のランダムサンプリングを取ったとき，それから誘導される分割表のサンプリングが，どの程度ランダムかである．そのためには，もとの群上のサンプリングが一様に近い必要がある．

(1) 群として対称群 S_n を取る．長さ n のサイクルと長さ $n-1$ のサイクルをランダムに取る．このときそれらの積は $S_n - A_n$ 上に一様に分布する．これはプログラムが楽そうだ．一様性も成り立っている．

(2) 一般の群 G 場合．消滅部分の収束率 $\rho_{G,\mu}$ は，$\chi \neq 1_G$ を G の既約指標としたとき，

$$\mathrm{Max}_\chi \left| \sum_{x \in G} \mu(x)\chi(x)/\chi(1) \right|$$

(χ は和の絶対値が 1 より小さいところを取る) であった．収束率が 0 に近いほどランダムウォークの収束が速い．当然そのような μ が求められる．とくに，$E = \mathrm{supp}(\mu)$ が単一の共役類として，$\rho_{G,\mu}$ はどこまで小さくできるのだろうか．$\rho_{G,\mu} \geq 2/n(n-3)$ と予想するのは自然である．最小を与える G, μ は，対称群と長さ $n-1$ のサイクルの場合である．一般の類関数 μ の場合にも，何か具体的な下限があるかもしれない．すべての既約指標に対し $|\chi(t)| \leq 1$ であるような $t \in G$ を見つけるのが鍵になるが，有限単純群論に出てきた例外指標の理論を思い起こさせる．G が単純群で単一の共役類の場合は，

$$\min_{x \neq 1} \max_{\chi \neq 1} \left| \frac{\chi(x)}{\chi(1)} \right|$$

を求める問題となる．

3.4 課題．

(1) **分割表のランダムウォーク．** 対称群上のランダムウォークから誘導される分割表たちの上のランダムウォークの収束性は重要な問題である．また対称群上のランダムウォークの収束の速さは公式があったが，分割表の場合はもう少し単純になる．

注意． 対称群の上のランダムウォークの振動部分は分割表に影響しない．ただし，周辺和 $\boldsymbol{a}, \boldsymbol{b}$ の成分がすべて 0,1 である場合を除く．

これはそのような分割表を与えるデータセット $[f, g]$ を取ったとき，は必ずある互換 τ によって $\mathsf{tab}[f, g\tau] = \mathsf{tab}[f, g]$ となることによる．

さらに分割表上のランダムウォークの収束率については，対称群上の $(n-3)/(n-1)$ や $2/n(n-3)$ のような収束率よりも小さなものが得られるはずである．これは対称群の Young 部分群に関する Hecke 環 $\mathbb{C}[S_f \backslash S_n / S_h]$ 上のランダムウォークの収束率と同等の問題になる．当然群環 $\mathbb{C}S_n$ よりも Hecke 環の方で考えるべきである．

(2) **対称半群．** 対称半群 T_n とは，$N\{1, 2, \cdots, n\}$ からそれ自身への写像全体が合成に関してなす半群のことである．これについても，T_n や $T_n \times T_n$ の作用

$$[f] \longmapsto [f\sigma], [f, g] \longmapsto [f\sigma, g\tau] \quad (\sigma, \tau \in T_n)$$

が考えられる．ただしこの作用によって周辺和は変化する．

とくに対称半群上のランダムサンプリングから 1 次元データセットのサンプリングが得られる．T_n からランダムに元 σ を取ることは，N から n 個の元 $\sigma(1), \sigma(2), \cdots, \sigma n$ を重複を許して選ぶことと同じである．これはまさにブートストラップ法の考えである．

ここに，対称半群の表現論が役立つかもしれない．半群環 $\mathbb{C}T_n$ は半単純でないが，$\mathbb{C}T_n / J(\mathbb{C}T_n)$ は $\mathbb{C}S_1, \mathbb{C}S_2, \cdots, \mathbb{C}S_n$ の直積である．したがって，対称半群の既約表現は対称群の表現論で記述できるはずである．

(3) **その他の代数系上のランダムウォーク．**距離正則グラフ，系統樹，超群，より一般に結合的概形 (association scheme) 上のランダムウォークは実用上からも研究の価値がある．系統樹のランダムサンプリングは生物の系統樹を作るときに必要である．得られた系統樹の信頼性がしばしば問題になる．これについて新しい方法が開発できるかもしれない．系統樹とは，葉にラベルの着いた根つき (または根なし) 木のことである．$n+1$ 枚の葉を持つ系統樹には S_{2n} が可移に作用しており，Gelfand 対 $(S_{2n}, S_n \times S_n)$ に関連している．さらに超平面配置上のランダムウォークも研究されている．

(4) **差集合** (difference set)．(v,k,λ)-**差集合**とは，位数 v の群 G と位数 k の部分集合 $D \subseteq G$ で，任意の $1 \neq z \in G$ に対し，$x^{-1}y = z$ を満たす $x, y \in D$ の対の個数が (z に無関係に) ちょうど λ 個存在することをいう．$\Delta := \sum_{x\in D} x \in \mathbb{C}G$, $\Delta' := \sum_{x\in D} x^{-1} \in \mathbb{C}G$ と置けば，差集合の条件は

$$\Delta'\Delta = k + \lambda(\hat{G} - 1)$$

となる．同じことだが，$z \in G$ を与え，$x, y \in D$ をランダムに取ったとき，$x^{-1}y = z$ となる確率は，$z = 1$ のとき $1/k$ であり，$z \neq 1$ のとき λ/k^2 である．これを用いても一様分布する群の元のサンプリングが得られる．

