統計数理（2003）
第 51 巻 第 2 号 261–295

特集「個票開示問題の統計理論」
［総合報告］

# 孤立個体数の推測

渋谷 政昭†

（受付 2003 年 2 月 4 日；改訂 2003 年 9 月 19 日）

目　　次



† 高千穂大学 経営学部：〒 168–8508 東京都杉並区大宮 2–19–1; sibuyam@takachiho.ac.jp

## 要　　旨

分類変量の分類数が非常に多く，各分類の確率よりは，確率全体の特徴が重視される分野がある．生態学における種の多様性，言語学における語彙，考古学における遺物類のパターン，などが典型例である．標本調査における個人データ保護もこれに含められる．

母集団個体の質的な属性に注目し，量的属性は区分して質的属性と同一視する．個体の識別子を除いて多重分割度数表に集約する．分割表の多重度が大きいとセルの数が多くなり，標本の大きさに匹敵し，超えることもある．

本稿では “母集団および標本で孤立している個体数の推測” という課題を議論する．標本の観測度数が 1 のセルがいくつかあるとき，そのなかで母集団の度数も 1 のものがいくつあるか，標本だけから予測したい．

最初にこの数を，調査データを公有化するときに生ずる個体データ漏洩危険の尺度として用いることを議論する．次に多数カテゴリーの多様性の統計学で，この課題が占める役割について議論し，この分野の主要成果を概観する．最後に最近の研究の成果と現在の方向を展望する．本文中の特殊な話題を付録で補足する．

キーワード： ジッフ法則，寸法指標，多数希少事象，多様性モデル，ミクロ統計の公有化，無限分解可能確率母関数．

## 1. まえがき

分類変量(categorical data)で分類数が非常に多く，各分類の頻度・比率による確率の推測よりは，分類確率の全体の状況が重視される分野がある．生態学における種の多様性，言語学における語彙，考古学における遺物類のパターン，などの研究が典型的である．標本調査における個人データ保護もこれに含めることができる．

母集団の個体についていくつかの質的な属性(attributes)に注目する．量的属性は区間に分けて質的属性と同一視する．個体の識別子(ID)を除くと，母集団が多重分割度数表に集約される．属性の種類が多い，つまり分割表の多重度が大きいと分類組合わせ(セル)の数が非常に多くなり，母集団の大きさに匹敵し，標本の大きさを超えることもある．以下の議論では特に断らない限り，分類変量の順序と分割表構造を問題とせず，単純な分類変量として議論する．

本稿では “母集団および標本で孤立している個体数の推測” という特殊な課題を議論する．標本の観測度数が 1 のカテゴリーがいくつかあるとき，そのカテゴリーの母集団の度数も 1 であるものがいくつあるか，を標本だけから予測したい．

最初にこの数が，調査データを公有化するときに生ずる個体データ漏洩危険の尺度としての役割を議論し，この課題が困難であることを示す(第 2 節)．

次に多数カテゴリーの多様性の議論で，この課題が占める役割について議論し，この分野におけるこれまでの成果をまとめる．諸種の Zipf 法則，特にカーリン・ルオー中心極限定理，多数希少事象を紹介する(第 3 節)．

最後に最近の研究の成果と現在の方向を展望する．特に無限分解可能離散分布の役割を議論

する(第 4 節).

本文中の特殊な話題を付録として補足する(第 5 節).

なお推測の具体的な方法は他の論文に譲り，モデルの構成を中心に議論する.

## 2. 孤立個体数

### 2.1 調査データ公有化における個体データの漏洩管理

本特集号のテーマは，標本調査の未加工データ，つまり “個票”あるいは “ミクロ統計”と呼ばれるものを，調査目的から外れた副次的解析に利用するために，調査主体の管理を離れた公共のものとするとき，被調査個体(個人，世帯，事業所など)の秘密を守りながら，データの情報をできるだけ活用する方法論である.

古典的な統計的方法では集団を観測するとき，もっぱらその中心，典型を集団の特徴とみなしている．個体の特徴を表わす数量であればその算術平均など，測定値を縮約，要約する統計量が重要である．そのために統計データの公表は層に分けた上で平均，比率を 2 重表で表示することが多い．このような統計量を公表する限り，個人データ漏洩の危険は比較的少ない.

種々のモデルを構築するためにできるだけ未加工の詳細なデータを必要とすることと，個体データを秘匿することは，対立する要求である．“木を見て森を見ず”というたとえがあり，統計調査の目的が集団を理解するためであって，集団を構成する個体の属性ではないことを強調するためにも引用される．しかしながら集団の理解は個体の観測，測定から出発するし，集団を調べれば，その中にはきわだった特徴をもつ個体が多数個存在する.

