研究ノート

分割表の一致率検定とFisherの正確確率法

吉 田 知 行

目次

[要旨]
このノートは，[8]への追加である。正方分割表 $X=(x_{\lambda,\mu}$ の対角和 $\mathrm{Tr}(X)$ を一致数という。与えられた周辺度数分布を持つ正方分割表のうち，一致数 $\mathrm{Tr}(X)\geq x$ となる確率 (p-値) を求めるための計算方法を述べる。これは ${}_2F_0$-型多項超幾何多項式を用いて簡明に表せる。また，2×2-型分割表の場合，一致数からカイ二乗統計量を計算出来る。したがって，2×2-型分割表に対するFisherの正確確率法と同じ p-値を求めることができる。

1　はじめに

1.1　比較言語学からの問題提起

比較言語学はいくつかの言語の比較によって，言語の系統や関係を明らかにする言語学の一分野である。歴史言語学と呼ばれることもある。いくつかの方法があるが，ここでは基礎語彙 (数詞，基礎 100 語，200 語) による比較を考える。

例えば，上古日本語・中期朝鮮語・アイヌ語幌別方言の基礎 200 語の一部は次で与えられている。一部発音記号は省略した。

	意味	日本語	朝鮮語	アイヌ語
1	all	mina	motĕn	oppita
2	ash	нанi	tʃĕi	uyna
⋮	⋮	⋮	⋮	⋮
99	woman	me	kyətʃip	mat
100	yellow	kĭ	nurŭ	siker
⋮	⋮	⋮	⋮	⋮

表 1.1 日本語・朝鮮語・アイヌ語の基礎語彙

このような語彙表をいくつかの言語について作成しておく。

比較の手続をまとめると次のようになる。

(1) ふたつの言語の基礎語彙表について，語頭音の一致数 x_0 を数える。
(2) 偶然による一致数 m を求める。
(3) x_0 以上の一致率が偶然得られる確率 (p-値) を求める。
(4) この確率が 0.05 より小さければ，5% 水準で有意と判断する。

著者は，安本美典氏が使ったこのような方法に強い興味を持った。ポリアの「いかにして問題を解くか 2」にヨーロッパの数詞を比較する例題が載っており，また子供の頃から比較言語学には関心があった。しかし安本はもっぱらオズワルトのシフト法を主に使っていた。

40 年近い昔のことで，そこには私の専門の「群論」(ぐんろん，代数学の一分野) の雰囲気が強く感じられた。実際，群論を用いて，偶然による一致数の平均値と分散の公式 (後

キーワード：一致率検定，正確確率法，データセットと分割表，比較言語学

述平均値公式）が得られた。

現代の統計学からは，シフト法はリサンプリング法のひとつであり，フィッシャーの並べ替え検定の流れをくむ。また，安本とオズワルトの私信には，ブートストラップ法（エフロン 1981）のアイデアが述べられていた。

このような群論と統計学の関係については殆ど意識されてこなかったようである。しかし今世紀になって，分割表のランダム生成に関する代数的方法（グレブナー基底を使う）が発表され，それにより分割表の独立性上側確率はいくらでも正確に求められることになった。そのため，群論の統計学への応用について今一度考えてみる価値があると考えた。

偶然による一致数を求める。オズワルトのシフト法では，片方の言語の単語をひとつずつずらしながら語頭子音を比較する。使う群は巡回群である。

項目	意味	日本語	朝鮮語
1	all	mina	motĕn
2	ash	нанi	tʃĕi
3	bark	кана	kəptʃir
4	belly	нara	pʊi
⋮	⋮	⋮	⋮
99	woman	me	kyətʃip
100	yellow	kï	nurŭ
1	all	mina	(motĕn)

表1.2 シフト法。片方の単語をずらしながら比較する。

1.2 現代統計学の発展

現代の統計学の発展はめざましい。その原動力として様々なものが考えられる。

（i）計算機とネットワークの大幅かつ急速な性能向上。
（ii）統計学の基礎の発展。
（iii）統計関係のソフトの性能向上。

それによって統計学の方法は一新され，多方面への応用範囲が広がった。一方，大量の計算が必要なため，かっては実用的でないとされていた方法が復活している（Fisher の並べ替え検定，正確確率法など）。ブートストラップ法など，現代の計算機の能力なしには使い物にならない。