4　分割表の一致率検定．

4.1　一致率検定．

サイズ n の $\Lambda\times\Lambda$ 型の正方分割表 $\boldsymbol{x} = (x_{\lambda,\mu})$ を考える．周辺和は $\boldsymbol{a} = (a_\lambda), \boldsymbol{b} = (b_\lambda)$ はとくに断らない限り固定しておく．その対角和 $\mathrm{Tr}(\boldsymbol{x}) = \sum_\lambda x_{\lambda,\lambda}$ を**一致数**という．比 $\mathrm{Tr}(\boldsymbol{x})/n$ を**一致率**という．

例えば，同じ競技を二人の審判が採点する場合，審判による採点の一致率で審判による採点のばらつきを計ることになる．本来なら，一致率は 1，すなわち二人の採点結果から得られる分割表は対角行列でなければならない．そうでなくとも，対角部分 (二人が同じ点数をつけた部分) に点数が集中していなければ，二人の審判の信頼性が問題になるだろう．

一致率検定として Cohen のカッパ検定がよく使われている．[富澤 06]．しかしやはり信頼度の問題が生ずる．そのため，ここでは前節までに述べてきた有限群のデータセットへの作用を使う．統計学的にかなり驚くことは，与えられた一致数が得られる確率が多項式計算で正確に得られることである．

4.2　データセットと分割表の一致数．

以下 $N = \{1, 2, \cdots, n\}$ で S_n は対称群，Λ は有限集合とする．データセット $[f, g]$ は $f, g : N \longrightarrow \Lambda$ なるものを考える．$[f, g]$ または対応する分割表 $\boldsymbol{x} = \mathsf{tab}[f, g]$ の**一致数**とは，

$$x_0[f, g] := \mathrm{Tr}(X) = \sharp\{i \in N | f(i) = g(i)\} = \mathrm{Tr}(\boldsymbol{x})$$

のことである．$[f, g]$ を固定して $x_0[f', g']$ ($[f'] \cong [f], [g'] \cong [g]$) の分布を知りたい．$x_0[f\sigma, g\tau] = x_0[f, g\tau\sigma^{-1}]$ なので，$x_0(\pi) := x_0[f, g\pi]\,(\pi \in S_n)$ の分布が分かればよい．

ここでは，より一般に，N に作用する有限群 G についての $x_0(\pi)(\pi \in G)$ の分布を調べる．とくに平均と分散は

$$\begin{aligned} m &:= \frac{1}{|G|}\sum_{\pi\in G} x_0(\pi), \\ s^2 &:= \frac{1}{|G|}\sum_{\pi\in G} (x_0(\pi) - m)^2 \end{aligned}$$

である．平均や分散などのモメントを，1 次元周辺度数分布 $a_\lambda := |f^{-1}(\lambda)|, b_\lambda := |g^{-1}(\lambda)|$ $(\lambda \in \Lambda)$ で表したい．

4.3　平均と分散．

定理． G が N に可移に作用していれば，

$$m = \frac{1}{n}\sum_{\lambda\in\Lambda} a_\lambda b_\lambda$$

この定理は,

$$\sharp\{(\pi,i,j)\in G\times N^2 \mid f(i)=g,\pi(i)=j\}$$

を二通りに数える方法で容易に証明される.

置換群 G の作用が t-**重可移**であるとは，相異なる N の元の t 組同士が移りあえることをいう：

$$\forall i_1,\cdots,i_t(\neq);\forall j_1,\cdots,j_t(\neq);$$
$$\exists\pi\in G;\pi(i_1)=j_1,\cdots,\pi(i_t)=j_t$$

例えば，巡回群 $C_n=\langle(1,2,\cdots,n)\rangle$ は (1-重) 可移，対称群 S_n は n-重可移，交代群 A_n は $(n-2)$-重可移である.

定理. G が 2 重可移なら，分散は次で与えられる：

$$s^2=\frac{1}{n-1}m(m+n)-\frac{1}{n(n-1)}\sum_\lambda a_\lambda b_\lambda(a_\lambda+b_\lambda)$$

証明は，$f(i)=g(j),f(i')=g(j'),\pi i=j,\pi i'=j'$ を満たすよつ組 $(\pi;i,i';j,j')\in G\times N^4$ を，二通りの方法で数えればよい.

確率 $p_\lambda=a_\lambda/n,q_\lambda=b_\lambda/n$，平均一致率 $p=m/n=\sum p_\lambda q_\lambda$ とすれば

$$s^2=\frac{n^2}{n-1}\left\{p(1+p)-\sum p_\lambda q_\lambda(p_\lambda+q_\lambda)\right\}$$

となるが，この式は，カッパ検定の分散公式と実質同じである．カッパ検定では，上の平均 m と分散 s^2 を持つ正規分布によって検定をしている.

