正方分割表の一致率検定のための代数的方法

吉田 知行 (北星学園大学 経済学部)
t-yoshida@hokusei.ac.jp

2017/1/29 金沢

概要: 本講演では，2 元以上の正方分割表の一致率 (対角和) の検定について，正確な p-値を求める方法を紹介する．この方法は，フィッシャーの並べ替え検定, フィッシャーの正確確率法の流れをくむ方法で，代数的には，対称群のデータセットへの作用と指標理論の応用分野である．また，有限群上のランダムウォーク (RW) とも関係している．

キーワード： 正確確率法，並べ替え検定，有限群上のランダムウォーク，分割表の列挙問題，有限群の表現.

1 一致係数

値 $1, 2, \cdots, r$ を取るふたつの確率変数 X, Y について，$\pi_{ij} = P(X = i, Y = j)$ と置く．このとき

$$p_o = P(X = Y) = \sum_i \pi_{ii}$$

であり，X と Y が独立なとき，偶然で $X = Y$ となる確率は $p_c = \sum_i \pi_{i+}\pi_{+i}$ である．ここで,

$$\pi_{i+} = \sum_j \pi_{ij}, \quad \pi_{+i} = \sum_j \pi_{jj}$$

Cohen の一致係数 κ は次で定義される．

$$\kappa := \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}$$

このとき κ の値は，0 と 1 の間にあり，1 に近いほど $X = Y$ の可能性が高いと言える．ただ，どの程度 κ の値が大きいと同一と判断できるのかは明瞭でない．0.8 以上だと，よく一致していると判断されているようだ.

2×2 の分割表を考える．

X \ Y	1	2	
1	a	b	$a+b$
2	c	d	$c+d$
	$a+c$	$b+d$	n

π_{ij} の推定値を

$$\widehat{\pi}_{11} = a/n, \widehat{\pi}_{12} = b/n, \widehat{\pi}_{21} = c/n, \pi_{22} = d/n$$

ですれば，κ の推定値は

$$\begin{aligned}\widehat{\kappa} &= \frac{(\pi_{11} + \pi_{22}) - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})}{1 - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})} \\ &= \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}\end{aligned}$$

となる． これより $\chi^2 \geq n\widehat{\kappa}^2$ なのでカイ二乗検定ができる.

フィッシャーの正確確率法を適用することも可能である．しかしカテゴリー数 r やデータ数 n が増えると計算は困難になる．とくに多元分割表の一致数検定に対しては方式も確立していないようである.

さらにブートストラップ法を使うことは可能だし，マルコフ連鎖モンテカルロ法 (MCMC 法) により与えられた周辺度数を持つ分割表を発生させてそれから p-値の (必要ならいくらでも正確な) 近似値を求めることも可能である.

2　基本的アイデア

本講演では，一致数検定のための正確な p-値を計算するための代数的方法を紹介する．まず計算方法と計算例を述べ，その後数学的背景を述べる．表立ってはいないが，フィッシャーの並べ替え検定を代数の言葉で書き直したものと言える．

なお代数学の予備知識は線形代数と確率・統計以外は想定していない．

基本的なアイデアは，以下の集合上のランダムウォーク (RW) を順に構成することである

(a) 対称群 S_n.
(b) 与えられた周辺度数を持つ 2 次元データセットの集合.
(c) 与えられた周辺度数を持つ 2 次元分割表の集合.

これによって分割表の列挙が可能になる．とくに一致率の検定も可能になる．有限群，とくに対称群の表現論を使えば，収束の速さも評価できる．

興味深いのは，グレブナー基底を用いた有名な列挙法に登場するマルコ基底が，対称群の互換の共役類を用いたものと結果的に等しいことである．なぜこうなるのかはまだ未解決の問題である．また，対称群の構造から，3 元以上の分割表の自明でない列挙は不可能である．

一致数の場合は特別で，正確な p-値を求める公式がある．${}_2F_0$ 型超幾何多項式を使ったアルゴリズムだが，計算量はデータのサイズ n でなく，n/r に依存している．残念なことに，カイ二乗統計量による独立性の検定では，厳密な p-値は求まらない

3　一致率検定のための正確な p-値

$\boldsymbol{x} = (x_{\lambda,\mu})_{\lambda,\mu\in\mathcal{L}}$ を正方分割表とする．ここで $\mathcal{L}$ はカテゴリーの有限集合である．このとき**一致数**とは，$\boldsymbol{x}$ の対角和である．

$$\mathrm{Tr}(\boldsymbol{x}) = \sum_{\lambda\in\mathcal{L}} x_{\lambda,\lambda}$$