とくに 1990 年代に「MCMC 革命」とでもいうべき大変革が起こった。ここで MCMC とはマルコフ連鎖モンテカルロ（Markov chain Monte Carlo）法の略である。原型は 1950 年代にすでに物理の分野（多重積分の数値計算）にあった。もともと汎用性のある方法だったので，統計と結び付くのは自然なことであった。その余波としてベイズ統計学の復興があった。現在主流になりつつあるベイズ統計も，30 年ほど前には，日本であまり研究されていなかったし評価も低かった。最近の書店の統計の棚を見るとベイズ統計が相当幅をきかせている。

統計学への応用として，次のような代数の分野が目に付く。

(1) MCMC 法による分割表の大量生成方式へのグレブナー基底の応用。多項式環や代数幾何の応用。Strumfels [4], 日比 [1] 参照。

(2) 有限群上のランダムウォークへの表現論の応用。これについては Diaconis [5] と Ceccherini-Silberstein [6] がよい。

実は (1) と (2) は密接に関係している。(1) の問題点は，収束性の議論が貧弱に感じられることである。また (2) は統計学という

より確率過程（ランダムウォーク）の分野に属する。

ところで，周辺分布を固定した分割表の集合は，対称群のヤング部分群による両側剰余類の集合と一対一に対応している。したがって，MCMC 法で分割表の大量生成しようとするなら，対称群の元の大量生成をすればよい。互換の集合は対称群の生成元になっているので，互換をランダムにとって次々に掛けていけば，対称群の元，したがって分割表がいくらでも得られる。しかし不思議なことに，(1) と (2) の関係はこれまで知られていないようである。

代数を使った統計の分野は「代数統計」とか「計算代数統計」と呼ばれている。代数の分野では，群論（有限群，線形群，表現），代数幾何（多項式環），可換環論（対称式，不変式），組合せ論（ヤング図形，数え上げ，母関数，有限幾何，グラフ，結合的概型），幾何学（凸多面体，ルート系）といったものが使えそうである。今この分野は第二期の爆発的発展の中にあるように感じる（第一期はグレブナー基底の登場）。

1.3 分割表

以下，$\mathcal{L}\times\mathcal{M}$ 型の（2 元）分割表と言うときは，$\mathcal{L}\times\mathcal{M}$ 型非負整数行列 $\boldsymbol{x}=(x_{\lambda\mu})$ を意味する。ここで λ,μ は質的変数で，それぞれ有限集合 $\mathcal{L},\mathcal{M}$ に属しているとする。総度数を n，周辺和を $x_{\lambda+},x_{+\mu}$ とする。

$$
\begin{aligned}
x_{\lambda+} &:= \sum_{\mu} x_{\lambda\mu},\\
x_{+\mu} &:= \sum_{\lambda} x_{\lambda\mu},\\
n &:= \sum_{\lambda\mu} x_{\lambda\mu} = \sum_{\lambda} x_{\lambda+} = \sum_{\mu} x_{+\mu}.
\end{aligned}
$$

ここでは，与えられた周辺度数 (a_λ) と (b_μ) を持つ分割表 $\boldsymbol{x}$ の分布は，**多項超幾何分布**

$$
\begin{aligned}
H(\boldsymbol{x}) &= P(X=\boldsymbol{x}\mid x_{\lambda+}=a_\lambda, x_{+\mu}=b_\mu)\\
&= \frac{\prod_\lambda a_\lambda!\prod_\mu b_\mu!}{n!\prod_{\lambda\mu}x_{\lambda\mu}!}
\end{aligned}
$$

にしたがうものと仮定する。

$\mathcal{L}=\mathcal{M}$ であるような分割表を $\mathcal{L}$-型の**正方分割表**という。その対角和（トレース）を**一致数**（measure of agreement）という。