4.4 高次モメント公式

定理. G が N に t-重可移に作用しているなら,

$$\frac{1}{|G|}\sum_{\pi\in G}\binom{x(\pi)}{t}=\frac{(n-t)!}{n!}\sum_{\Sigma t_\lambda=t}\prod_\lambda\binom{a_\lambda}{t_\lambda}\binom{b_\lambda}{t_\lambda}t_\lambda!$$

証明. $f^{[t]},g^{[t]}:N^{[t]}\longrightarrow\Lambda^t$ に平均公式を適用する．ここで，$N^{[t]}$ は，相異なる N の元の t-組の集まり．t-重可移性より，G は $N^{[t]}$ に可移に作用していることに注意する.

組合せモメントを使うのは，坂内英一 (九大)，榎本彦衛 (慶大) の示唆．ただし上の証明は吉田.

この公式を使えば，$t\geq 4$ の場合の歪度 $\gamma_1:=m_3/s^3$ と尖率 $\gamma_2:=m_4/s^4-3$ が求まる (『数理科学』1984)．ここで,

$$t_0:=\sum a_\lambda b_\lambda,\ t_1:=\sum a_\lambda b_\lambda(a_\lambda+b_\lambda);$$
$$m=t_0/n,\ s^2=m_2=\frac{t_0(n^2+t_0)}{n^2(n-1)}-\frac{t_1}{n(n-1)}$$

$$t_2:=\sum a_\lambda b_\lambda(a_\lambda^2+b_\lambda^2),\ t_3:=\sum a_\lambda b_\lambda(a_\lambda^3+b_\lambda^3),$$
$$t_4:=\sum a_\lambda^2 b_\lambda^2,\ t_5:=\sum a_\lambda^2 b_\lambda^2(a_\lambda+b_\lambda).$$

$$\begin{aligned}m_3:=&-\frac{2m(n^2+3mn+m^2)}{(n-1)(n-2)}+\frac{3(n+2m)}{n-2}s^2\\&+\frac{2(t_2+3t_4)}{n(n-1)(n-2)}.\\m_4:=&\frac{6m(m^3+n^3+6m^2n+6mn^2+mn)}{(n-1)(n-2)(n-3)}\\&-\frac{36m^2+60mn+n(11n-1)}{(n-2)(n-3)}s^2\\&+\frac{3n(n-1)}{(n-2)(n-3)}s^4+\frac{6(2m+n)m_3}{n-3}\\&-\frac{6(t_3+t_4+6t_5)}{n(n-1)(n-2)(n-3)}\end{aligned}$$

高次のモメントが分かれば，一致率の分布の Edgeworth 展開が分かり，したがってより正確な検定が可能になる.

4.5 一致数に対する正確な p 値.

一致数については正確な p 値を計算する実用的な方法がある．分割表のサイズ n が大きく，各セル x_{ij} が小さくない (≥ 6) なら，カッパ検定のように正規分布にもとづく検定が出来る．n が大きいが，小さなセルのある場合は Yates の補正が有効である．反対に n が小さい場合は，Fisher の正確確率法が使える．ここで説明する方法はちょうどその中間で威力を発する．今，観測されたデータセットを $[f_0,g_0:N\longrightarrow\Lambda]$ とする．周辺和を

$$\boldsymbol{a}=(a_\lambda)=(|f_0^{-1}(\lambda)),\boldsymbol{b}=(b_\mu)=(|g_0^{-1}(\mu)|)$$

であるとする．このデータセットから得られる正方分割表を $\boldsymbol{x}_0=\mathsf{tab}[f_0,g_0]$ とする．$\mathsf{TAB}(\boldsymbol{a},\boldsymbol{b})$ の元

は，$\mathsf{tab}[f_0, g_0\pi]$ $(\pi \in S_n)$ の形をしていることに注意する．$\mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b})$ に属する分割表の一致数が r 以上になる確率は

$$\begin{aligned} P(r) &:= \mathrm{Prob}(\mathrm{Tr}(\boldsymbol{x}) \geq r) = \sum_{\mathrm{Tr}(\boldsymbol{x}) \geq r} H(\boldsymbol{x}) \\ &= \mathrm{Prob}([f,g] \in \mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b}) \mid x_0[f,g] \geq r) \\ &= \frac{\sharp\{[f,g] \in \mathsf{TAB}(\boldsymbol{a}, \boldsymbol{b}) \mid x_0[f,g] \geq r)\}}{|\mathsf{TAB}\ (\boldsymbol{a}, \boldsymbol{b})|} \\ &= \frac{1}{n!}\sharp\{\pi \in S_n \mid x_0[f_0, g_0\pi] \geq r\} \end{aligned}$$

である．ここで，${}_2F_0$ 型超幾何多項式とその積を次で定義する，

$$F_{a,b}(z) = {}_2F_0(-a, -b; z) = \sum_{k \geq 0} \binom{a}{k}\binom{b}{k} k!\, z^k$$

と置き，さらに

$$F(z) := \prod_{\lambda} F_{a_\lambda, b_\lambda}(z) = \sum_{k \geq 0} \binom{n}{k} k! q(k)\, z^k$$

と展開する．$q(0) = 1, q(1) = m$．このとき P 値

$$\begin{aligned} P(r) &:= \mathrm{Prob}(x_0(\pi) \geq r) \\ &= \frac{1}{n!}\sharp\{\pi \in S_n \mid x_0(\pi) \geq r\} \end{aligned}$$