2 元分割表のふたつの周辺和とサイズを

$$\boldsymbol{a} = (a_\lambda)_\lambda, \quad a_\lambda := \sum_\mu x_{\lambda,\mu},$$
$$\boldsymbol{b} = (b_\mu)_\mu, \quad b_\mu := \sum_\lambda x_{\lambda,\mu},$$
$$n = \sum_{\lambda,\mu} x_{\lambda,\mu} = \sum_\lambda a_\lambda = \sum_\mu b_\mu$$

とする．

$\mathcal{L}\times\mathcal{L}$ 型で周辺和 $\boldsymbol{a},\boldsymbol{b}$ を持つ分割表の集合を $\mathsf{tab}(\boldsymbol{a},\boldsymbol{b})$ とする．$\boldsymbol{x}\in\mathsf{tab}(\boldsymbol{a},\boldsymbol{b})$ の生起確率は

$$H(\boldsymbol{x}) = \frac{\boldsymbol{a}!\boldsymbol{b}!}{n!\boldsymbol{x}!} = \frac{\prod_\lambda \cdot a_\lambda! \times \prod_\mu b_\mu!}{n! \times \prod_{\lambda,\mu} x_{\lambda,\mu}!}$$

周辺度数 $\boldsymbol{a},\boldsymbol{b}$ を固定したとき，一数の分布を知りたい．ここでは，結果だけを書いておく．${}_2F_0$ 型超幾何多項式を

$${}_2F_0(-a,-b;z) := \sum_{n=0}^{\infty} \binom{a}{n}\binom{b}{n} n! z^n$$

で定義する．これは a,b が非負整数のとき多項式になる．次数は $\min(a,b)$ である．

数列 $\{q(k)\},\{p(k)\},\{P(k)\}$ $(k=0,1,\cdots,n)$ を次で定義する：

$$\prod_\lambda {}_2F_0(-a_\lambda,-b_\lambda;z) = \sum_{k\ge 0}\binom{n}{k} k! q(k) z^k,$$
$$p(k) = \sum_{j=k}^{n} (-1)^{j-k}\binom{j}{k} q(j),$$
$$\sum_{k=0}^{n} p(k) z^k = \sum_{k=0}^{n} q(k)(z-1)^k,$$
$$\sum_{k=0}^{n} P(k) z^k = 1 + z\sum_{k=1}^{n} q(k)(z-1)^{k-1}$$

最後の $P(k)$ が，一致数が k 以上になる確率である:

$$P(k) = \mathrm{Prob}(\boldsymbol{x}\in\mathsf{tab}(\boldsymbol{a},\boldsymbol{b}) \mid \mathrm{Tr}(\boldsymbol{x}) \ge k).$$

4　言語への応用

私にとって問題の発端は比較言語学であった．ポリアの本 [Polya 59] にヨーロッパの 10 の言語を，数詞の語頭文字の一致数で比較した羚が載っている．

その後，日本語の誕生について，安本美典氏の研究が『数理科学』に載り非情に京見を持った．彼の研究の中に，日本語，アイヌ語，朝鮮語の基礎 200 語についての音韻対応表がある．ここで似た音はひとまとめにしてにして数えてある．

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

表 1. 上古日本語・中期朝鮮語の音韻対応表

カッパ検定では，有意性は認められない．

$$p_o = 53/200 = 0.265,$$
$$p_c = 36.035/200 = 0.180175,$$
$$\kappa = 0.10347,$$
$$\chi^2 \approx 200 \times 0.10347 = 2.14121$$

しかし，言語学者 Oswalt のシフト法 (リサンプリング法の一種) では有意性が認められる．

一致数 $x_0 = 53$
背景点平均 $m = 36.155,$
背景点平均 $\sigma = 5.1647,$
$z = 3.2615, \quad P = 0.0^3554$

背景点とは，偶然による一致数である．

いよいよ，いくつかの方式での一致率の p-値を見てみる．P(正規) とあるのがオズワルトの方法で，正規分布の上側確率である．P(正確) とあるのが，前節で述べた正確な確率である．P(二項) とあるのが，ポリヤが使った二項検定法である．

	J × A	J × K	A × K
x_0	41	53	56
m	36.535	36.035	37.53
s	5.1070	5.1635	5.2094
z	0.8743	3.2615	3.5455
P(正規)	0.1910	0.0^3509	0.0^3196
P(正確)	0.2163	0.0^2156	0.0^3479
P(二項)	0.2312	0.0^2188	0.0^3943

表 2. 日本語 (J)，アイヌ語 (A)，朝鮮語 (K) の比較 (1)