この論文の主目的は，正方分割表の対角和の分布と，正確な p-値を求めることである。

本論講においては，有限群論と圏論（カテゴリー論）のごく初歩の概念を使っている。これについては，代数の教科書を参照して欲しい。

2 データセットと分割表

2.1 データセットの代数

$\mathcal{K}$ 上のデータセットとは，有限集合の間の写像 $[h:N\to\mathcal{K}]$ のことである。単なる写像と区別するために括弧でくくってある。$n=|N|$ を**サイズ**という。統計学的には，N はサンプルの集合，$\mathcal{K}$ は名義尺度，h は確率変数で，$h(i)$ は $i\in N$ の属するカテゴリー（統計的な意味）である。すなわち，$\mathcal{K}$ 上のデータセットは $\mathcal{K}$ 上の有限集合，すなわちスライスカテゴリー $\mathbf{set}/\mathcal{K}$ の対象である（**set** は有限集合と写像のなすカテゴリー）。したがって $\mathcal{K}$ 上のデータセット同士の射 $\theta:[h:N\to\mathcal{K}]\to[h':N'\to\mathcal{K}]$ は，写像 $\theta:N\to N'$ で，$h'\circ\theta=h$ を満たすものである。カテゴリー論の術語を使ってデータセットについて同型射，全射，単射，等化などの概念が定義できる。

スライスカテゴリーにおける直既約な対象は，$i_\kappa:\{*\}\to\mathcal{K};*\mapsto\kappa$（ここで $\kappa\in\mathcal{K}$）の形をしている。したがって κ に対応する Burnside 準同型は

$$
\mathsf{tab}_\kappa:[h:N\to\mathcal{K}]\mapsto|\hom([i_\kappa],[h])|=|h^{-1}(\kappa)|
$$

で与えられる。したがって Burnside 準同型は写像

$$\mathsf{tab} : \mathbf{set}/\mathcal{K} \to \mathbb{N}_0^{\mathcal{K}}; [h] \mapsto \mathsf{tab}_\kappa([h]) = (|h^{-1}(\kappa)|)_\kappa$$

を与える。配列 $\mathsf{tab}[h] = (|h^{-1}(\kappa)|)_\kappa$ をデータセット $[h]$ の**度数分布表** (table) という。こう考えると抽象バーンサイド環の理論が使える。

$$\mathsf{TAB}(n;\mathcal{K}) := \{\boldsymbol{c} = (c_\kappa)_{\kappa\in\mathcal{K}} \mid c_\kappa \in \mathbb{N}_0, \textstyle\sum c_\kappa = n\}$$

と置くと，上の写像 tab は，$\mathsf{DS}(N;\mathcal{K}) \to \mathsf{TAB}(n;\mathcal{K})$ (ここで $n := |N|$) を与える。

$\boldsymbol{c} \in \mathsf{TAB}(n;\mathcal{K})$ に対し，

$$\mathsf{DS}(\boldsymbol{c}) := \mathsf{tab}^{-1}(\boldsymbol{c}) = \{[h : N \to \mathcal{K}] \mid \mathsf{tab}[h] = \boldsymbol{c}\}$$

と置く。N 上の対称群 S_N は $\mathsf{DS}(N;\mathcal{K})$ に，右から合成によって作用する：$[h : N \to \mathcal{K}]\sigma := [h \circ \sigma]$。

補題 2.1 N はサイズが n とする。

(1) 2 つのデータセット $[h : N \to \mathcal{K}]$, $[h' : N' \to \mathcal{K}]$ について次の条件は同値である：

(a) スライスカテゴリー $\mathbf{set}/\mathcal{K}$ において $[h] \cong [h']$.

(b) 全単射 $\pi : N \to N'$ があって, $h = h' \circ \pi$.

(c) $\mathsf{tab}[h] = \mathsf{tab}[h']$.

(2) $\mathsf{tab} : \mathsf{DS}(N;\mathcal{K}) \to \mathsf{TAB}(n;\mathcal{K}); [h] \mapsto \mathsf{tab}[h]$ は，全単射

$$\mathsf{DS}(\boldsymbol{c})/S_N \cong \mathsf{TAB}(N;\mathcal{K})$$

を誘導する。とくに $\mathsf{DS}(\boldsymbol{c})$ は S_N-軌道になっている。

(3) $\boldsymbol{c} = (c_\kappa) \in \mathsf{TAB}(N;\mathcal{K})$ で $[h] \in \mathsf{DS}(\boldsymbol{c})$ とする。このとき

$$S_h\backslash S_N \to \mathsf{DS}(\boldsymbol{c}); S_h\pi \mapsto [h \circ \pi] \qquad (1)$$