は次の定理から計算できる：

定理 (モメント公式の別形)：

$$\sum_{k \geq 1} P(k)\, z^k = 1 + z \sum_{k \geq 1} q(k)\, (z-1)^{k-1}$$

4.6 その他の方式—対称半群，準群，超群を使った一致率検定

かき混ぜる代数系として N 上の多重可移群 (とくに対称群) を用いたが，他の代数系も考えられる．

周辺和を $\boldsymbol{a}, \boldsymbol{b}$ とするデータセット $[f, g]$ を用意する．偶然による一致数として

$$x_0(\sigma, \tau) := \sharp\{i \in N \mid f(\sigma(i)) = g(\tau(i))\}, (\sigma, \tau \in T_n)$$

を考える．$(\sigma, \tau) \in T_n \times T_n$ をランダムに取ることは，重複を許して $\sigma(k), \tau(k) \in N$ $(k = 1, \cdots, n)$ をランダムに取ることである．これはブートストラップ法の考えである．$[f, g]$ と $[f\sigma, g\tau]$ の周辺分布は必ずしも等しくないが，それでも考え方は合理的である．実際 $x_0(\sigma, \tau)$ の分布は，二項分布 $B(n, m/n)$ (ここで $m = (1/n)\sum a_\lambda b_\lambda$) で与えられる．対称半群でなく対称群を取った場合，$(\sigma, \tau) \in T_n \times T_n$ をランダムに取ることは，重複を許さずにm $\sigma(k), \tau(k) \in N$ $(k = 1, \cdots, n)$ をランダムに取ることである．この検定方式は並べ替え検定で，分布は (独立でない) 多項超幾何分布の和である．

より一般化するなら，対称群や対称半群の代わりに，$N = \{1, \cdots, n\}$ に作用する集合 S (同じことだが，対称半群 T_n 上の重複集合) を用いることが考えられる．

N に作用する集合 S が**擬斉次**であるとは，$\sharp\{\sigma \in S \mid \sigma i = j\}$ が $i \neq j \in N$ によらないことをいう．これについても一致数の平均値を与える公式がある．

例． (1) S_n の共役類は擬斉次である．

(2) **ラテン方陣** (数独 の解) $S = N$ は N 上擬斉次．

(3) **超群** $G \times G \longrightarrow \mathbb{R}G; (x, y) \longmapsto \sum_z p_{xy}^z z$ $(p_{xy}^z \in [0, 1])$ の可移な作用 $G \times N \longrightarrow \mathbb{R}N$ は擬斉次．ここで，p_{xy}^z は，x と y を乗じて z になる確率を表す．**アソシエーションスキーム**も超群とほとんど同じ概念である．

注意：S が N に作用しているということは，写像 $\alpha : S \longrightarrow T_n$ の存在と同じである．したがって，各 $\sigma \in T_n$ の重みを $|\sigma| := |\alpha^{-1}(\sigma)|$ で定義すれば，T_n は T_n 上の**重複集合**となる．また，$A \subset E_n$ に対し，$\mu(A) := |\alpha^{-1}(A)|/|S|$ は E_n 上の確率測度である．このように，N に作用する集合の概念，T_n への写像の概念，T_n 上の重複集合の概念，T_n 上の確率測度の概念はきわめて近い概念と考えられる．

4.7 ３次元分割表の一致数検定

３次元データセット $[f, g, h : N \to \Lambda]$ の分割表は

$$\mathsf{tab}[f, g, h] := (|f^{-1}(\lambda) \cap g^{-1}(\mu) \cap h^{-1}(\nu)|)_{\lambda, \mu, \nu \in \Lambda}$$

で定義される．(1 次元) 周辺分布は $a_\lambda := |f^{-1}(\lambda)|, b_\mu := |g^{-1}(\mu)|, c_\nu := |h^{-1}(\nu)|$ で定義される 3 yつのベクトル $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ のことである．

一致数はいくつかの定義が可能である．

方式 1．$x_0[f,g,h] := x_0[g,h] + x_0[f,h] + x_0[f,g]$. ここで $x_0[f,g]$ などは，2 次元データセットの一致数である．$[f,g,h]$ と同じ 1 次元周辺分布を持つすべての 3 次元データセットについて，平均と分散が計算できる．

平均：$m' = m[g,h] + m[f,h] + m[f,g] = (1/n)\sum(b_\lambda c_\lambda + a_\lambda c_\lambda + a_\lambda b_\lambda)$
分散：$s'^2 = s^2[g,h] + s^2[f,h] + s^2[f,g]$．なぜこのようにきれいな公式なのだろう？

これらの計算は，与えられた 1 次元周辺和を持つデータセットが $[f\sigma, g\tau, h\rho]$ $(\sigma,\tau,\rho \in S_n)$ と表せることを使う．

方式 2．$x_0[f,g,h] := \sum_\lambda x_{\lambda\lambda\lambda} = \sharp\{a \in N \mid f(a) = g(a) = h(a)\}$.

平均 $m'' = (1/n^2)\sum a_\lambda b_\lambda c_\lambda$.