安本氏は，このような観測から，3 言語には，共通の核 (JAK(2)) があり，核を除けば相関がないとの結果を出した．これから想像されることがいくつかある．極東地域に 3 言語の元になる言語 (安本の古極東アジア語) 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．

3 言語に対する一致率検定も考えられる．これは日本語・アイヌ語朝鮮語のまとまり具合を調べるものである．一致数として $x_{JA} + x_{AK} + x_{KJ}$ を取ったものが表 3 の JAK(1) である．また，3 言語すべての一致数を差一ようしたものが AJK(2) である．

	JAK(1)	JAK(2)
x_0	151	23
m	109.88	8.2465
s	8.9160	2.9833
z	4.6119	4.9454
P(正規)	0.0^5252	0.0^6380
P(二項)	0.0^4227	$0.^4104$

表 3. 日本語 (J)，アイヌ語 (A)，朝鮮語 (K) の比較 (2)

JK の一致率のの正確な P-値 $P(x \geq 53)$ は

$$\frac{\begin{pmatrix} 971072955976527370489065992048363534 6 \\ 952586831809163300169726237291628421702 7 \\ 250955246737090436665097707182327060666 3 \end{pmatrix}}{\begin{pmatrix} 621808756731104460234413202960802888298 3 \\ 704548417072721106375207655265584722289 0 \\ 439682694827900288368064738734554630000 0 \end{pmatrix}}$$

である．117 桁/120 桁の数である．3 つの方法 (シフト，二項，正確) の数値を比べると次のようになる．

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

0 どの方法でも，有意性が認められる．とくに 3 言語 JAK のまとまりと AK の近さが目に付く．

二項分布による値が正確な値にかなり近い．面倒なシフト法を使うより，二項検定で十分であろう．

近年，日本人の起源問題に大きな進展があり，日本列島の好機 k ッ湯石器時代人と縄文人の南方由来説が研究されている．これ F については安本氏が日本語とインドネシア語やカンボジア語との関係を発見している．これについても二項検定や推定の計算結果を紹介したい．

5　分割表の列挙問題

$I \times J$ 型分割表とは，非負整数行列 $\boldsymbol{x} = (x_{ij})$ のことである．その周辺和とサイズを次で定義する:

$$x_{+j} := \sum_{i \in I} x_{ij}, \quad x_{i+} := \sum_{jinJ} x_{ij}$$
$$n := \sum_{ij} x_{ij} = \sum_{i} x_{i+} = \sum_{j} x_{+j}$$

ここでは添え字を i, j にした．

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

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

であるとする．(A_i) と (B_j) を，サイズ n の集合 N の $\boldsymbol{a}$ 型と $\boldsymbol{b}$ 型のランダムに選んだ分割としたとき，$H(\boldsymbol{x})$ は $|A_i \cap B_j| = x_{ij}$ $(\forall i, j)$ となる確率である．

独立性の検定で，p-値は次で定義される：

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

ここで，和は $\chi^2(\boldsymbol{x}) \geq \chi_0^2$ を満たす分割表 $\boldsymbol{x} \in \mathsf{tab}(\boldsymbol{a}, \boldsymbol{b})$ についての和である．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**)．しかしこの方法はちょっと $|I|, |J|$ 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．

また，分割表の大量のランダムサンプリングが得られれば，p-値の近似値が得られる．ランダムに分割表を作るなら**マルコフ鎖モンテカルロ法** (MCMC 法 (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

$$\boxed{\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix}}$$

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．

6　分割表と対称群

分割表の列挙やランダムサンプリングへの対称群上のランダムウォーク (RW) を使う方法がある．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，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．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 (N 上の置換全体のなす群) 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．すなわち，ふたつ要素は互いに移り得る．一点 $[f]$ の固定部分群 G_f は分割 $N = \coprod_{i \in I} f^{-1}(i)$ に対応する Young 部分群と呼ばれる．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．すなわちふたつの要素は $\mathsf{tab}[f, g] = \mathsf{tab}[f', g']$ 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．

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

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

$$\frac{\sharp\{\pi \in S_n \mid \mathsf{tab}_{f_0, g_0}(\pi) = \boldsymbol{x}\}}{n!} = \frac{\boldsymbol{a}!\boldsymbol{b}!}{n!\boldsymbol{x}!} = H(\boldsymbol{x})$$