は全単射である。ここで S_h は $[h] \in \mathbf{set}/\mathcal{K}$ の自己同形群である。さらに

$$S_h := \{\pi \in S_N \mid h \circ \pi = h\} \qquad (2)$$

(4) $|\mathsf{DS}(\boldsymbol{c})| = n! \Big/ \prod_{\kappa\in\mathcal{K}} c_\kappa!$

注意. (1) S_h はヤング部分群である。対応する $n = |N|$ の分割は $\mathsf{tab}[h]$ である。

(2) この程度のことにカテゴリー論を持ち出すこともないのだが，$\mathcal{K}$ が順序集合や直積集合のように構造を持っている場合，一般論を用意しておくのが便利である。

2.2 2 元データセット

2 元データセットは，(1 元) データセットの対 $([f : N \to \mathcal{L}], [g : N \to \mathcal{M}])$ (普通は $[f,g]$ のように書く) のことである。$\mathsf{DS}(N;\mathcal{L},\mathcal{M}]$ によってそのような 2 元データセットの集合を表す。$[f,g]$ と $[\langle f,g\rangle : N \to \mathcal{L}\times\mathcal{M}]$ を対応させることにより，2 元データセットは 1 元データセットと 1 対 1 に対応する：

$$\mathsf{DS}(N;\mathcal{L},\mathcal{M}) \cong \mathsf{DS}(N;\mathcal{L}\times\mathcal{M}]$$

対称群の直積 $S_N \times S_N$ が $\mathsf{DS}(N;\mathcal{L},\mathcal{M}]$ に作用する：$[f,g](\sigma,\tau) := [f\circ\sigma, g\circ\tau]$。他方 1 元データセットへの S_N の作用を考えると，$[f,g]\pi := [f\circ\pi, g\circ\pi]$ となっているが，これは対角作用である：$[f,g]\pi = [f,g](\pi,\pi)$。しかも $[f,g],[f',g'] \in \mathsf{DS}(N;\mathcal{L},\mathcal{M})$ が同型であることと，それらが S_N の対角作用により同じ軌道に含まれることは同値である。

2.3 分割表と周辺分布

$\mathsf{TAB}(n;\mathcal{L},\mathcal{M}) := \mathsf{TAB}(n;\mathcal{L}\times\mathcal{M})$ と書く。$\boldsymbol{x} = (x_{\lambda,\mu}) \in \mathsf{TAB}(n;\mathcal{L},\mathcal{M})$ に対し，

$$x_{\lambda+} := \sum_{\mu\in\mathcal{M}} x_{\lambda\mu},\ x_{+\mu} := \sum_{\lambda\in\mathcal{L}} x_{\lambda\mu},$$
$$x_{++} := \sum_{\lambda,\mu} x_{\lambda\mu} = \sum_{\lambda} x_{\lambda+} = \sum_{\mu} x_{+\mu}$$

と置く。$n := x_{++}$ を $\boldsymbol{x}$ のサイズという。$(x_{\lambda+}) \in \mathsf{TAB}(n;\mathcal{L})$ と $(x_{+,\mu}) \in \mathsf{TAB}(n;\mathcal{M})$ を**周辺分布** (marginal distribution) という。

次の可換図式を得る：

$$\begin{array}{ccc} & & \mathsf{DS}(N;\mathcal{L},\mathcal{M}) \\ & & \| \\ \mathsf{DS}(N;\mathcal{L}\times\mathcal{M}) & \xrightarrow{\cong} & \mathsf{DS}(N;\mathcal{L})\times\mathsf{DS}(N;\mathcal{M}) \\ \downarrow{\scriptstyle \mathsf{tab}_{\mathcal{L}\times\mathcal{M}}} & & \downarrow{\scriptstyle \mathsf{tab}_{\mathcal{L}}\times\mathsf{tab}_{\mathcal{M}}} \\ \mathsf{TAB}(n;\mathcal{L}\times\mathcal{M}) & \xrightarrow{\mathsf{mar}} & \mathsf{TAB}(n;\mathcal{L})\times\mathsf{TAB}(n;\mathcal{M}) \end{array}$$