分散 $s''^2 = \dfrac{1}{n^2(n-1)^2}\left\{(2n-1)n^2 m''^2 + (n-2)n^3 m'' - \sum a_\lambda b_\lambda c_\lambda (a_\lambda + b_\lambda + c_\lambda)\right\}$

注意．(1) 一般に $\mathsf{tab}[f\sigma, g\tau, h\rho]$ の 2 次元周辺分布は $\mathsf{tab}[f,g,h]$ のものと異なる．ただし，互換 τ に対する $\mathsf{tab}[f\tau,g,h] - \mathsf{tab}[f,g\tau,h] + \mathsf{tab}[f,g,h\tau]$ は $\mathsf{tab}[f,g,h]$ と同じ 2 次元周辺分布を持つ．ただし，$\mathsf{tab}[f\tau,g,h] - \mathsf{tab}[f,g\tau,h] + \mathsf{tab}[f,g,h\tau]$ は 3 次元データセットの分割表であるどころか，成分が負になることもあり得る．3 次元分割表の**マルコフ基底**にこのような変換を与えるものがある．つまり高次元のデータセットについては,「一般的」(virtual な) もの $[f\tau,g,h] - [f,g\tau,h] + [f,g,h\tau]$ を考える必要がありそうだ．
(2) 方式 2 については，t 次モメントの公式がある．${}_3F_0$ を使った書き換えもある．
(3) 写像のファミリー $\{f_\alpha\}, \{g_\beta\}$ 同士についても公式がある．一致数 $x_0[\{f_\alpha\}, \{g_\beta\}]$ を $\sum_{\alpha,\beta} x_0[f_\alpha, g_\beta]$ で定義したとき，一致数の平均は $\sum_{\alpha,\beta} m[f_\alpha, g_\beta]$ であり，分散は $\sum_{\alpha,\beta} s^2[f_\alpha, g_\beta]$ で与えられる．

5 今後の課題

5.1 高次元分割表の難しさ

立方的とは限らない 3 次元データセット $[f : N \to I, g : N \to J, h : N \to K]$ を考える．3 次元分割表は

$$\mathsf{tab}[f,g,h] := (|f^{-1}(i) \cap g^{-1}(j) \cap h^{-1}(k)|)$$

である．2 次元周辺和について，

$$\mathsf{tab}[f,g] = \mathsf{tab}[f',g'] \Leftrightarrow \exists \pi \in S_n; f' = f\pi, g' = f\pi$$

なので，$f' = f\sigma, g' = g\tau, h' = h\rho$ $(\sigma,\tau,\rho \in S_n)$ と書いたとき, $\mathsf{tab}[f,g,h]$ と $\mathsf{tab}[f',g',h']$ の 2 次元周辺和が一致するための必要十分条件は，$\alpha,\beta,\gamma \in S_n$ で次の条件を満たすものが存在することである：

$$\begin{cases} f\sigma = f\alpha, g\tau = g\beta, h\rho = h\gamma \\ g\tau = g\alpha, h\rho = h\beta, f\sigma = f\gamma \end{cases}$$

この条件は，次のようにも表せる：

$$\begin{cases} S_f\sigma \cap S_g\tau \neq \emptyset \\ S_g\tau \cap S_h\rho \neq \emptyset \\ S_h\rho \cap S_f\sigma \neq \emptyset \end{cases}$$

この条件は，$(S_f\sigma, S_g\tau, S_h\rho)$ が，コセット幾何 $(S_n/S_f, S_n/S_g, S_n/S_h)$ の flag をなす事を意味する．つまり与えられた 2 次元周辺分布を持つ 3 次元分割表のランダムサンプリングは，この幾何の flag のランダムサンプリングをする必要がある．3 次元の場合は，$\rho = 1$ に固定して考えればよいことが分かる．2 次元の場合と違って，これでも議論は難しい．flag のサンプリングの代わりに，方程式 $\alpha\beta\gamma = 1$ $(\alpha \in S_f, \beta \in S_g, \gamma \in S_h)$ の解 (α,β,γ) の集合からのサンプリングとしてもよい．

3 次元の場合の MCMC 法ではいろいろなマルコフ基底が得られている．代数的の立場からの研究が望まれる．また収束の速さをどう求めるかも問題である．

5.2 q-アナログ，系統樹

データセットや分割表の理論の q-アナログ (q は素数ベキ) も考えられる．ごく簡単な例として 2×2 型分割表の生起行列 $H(X)$ の q -アナログは

$$\begin{aligned} H_q(X) &= \frac{a_1!a_2!b_1!b_2!}{n!x_{11}!\cdots x_{22}!}\times q^{x_{12}x_{21}}, \\ n! &:= (q-1)\cdots(q^n-1). \end{aligned}$$

とするのが自然である．生起確率の q-アナログが得られたとなると，分割表や対角和 (一致数) の q-アナログを考えたくなる．これはいったい何だろう．${}_2F_0$ 型超幾何多項式の積公式に対応する q-アナログもある．

一方，系統樹 (葉がラベル付けられた木) は生物などの分類の基本的表現方法である．例えば，生物同士の遺伝的距離を DNA などから測り,「非類似度行列」を作り，それから系統樹を描くことをする．数学的には，非類似度の定義された集合 N (例えばいくつかの生物の種) を，N 上の超距離空間にもっともストレスなく埋め込む方法を問題にする．こちらの方面は，膨大な研究がある．最近では，超距離空間

$$d(x,y)\le \max(d(x,z),d(a,y))$$

よりも，4 点条件

$$d(w,x)+d(y,z)\le \max\begin{pmatrix} d(w,y)+d(x,z) \\ d(w,z)+d(y,z)\end{pmatrix}$$