とくに，tab_{f_0, g_0} 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．

一致数の正確な p-値の場合も，分割表についての和のかわりに対称群の和に書き直し手証明できる．

7 有限群上のランダムウォーク

この節はやや専門的な用語が入る．有限群の作用する集合 (例えば与えられたデータセット) 上のランダムウォークを考えるべきだが，簡単のため，有限群 G 上の RW を考える ([Diaconis 88])．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：すなわち，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．ここで，$\mathbb{C}G$ は複素数係数群環であり，すなわち G の要素の形式的一次結合で，G の積を拡張した積の演算を持つ．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 (N は $\mathbb{C}G$ のどの要素とも可換) 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．こ/7 のことは，右辺の $|\chi(x)/\chi(1)|$ の挙動によって RW の収束や収束の速さが決まることを意味する．

E を μ のサポートとする．このとき $k \to \infty$ での M^k の漸近的挙動は次のようになる：

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

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 $\chi \neq 1_G$ 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 (代表元が t) の場合，この条件は G が E と t で生成されることと同値である．

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

8 代数学者から見た統計学とその応用

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．

抽象数学の代表と思われている代数学が活躍するのは嬉しいものである．

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 (とくに代数学) との相性は悪くないと思われる．しかし，数学からの積極的関与は今も乏しい．例えば，生物の系統樹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氏や根井正利氏が，基本原理を考え，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が，数学者の貢献はあまりなかったらしい．

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．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．0

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 1 0 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．

このような研究によって意外な分野が結びつくこ

とがある．例えば，一致数の平均値公式

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

の右辺は暗号理論で一致反復率と呼ばれている量である．また，グレブナー基底を使った MCMC 法で，マルコフ基底 $\boxed{\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix}}$ は，対称群の互換のランダムサンプリングから得られるマルコフ連鎖と同一である．

そのほか分割表上の RW と対称群の表現の関係もある．前者は代数統計で活発に研究されている分野であり，後者は Diaconis たちによって研究されて板分野である．しかしこれまで両者を結び付けた研究はなかったようである．

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．

数学者から見て，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関わっているかを紹介した．

9　確率論と有限群論

ただ注意しておきたいのは，代数学と確率・統計での術語の違いである．そのため量子確率論や，量子ランダムウォークの分野ではもっぱら代数の言葉で確率論やランダムウォークを記述している．

ここでもそのような例を挙げておく．記号は新しいものが多いので警戒してほしい．Δ で単位区間を表す．また，X, Y などは (確率変数でなく) 有限集合とする．∇X によって X 上の確率測度全体の集合を表す：

$$\nabla X := \left\{ \mu : X \to \Delta \,\middle|\, \sum_{x\in X} \mu(x) = 1 \right\}$$

ただ，このままでは群の作用などの取り扱いが易しくないので，X を頂点集合とする単体を考える：

$$\Delta X := \left\{ \widehat{\mu} := \sum_{x\in X} \mu(x)x \,\middle|\, \mu(x) \geq 0, \sum_{x\in X} \mu(x) = 1 \right\}$$

これを X を基底とする**マルコフ空間**と呼びたい．この集合は対応 $\mu \leftrightarrow \widehat{\mu}$ によって ∇X と一対一に対応している．どちらの集合も凸集合である．

G が有限群なら，群の演算 $G \times G \to G$ はマルコフ群環 ΔG 上の積に拡張できる．また，有限群 G の有限集合 X への作用 $G \times X \to X$ はマルコフ群 ΔG のマルコフ集合 ΔX への作用に拡張できる．確率論では，体上の線形空間の代わりはマルコフ集合であり，線形写像の代わりは，マルコフ写像 (凸写像) $f : \Delta X \to \Delta Y$ である．代数的には，群環 $\mathbb{C}G$ の代わりにマルコフ群環を使った表現論を作ることになる．当然 $\Delta G \subset \mathbb{C}G$ なので，マルコフ表現から通常表現に移したときの様子を観察することになる．このように，確率論・統計学も有限群論との関係をうかがわせるいくつかの事実がある．[吉田 09] 参照.

参 考 文 献

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

[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.

[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
[竹村 84] A.Takemura, ”Zonal Polynomials,” Inst.Math.Stat.LN., Monograph Series 4 (1984).
[竹村・日々 15] 竹村彰通・日比孝之他『グレブナー教室』共立出版 (2015)
[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』
[吉田 07] 吉田知行，分割表の一致率検定への有限群論と組合せ論の応用，代数学シンポジウム，於神戸大学．ネットに pdf ファイルあり．
[吉田 09] 吉田知行，Finite Gelfand pairs and Markov chain Monte-Carlo method, 「表現論と組合せ論」RIMS 講究録第 **1689** (201),164-170,

(2016/12/21)

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/
・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/