ここで $\mathsf{mar}:(x_{\lambda\mu})\mapsto((x_{\lambda+}),(x_{+\mu}))$ である。また $\mathsf{tab}[f,g]=(|f^{-1}(\lambda)\cap g^{-1}(\mu)|)_{\lambda,\mu}$ はデータセット $[f,g]$ の分割表 (contingency table) である。

$(\boldsymbol{a},\boldsymbol{b})\in\mathsf{TAB}(n;\mathcal{L})\times\mathsf{TAB}(n;\mathcal{M})$ に対し

$$\mathsf{DS}(\boldsymbol{a},\boldsymbol{b}):=\mathsf{mar}^{-1}(\boldsymbol{a},\boldsymbol{b})$$

と置く。これは，周辺分布が $\boldsymbol{a},\boldsymbol{b}$ であるような分割表の集合である。

補題 2.2 $[f,g],[f',g']\in\mathsf{DS}(N;\mathcal{L}\times\mathcal{M})$ で $|N|=n$ とする。

(1) $[f,g],[f',g']$ が同じ周辺分布を持つ，すなわち $\mathsf{mar}[f,g]=\mathsf{mar}[f',g']$)，ための必要十分条件は，$\sigma,\tau\in S_N$ で，$f=f'\sigma, g=g'\tau$ を満たすものが存在することである。とくに $S_N\times S_N$ は $\mathsf{DS}(,\boldsymbol{b})$ に可移に作用する。

(2) $\mathsf{tab}[f,g]=\mathsf{tab}[f',g']$ であるための必要十分条件は，$\pi\in S_N$ で $f=f'\pi, g=g'\pi$ を満たすものが存在すること（同じことだが，$\mathbf{set}/\mathcal{L}\times\mathcal{M}$ において同型なこと）である。

(3) $\sigma,\tau,\pi\in S_N$ に対し，

$$\mathsf{tab}[f\circ\pi,g\circ\pi]=\mathsf{tab}[f,g],$$
$$\mathsf{tab}[f\circ\sigma,g\circ\tau]=\mathsf{tab}[f\circ\sigma\tau^{-1},g]$$

系 2.3 $[f,g]\in\mathsf{DS}(N;\mathcal{L},\mathcal{M})$ に対し $\pi\mapsto[f\pi,g]$ と $(\sigma,\tau)\mapsto[f\sigma,g\tau]$ は次の一対一対応を引き起こす：

$$S_{f_0}\backslash S_N/S_{g_0}\xrightarrow{\cong}(S_f\times S_g)\backslash(S_N\times S_N)/S_N^{\mathrm{diag}}\xrightarrow{\cong}\mathsf{TAB}(\boldsymbol{a},\boldsymbol{b})$$

系 2.4 (1) $\mathsf{tab}\times\mathsf{tab}$ による $\mathsf{DS}(\boldsymbol{a},\boldsymbol{b})$ 上の一様分布の像は超幾何分布である：

$$\mathrm{Prob}(\mathsf{tab}[f,g]=\boldsymbol{x})=\frac{\boldsymbol{a}!\,\boldsymbol{b}!}{n!\,\boldsymbol{x}!}=:H(\boldsymbol{x})$$

ここで $\boldsymbol{x}!=\prod_{\lambda\mu}x_{\lambda,\mu}!$ など。

(2) $[f,g]\in\mathsf{DS}(\boldsymbol{a},\boldsymbol{b})$ とする。このとき $(\sigma,\tau)\mapsto\mathsf{tab}[f\sigma,g\tau]$ による $S_N\times S_N$ 上の一様分布の像は同じ超幾何分布である。同じことだが，

$$\frac{1}{n!^2}\sharp\{(\sigma,\tau)\mid\mathsf{tab}[f\sigma,g\tau]=\boldsymbol{x}\}=H(\boldsymbol{x})$$