を満たす距離空間が使われている．この方面でもっとも有効と言われる近隣結合法は，日本人の分子生物学者 (根井，斎藤) が開発したものである ([根井]). 数学者の関与が少ないのは残念である．

さて，$n+1$ 枚の葉を持つ系統樹の頂点に $1,2,\cdots,n+1$ までのラベルを付け，内点には $n+2$ から $2n$ までのラベルを付ける．これによって，このような系統樹と完全マッチングとが 1 対 1 に対応する．すなわち $S_{2n}/2^n\cdot S_n$ と対応する ([[Diaconis 02]]). そうすると，S_{2n} 上のランダムウォークから系統樹のランダムウォークが誘導されることになる．したがって，分割表の場合と同様に収束の速さの議論が可能になる．

5.3 夢よもう一度

分割表の一致率検定では，分割表のトレースという線形統計量の正確な p 値を求めた．そうなると，分割表の独立性についてのカイ二乗統計量

$$\chi^2=\sum_{i,j}\frac{(x_{i,j}-a_ib_j/n)^2}{a_ib_j/n}$$

に関する正確な p-値を考えたくなる．この問題は，$I\cup J$ 上の二部グラフ構造と，与えられた長さのサイクルの数え上げの問題になる．残念ながら今のところ，役に立つ公式を得るのは難しそうである．

さらに，重み付き一致数の分布のモメントを求める問題も，実用上の要請から研究の価値が十分ある．分割表 $\boldsymbol{x}=\mathsf{tab}[f,g]=(x_{ij})$ と重み関数 $w: I\times J\to\mathbb{R}$ に対し，

$$\begin{aligned} x_0^w[f,g] &:= \sum_{i\in N} w_{f(i),g(i)} \\ &= \sum_{ij} w_{ij}|f^{-1}(i)|\cdot|g^{-1}(j)| \end{aligned}$$

と置く．このとき $x(\pi):=x_0^w[f,g\pi]$ $(\pi\in S_n)$ の平均値は $m=\frac{1}{n}\sum_{ij}w_{ij}a_ib_j$ で与えられる．さらに分散の公式も知られている．

$U_{ij}:=u^{w_{f(i),g(i)}}$ と置く (u は不定元)．ξ を S_n の既約指標としたとき，

$$\sum_\pi \xi(\pi)u^{x(\pi)}=\sum_\pi \xi(\pi)\prod_i U_{i,\pi(i)}$$

である．とくに $\xi=1$ の場合はパーマネントであり，$\xi=\mathrm{sgn}$ の場合は行列式が出てくる．このように対称群の表現論や対称関数の理論と関係ありそうだ．

6 比較言語学への応用

6.1 比較言語学における数理的方法

いくつかの言語を比較してその系統関係を明らかにする方法はいくつかある．これについては [安本 83] [安本 95] に詳しい．

基礎語彙を用いた比較方法としては，Polya の二項検定法がある [Polya 59]．彼はヨーロッパの十の言語の数詞のリストを用意し，各言語の組に対して，語頭文字の一致する数字の数を数えた．ふたつの言語での一致数が偶然の範囲を超えて大きいかどうかを，二項検定法で調べている．ハンガリー語だけが小さな一致数であり，他のどの言語とも偶然以上の関係が見いだせない．実際ハンガリー語だけはインドヨーロッパ系ではない．

もう一つの方法として，Oswalt のシフト検定法がある (1970)．統計学的には，一致数に対する並べ換え検定である．吉田が平均と分散の公式を群論的方法で求めた (1980 頃)．

さらに安本がこれらの方法を改良して使いやすいものにした (1980 頃から)．これは Polya の二項検定法で，ふたつの語頭音の一致の確率を吉田の公式で置き換えたものである．適用範囲の広さ，計算 (二項検定) が簡単で速いこと，タフなこと (データの欠損などに強い)，関係ない言語を関係ありと判断する間違いが少ないことなど多くのすぐれた点を持つ．

ただ，より高い精度の結果を求めたいなら，対称群を用いたシフト法を使って，正確な確率が計算できる．この方式はすでに述べたが，具体的な計算例として紹介したい．

6.2 アイヌ語 (A)・日本語 (J)・朝鮮語 (K) の比較

応用例としてこれら 3 言語の基礎語彙による比較を行う．現代アイヌ語幌別方言 (A)，奈良時代の上古日本語 (J)，15 世紀李氏朝鮮時代の中期朝鮮語 (K) について，200 語の基礎語彙表を用意する．これは安本氏の作成したものを使った．これらの言語で，語頭子音を対応させる写像をそれぞれ $f, g, h : N \longrightarrow \Lambda$ とする．ここで Λ は発音記号の集合だが，似た音同士はまとめておく．例えば 's' と 't' を同一視するとか，清音と濁音は同一視するなどである．基礎語彙表から分割表 (音韻対応表) を作る．例えば日本語と朝鮮語では次のようになる．例えば，'m' 行 'k' 列の 7 という数は，日本語で 'm' 音で始まり，朝鮮語で 'k' 音で始まる単語が 7 個あることを意味する．