(3) $\sigma\mapsto\mathsf{tab}[f\sigma,g]$ による S_N 上の一様分布の像も同じ超幾何分布である。

$N=\{1,2,\cdots,n\}$, $\boldsymbol{a}\in\mathsf{TAB}(n;\mathcal{L}),\boldsymbol{b}\in\mathsf{TAB}(n;\mathcal{M})$ とする。$[f,g]$ を $DS(\boldsymbol{a},\boldsymbol{b})$ から取っておく。例えばこれは観測データから得られたデータセットでよい。対称群 S_n の元 $\sigma_1,\sigma_2,\cdots$ をランダムに取ってゆくと $\mathsf{tab}[f\sigma_1,g],\mathsf{tab}[f\sigma_2,g],\cdots$ は $TAB(\boldsymbol{a},\boldsymbol{b})$ のランダムウォークである。生起確率は超幾何分布にしたがう。これで分割表のサンプリングが得られる。対称群の元のランダムサンプリングが問題になるが，Diaconis の方法では，ランダムに互換 $\tau_1,\tau_2,\cdots$ を取ってそれを順に掛けて行く：$1,\tau_1,\tau_2\tau_2,\cdots$。ただし奇置換と偶置換が交互に入れ替わるので，このランダムウォークはいわゆるエルゴード性を満たしておらず，定常分布を持たない。幸いなことに $\sigma\mapsto\mathsf{tab}[f\sigma,g]$ で誘導される分割表のウォークは，まれな例外（周辺度数 $x_{\lambda+},x_{+\mu}$ がつねに 0 か 1）を除いてエルゴード性を満たし，したがって超幾何分布に収束するランダムウォークである。

ランダムな互換 τ によって $[f,g]$ から $[f,\tau,g]$ を作ることは，分割表 $\mathsf{tab}[f,g]$ のレベルでいうと，$\mathsf{tab}[f,g]$ に $B=\begin{bmatrix}-1 & 1\\ 1 & -1\end{bmatrix}$ を加えること（非負成分は 4 つ）である。したがってこの遷移法則は完全に分割表だけで記述できる。もとのデータセットの取り方に依らない。

この行列 B は，MCMC 法の有名な行列（マルコフ基底，基本移動 basic move）である。収束は遅い（収束率 $(n-3)/(n-1)$ は 1

に近い)。不思議なことに，対称群上の RW から誘導される分割表たちの RW について書いた文献が見つからない。Diaconis など両方の研究で重要な成果を上げているのにである。しかし考えをさかのぼれば，両者を結びつける方法はフィッシャーの並べ替え検定 (permutation test) そのものである。フィッシャーの時代 (1930 年代) は計算機もなく実用性はほとんどなかった。今では再評価されている。このようなリサンプリング法の流れの先にブートストラップ法がある。代数的に解釈するなら，並べ替え検定で対称群を対称半群に取り替えものである。

3 一致数のモーメント公式

3.1 一致数の平均値公式

対角データセット $\langle f,g\rangle : N \to \mathcal{L}\times\mathcal{L}$ の等化 (equalizer) を

$$\mathsf{Eq}[f,g] := \{i \in N \mid f(i) = g(i)\} \subseteq N \quad (3)$$

で定義し，一致数をそのサイズとする:

$$x[f,g] := \sharp\{i \in N \mid f(i) = g(i)\} \quad (4)$$

一致数は，$[f,g]$ の分割表の対角和である：

$$x[f,g] = \mathrm{Tr}(\mathsf{tab}[f,g]) \quad (5)$$

問題は，与えられた周辺分布 $\boldsymbol{a},\boldsymbol{b}$ を持つデータセット $[f',g']$ についての一致数 $x[f',g']$ の分布である。まず平均は補題 2.2 により次のように表せる：

$$m := \frac{1}{n!}\sum_{\pi\in S_n} x(\pi), \quad x(\pi) := x[f, g\circ\pi]$$

一致数の平均について次の公式がある：

定理 3.1（平均値公式）

$$m = \frac{1}{n}\sum_{\lambda\in\mathcal{L}} a_\lambda b_\lambda \quad (6)$$

定理の証明。対称群 S_n の可移性から，任意の $i,j \in N$ に対し，

$$\sharp\{\pi \in G \mid \pi(i) = j\} = n!/n = (n-1)!$$

に注意しておく。

$$\begin{aligned}\sum_{\pi\in G} x[f\ ,\ g\pi] &= \sharp\{(\pi,i) \mid f(i) = g(\pi(i))\}\\ &= \sharp\{(\pi,i,j) \mid f(i) = g(j), \pi(i) = j\}\\ &= \sharp\{(i,j) \mid f(i) = g(j)\} \times (n-1)!\\ &= \frac{n!}{n}\sum_{\lambda\in\mathcal{L}} |f^{-1}(\lambda)|\cdot|g^{-1}(\lambda)|\\ &= \frac{n!}{n}\sum_{\lambda\in\mathcal{L}} a_\lambda b_\lambda\end{aligned}$$

3.2 モーメント公式

データセット $[f,g] \in \mathsf{DS}[N;\mathcal{L}.\mathcal{L}]$ の周辺度数分布を

$$\begin{aligned}&\boldsymbol{a} = (a_\lambda), a_\lambda = |f^{-1}(\lambda)|,\\ &\boldsymbol{b} = (b_\mu), b_\mu = |g^{-1}(\mu)|\end{aligned}$$

とする。

$$x(\pi) := \mathsf{tab}[f, g\circ\pi], \quad \pi \in S_N$$

とおく。

定理 3.2（モーメント公式）

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

左辺の和を組合せモーメントという。とくに $t=1$ の場合が平均値公式である。

この公式の証明は，ちょっとテクニックを要する。まず平均値公式を，対称群の代わりに N に回移に作用する有限群 G の場合に拡張する。N の代わりに N^T, t-点集合 T から N への単射全体の集合) に平均値公式を適用する。これによってモーメント公式が得られる。

系 3.3 分散は次で与えられる：

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

4 一致数に対する正確な p-値

4.1 超幾何多項式による表現

モーメント公式の存在は，$x(\pi)$ の分布が周辺分布 $\boldsymbol{a},\boldsymbol{b}$ によって完全に決まること

を意味する。したがって正確な p-値の計算公式があるはずである。実際モーメント公式より

$$\frac{1}{(n-k)!}\sum_{\pi\in S_n}\binom{x(\pi)}{k}=\sum_{\Sigma k_\lambda=k}\prod_\lambda\binom{a_\lambda}{k_\lambda}\binom{b_\lambda}{k_\lambda}k_\lambda!$$

多項式型母関数を取ると

$$\begin{aligned}F(z)&:=\sum_{k\geq 0}\frac{1}{(n-k)!}\sum_{\pi\in S_n}\binom{x(\pi)}{k}z^k\\&=\prod_\lambda F_{a_\lambda,b_\lambda}(z)\end{aligned}$$

ここで，$F_{a,b}(z)$ は ${}_2F_0$-型超幾何多項式である。

$$F_{a,b}(z):={}_2F_0(-a,-b;z).$$

a,b が非負整数のときのみ多項式で，それ以外の場合は，収束半径 0 の整級数である。

一致数が x に等しくなる確率は

$$\begin{aligned}p(x)&=\mathrm{Prob}(\mathsf{Tr}(\mathsf{tab}[f,g])\\&=\frac{\sharp\{\pi\in S_n\mid x(\pi)=x\}}{n!}\end{aligned}$$

$F(z)$ の展開式

$$F(z)=\sum_{k\geq 0}\binom{n}{k}k!q(k)z^k$$

から数列 $\{q(k)\}$ を求め，最後に求める確率 $\{p(x)\}$ を:

$$p(x)=\sum_{k=x}^{n}(-1)^{k-x}\binom{k}{x}q(k)$$

で求める。正確な p-値は

$$P(x)=\mathsf{Prob}(x\geq x_0)=\sum_{x\geq x_0}p(x)$$

で求まる。

なお，

$$q(k)=\sum_{x\geq k}\binom{x}{k}p(x)$$

である。母関数で表すなら，

$$\sum_{k=0}^{n}p(k)z^k=\sum_{k=0}^{n}q(k)(z-1)^k$$

あるいは，

$$\sum_{k=0}^{n}P(k)z^k=1+z\sum_{k=1}^{n}q(k)(z-1)^{k-1}$$

とも表現できる。高速の多項式計算ができる数式処理システムがあるなら，多項式のままプログラムするのが楽である。

前稿で述べた日本語・アイヌ語朝鮮語の正確な一致数はここで述べた公式を使った。そこで述べたように，得られた確率は二項確率より小さい値を取るが，言語によっては二項確率の方が小さい値を取ることもある。