J\K	k	m	n	p	r	t	w	y	-	
k	9	6	6	9	3	11	0	0	4	48
m	7	4	1	4	2	5	0	1	1	25
n	3	4	3	3	1	2	0	0	1	17
p	6	3	7	10	0	6	0	0	1	33
r	0	0	0	0	0	2	0	0	0	2
t	4	5	8	11	1	27	0	1	0	57
w	1	0	3	2	1	3	0	0	0	10
y	1	1	2	2	0	1	0	0	1	8
-	0	0	0	0	0	0	0	0	0	0
	31	23	30	41	8	57	0	2	8	200

対角和 (一致数) $x_0 = 53$ である．このようなやり方で一致数，平均，分散，歪度，尖度，偏差値を求め，正規・正確・二項検定の p 値を求める．3 言語について次のようにまとめられる．

	J × A	J × K	A × K	JAK(2)
x_0	41	53	56	23
m	36.535	36.035	37.53	8.2465
s	5.1070	5.1635	5.2094	2.9833
γ_1	0.0981	0.1035	0.1001	—
γ_2	-$0.0^3$862	$0.0^3$104	$0.0^3$481	—
z	0.8743	3.2615	3.5455	4.9454
P_n	0.1910	$0.0^3$509	$0.0^3$196	$0.0^6$380
P_e	0.2163	$0.0^2$156	$0.0^3$479	—
P_b	0.2312	$0.0^2$188	$0.0^3$943	$0.^4$104

上で，P_n, P_e，P_b はそれぞれ，正規検定，正確確率法，二項検定による P 値である．この表の P 値を観察すれば，次のようなことが分かる．

・アイヌ語は日本語よりも朝鮮語に近い．
・共通の核 (JAK(2)) がある，

これから想像されること．極東地域に 3 言語の元になる言語 (古極東アジア語 [安本 83]) があった．まず日本語が分かれ，その後アイヌ語・朝鮮語が分かれた．3 言語は各地域で独立に発展した．アイヌ語は縄文時代の言語という人もいる．

ただ，その成立が割と新しい (13 世紀頃擦文文化期以降) 可能性もあると思う．あまり知られていないようだが，１３世紀末から１４世紀前期にかけて，アイヌとモンゴル帝国がアムール川下流域とサハリンをめぐって一進一退の攻防を繰り広げていた (北の元寇)．その後大陸との連絡の切れたオホーツク人

(アイヌと仲が悪かった) などの北方系民族の言語がアイヌ語の基礎語彙にいくつか取り込まれた可能性がある．この北方系の言語が朝鮮語系 (こちらの可能性が高い) であるか，そうでなくても朝鮮語とアイヌ語両方に影響したとすれば，アイヌ語と朝鮮語の意外な近さが説明できる.『ユーカラ』にあるアイヌの英雄叙事詩が，アイヌとオホーツク人との戦いを描いたものだという説 (知里真志保) もある．

閑話休題．上古日本語と中期朝鮮語の間の一致数 $x_0 = 53$ について，正確な p 値は，次のようになる．

$$\frac{\begin{array}{c}971072955976527370489065992048363534 6\\ 952586831809163300169726237291628421702 7\\ 250955246737090436665097707182327060666 3\end{array}}{\begin{array}{c}621808756731104460234413202960802888298 3\\ 704548417072721106375207655256584722289 0\\ 439682694827900288368064738734554630000 0\end{array}}$$

分母 120 桁，分子 117 桁である．$P(x \geq 53)$ は，

$$0.000554(\text{正規}), 0.00156169(\text{正確}), 0.00238(\text{二項})$$

となっている．予想通り，正規検定では p 値が過小に出ている．正規検定を使う場合は，有意性判断の危険率のボーダーラインを 0.001 するとかの対策が必要である．何十という言語の比較ならもっと小さくする必要が出てくる．正確検定は正しい p 値が得られるので，できればこの方式を使いたい．計算時間も 1 秒もかからない (R 言語と Asie/Risa)．汎用の数式処理システムでも似たようなものである．安本方式の二項検定法は，p 値が正確確率よりも少し大きく出ている．これは計算方式の精度のためではなく，考え方の違いである．基礎 200 語の語彙リストを，すべての単語からの標本と見るなら，二項検定法 (復元抽出) になるし，それ自体母集団と見るなら対称群を使った一般シフト法 (非復元抽出) となる．正確も二項もどちらも正確な p 値である．

言語については多くの課題が残っている．例えば Swadesh の公式 $x_0(t) = x_0(0)r^{2t}$ ($r \fallingdotseq 0.81$)．数学的に見てもこの公式は問題がある．分岐年代 $t \to \infty$ とすれば，一致数 $x_0(t) \to m$ (平均一致数) のはずだが，Swadesh の公式ではそうなっていない．これについては北大理学部ホームページにある『サイエンストピックス』のどこかに私の書いたものがある (古い版で間違いが多い)．

参 考 文 献

[Agresti 92] A.Agresti, A survey of exact inference for conitingency tables, *Stat. Sci.*, **7** (1992), 131–177.

[Brown 00] K.S.Brown, Semigroups, Rings, and Markov Chains, *J.Theoretical Probability*, **13** (2000), 871–938.

[Diaconis 81] P.Diaconis and M.Shahshahani, Generating a random permutation with random transpositions, *Z.Whar.* **57** (1981), 159–179.

[Diaconis 88] P.Diaconis, "Group Representaions in Probability and Statistics," LN-Monograph series 11, Institute of Math.Stat., 1988.

[Diaconis 94] P.Diaconis and L.Saloff-Coste, Random walks on finite groups : a survey of analytic techniques, 44–75, in "Probability Measures on Group and Related Structres", H.Heyer(ed), 1994.

[Diaconis 02] P.Diaconis and S.P.Holmes, Random Walks on Trees and matchings, *Electric Jounal of Probability*, **7** (2002), 1–17.

[Diaconis 05] P.Diaconis and I.M.Isaacs, Supercharacters and superclasses for algebra groups, 2005 (preprint).

[Gluck 97] D.Gluch, Characters and random walks on finite classical groups, Adv.Math., **129** (1997), 46–72.

[Good 94] P.Good, "Permutation, Parametric, and Bootstrap Tests of Hypothesis," Springer, 1994, 2000, 2005.

[Heyer 94] H.Heyer(ed), "Porbability measures on groups and related structures," Proc. Oberwolfach 1994. [Hinkley 97] D.V.Hinkley, "Bootswtrap Methods and their Application," Cambridge, 1997.

[Liebeck] M.W.Liebeck and A.Shalev, Character degrees and random walks in finite groups of Lie type, 1–31.

[Mielke 01] P.W.Mielke, K.J.Berry, "Permutation Methods", Springer, 2001, 2007.

[Muirhead 82] R.B.Muirhead, "Aspects of Multiplicative Statistical Theory," Wiley 1982, 2005.
[Polya 59] ポリア『発見的推論　そのパターン—数学における発見はいかになされるか２』丸善 (1959)
[Saloff 03] L.Salloff-Coste, Random Walks on Finite Groups, in "Probability on Discrete Structures " (Encyclopaedia of Mathematical Sciences), 264–346, 2003.
[Semple 03] C.Semple and M.Steel, "Phylogenetics", Oxford, 2003.
[Sturmfels 05] L.Prachter, B.Sturmfels (編), "Algebraic Statistics for Computational Biology," Cambridge, 2005
[青木 07] 青木敏，竹村彰道，統計学とグレブナー基底—計算代数統計の発展と展開—,「数学」**59**, No.3 (2007), 283–302.
[伊庭 05] 伊庭，種村『計算統計 2— マルコフ連鎖モンテカルロ法とその周辺』岩波 2005
[汪 03] 汪 ほか『計算統計 I—確率計算の新しい手法』岩波 2003
[水川 04] H.Mizukawa, Zonal spherical functions on the complex reflection groups and $(n+1, m+1)$-hypergeometric functions, *Adv. Math.,* **184** (2004), 1-17.
[竹村 84] A.Takemura, "Zonal Polynomials," Inst.Math.Stat.LN., Monograph Series 4 (1984).
[富澤 06] 富澤貞男，統計学における正方分割表の解析,『数学』**58**, No.3 (2006), 263–287.
[根井 06] 根井正利 and S. クマー,「分子進化と分子生物学」培風館 (2006).
[日比 06] 日比孝之 (編)『グレブナー基底の現在』数学書房 2006
[安本 83] 安本美典『日本語の誕生』大修館書店 1983.
[安本 95] 安本美典『言語の科学』朝倉書店 1995』

あとがき．この講演は，統計関係の研究集会で話して来たことをまとめたものです．Polya の二項検定法，Oswalt のシフト法については省略しました．[安本 95] を見てください．また文献を追加しました．

最後に，代数とはかなり遠い分野の研究に講演の機会を与えていただいたことに感謝します．

(2007/11/21)

役に立つサイト

・Persi Diaconis．多数の論文．
http://stat.stanford.edu/ cgates/PERSI/index.html
・竹村彰道．プレプリント，講演資料など．
http://www.e.u-tokyo.ac.jp/ takemura/
・青木敏．学位論文，3D マルコフ基底のアニメ．
http://www.sci.kagoshima-u.ac.jp/ aoki/
・Bernd Sturmfels．多数の論文．
http://math.berkeley.edu/ bernd/
・大森裕浩．MCMC 法．
http://www.e.u-tokyo.ac.jp/ omori/
・統計解析システム『R』に関する Wiki.
http://www.okada.jp.org/RWiki/?RjpWiki
・斉藤成也．論文，解説記事など多数．
http://sayer.lab.nig.ac.jp/ saitou/index-j.html
・安本美典．講演会記録の中に言語関係がある．
http://yamatai.cside.com/

言語・生物関係和文献追加

[風間 78] 風間喜代三『言語学の誕生－比較言語学小史』岩波新書 (1978)
[亀井 88, 97] 亀井孝・河野六郎・千野栄一『言語学大辞典』,『日本列島の言語』三省堂 (1988〜, 1997)
[斎藤 05] 斉藤成也『ＤＮＡから見た日本人』ちくま新書 (2005)
[斎藤 07a] 斎藤成也『ゲノム進化を考えるー系統樹の数理から脳神経系の進化までー』数理科学、SGC ライブラリ５３，サイエンス社（2007）
[斎藤 07b] 斎藤成也『ゲノム進化学入門』共立出版 (2007/12 月)．(第 12 章「遺伝子系統樹の作成」は，数学専門の人でも読める)
[三中 06] 三中信宏『系統樹思考の世界〜すべてはツリーとともに〜』講談社現代新書 (2006)

(2007/11/25 追記)