4.2 三つのデータセットの一致数

3 元データセット $[f,g,h:N\to\mathcal{L}]$ の一致数はいろいろ考えられる。

まず，二つのデータセットに対する一致数 x_{fg},x_{gh},x_{hf} から得られる一致数

$$x^I_{fgh}:=x_{fg}+x_{gh}+x_{hf}$$

がある。奇妙なことにこの一致数に対する分散は

$$s'^2_{fgh}=s'^2_{fg}+s'^2_{gh}+s'^2_{hf}$$

を満たす。正規分布などでよく見る式だが，なぜこのようなきれいな式があるのかはよく分からない。残念ながら，この一致数に対しては正確な確率を求める良い公式はないようだ。

またもうひとつの一致数

$$x^{\mathrm{II}}_{fgh}:=\sharp\{i\in N\mid f(i)=g(i)=h(i)\}$$

については ${}_3F_0$-型超幾何多項式を使った公式がある。

4.3 言語学からのさらなる要請

シフト法の方式は様々なものが考えられる。言語群の比較。ふたつの語族（あるいは言語の集合）$A:f_1,f_2,\cdots$ と $B:g_1,g_2,\ldots$ が与えられているとする。このとき語族の一致数を

$$x_{A,B}:=\sum_\alpha\sum_\beta x[f_\alpha,g_\beta]$$

で定義する。このときシフト法による平均と分散は

$$m_{A,B} = \sum_{\alpha,\beta} m[f_\alpha, g_\beta],$$
$$s^2_{AB} = \sum_{\alpha,\beta} s[f_\alpha, g_\beta]^2$$

で与えられる。語族間の距離を測るのに使えそうだ。ただ方法論的に改良の余地がある。

そのほか，日本語からアルタイ系の言語の影響を除いて，そこに南方系の言語の影響がどの程度あったかを探るのは今後の課題である。

5 2×2 分割表の一致数と独立性

周辺度数の決まっている 2×2-分割表は，$(1,1)$-成分（下図の a），または一致数 $a+d$ のどちらか一方を与えるだけで，すべての成分が決まる。

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

上の分割表において

$$A := (a+b)(a+c),$$
$$B := (a+b)(b+d),$$
$$C := (a+c)(c+d),$$
$$D := (b+d)(c+d).$$

とおく。このとき

$$a+b+c+d = n,$$
$$A+B+C+D = n^2$$

である。この分割表のカイ二乗統計量とカッパ係数は

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)},$$
$$\kappa := \frac{(a+d)/n - A/n^2 - D/n^2}{A/n^2 + D/n^2}$$

で定義されている。簡単な計算で

$$2(ad-bc) = \kappa(B+C)$$

となり，結局次の定理を得る：

定理 5.1 (1) $\chi^2 = n\dfrac{(B+C)^2}{4BC}\kappa^2 \geq n\kappa^2$. 等号成立は $b=c$ のときに限る。
(2) 一致数（したがって κ）の正確な p-値が分かれば，χ^2 の正確な p-値も分かる。逆も言える。

参考文献

[1] 日比『グレブナー道場』共立（2011）.
[2] Agresti, A., “An Introduction to Categorical Data Analysis”, Wiley, 2007.
[3] Diaconis, P. and Gangolli, A. , Rectangular arrays with fixed margins, in “Discrete Probability and Algorithms（D. Aldous et al., eds.）”, 15-41, Springer, New York, 1995.
[4] L.Prachter, B.Sturmfels (編), “Algebraic Statistics for Computational Biology,”Cambridge, 2005.
[5] P.Diaconis, “Group Representaions in Probability and Statistics,” LNMonograph series 11, Institute of Math.Stat., 1988.
[6] T.Ceccherini-Silberstein, etc.,“Harmonic Analysis on Finite Groups,” Cambridge, 2008.
[7] L.Salloff-Coste, Random Walks on Finite Groups, in Encyclopaedia of Math.Sci., 264-346, 2003.
[8] 吉田知行「比較言語学における数学的方法」『北星論集』（北星学園大学経済学部）56（2017），121－135.