個体の秘密が漏洩しないように，特異な個体が存在する事実をできるだけ損わず，しかもできるだけ多くの人が利用できるようにするのが “ミクロ統計の漏洩管理(Statistical Disclosure Control)”である．たとえば Willenborg and de Waal (1996，2001)，渋谷(1999)を参照.

### 2.2 漏洩の危険を測る尺度

被調査個体の秘密漏洩で影響が大きいのは個人の場合である．個人の秘密データを保護するとき，匿名を保証することと，それと関連した秘密データをかくすことの 2 側面がある．たとえば年齢，職業，住所の地域，とともに高額所得が記載されていれば個人が識別されやすい．すべての項目の分類を粗くするのが一つの方法であるが，それは副次的分析のための情報を損うことになる．高額所得という，隠したい，あるいは誇示したくない，データを，ある金額以上とあいまいにすることになる.

なま(生)の加工されていないデータにたいして，欠測，グループ化，など変換されたデータすべてを “不完全データ incomplete data”という概念にまとめ，失われたデータを復元する研究がある．秘匿と正反対の研究であり，両刃の剣である．不完全データ分析で “この地域に，ある金額以上の所得者が何人いる”という推測を高い確率で行えるとしたとき，これが個人の秘密をおびやかすとも言えるが，危険を測りにくい.

そのため本稿では氏名，住所，電話番号などを消去して匿名としたつもりの被調査者の氏名が公表した個票データから識別できる “再識別 re-identification”だけを危険とする.

このように限定しても，悪意をもって再識別しようと試みる “侵入者 intruder”が，特定個人の属性について，個票データと対照できる項目をいくつ知っているかによって可能性が大きく変わる．そもそもある特定個人が公開されるデータセットに入っていることを確信していることと，データを分析していて特異なデータに気付くなど，では “危険のシナリオ”が違う.

諸個人についてのデータベースにアクセスでき，それと個票データと対照することにより，できるだけ多くの特異な個人を見出そうという IT 技術者の犯罪がもっとも危険である．データベースのなかの特異な現象を発見する “データマイニング”は，個人データ保護と逆向きの仕

事を目的としており，方法論に共通することが多いであろう．

侵入者が再識別に役立つ項目のデータを母集団全部について知っているという極端な場合を想定し，公開された個票データから本人を再識別できる確率により，漏洩の危険度を計るのが本論文の課題である．その根拠は問題の定式化が明確であること，それでもなお難問であること，にある．

したがって，この尺度を適切に評価できたとして，それをどのように解釈するかは別問題であること，あくまで一つの尺度であることを強調しておきたい．

### 2.3 推測の困難

#### 2.3.1 有限母集団モデルと寸法指標

大きさ $N$ の母集団の個体が $K$ 種のカテゴリー(category，あるいはセル cell，類別 homology)に分けられ，各カテゴリーに属する個体数が $M_k \geq 0,\ k = 1, \ldots, K,$ であるとする．この母集団から単純非復元確率抽出により得た，大きさ $n$ の標本の分類変量を $X = (X_1, \ldots, X_K)$ とする．これは多変量超幾何分布に従う:

$$P\{X = (x_1, \ldots, x_K)\} = \binom{M_1}{x_1} \cdots \binom{M_K}{x_K} \Big/ \binom{N}{n} = \frac{n!(N-n)!}{\prod_k x_k!(M_k - x_k)!} \Big/ \frac{N!}{\prod_k M_k!} = \binom{n}{x_1, \ldots, x_K} \frac{\prod_{k=1}^K M_k^{\underline{x_k}}}{N^{\underline{n}}}. \tag{2.1}$$

ただし $N = \sum_k M_k, n = \sum_k x_k; N^{\underline{n}} = N(N-1)\cdots(N-n+1)$ である．

もしも各分類の意味を無視して，$\{M_1, \ldots, M_K\},\ \{x_1, \ldots, x_K\}$ を集合として考えるならば，これらの離散的な順序統計量を考えることになる．大きい方に関心があれば降順に，逆ならば昇順に並べる．さらにそれを見やすくするために，母集団および標本における，個体数が $\nu$ のセルの数を，それぞれ $T_\nu$ および $S_\nu$ とする．つまり，述語(predicate) I[·] ([ ] で包まれる事象が生じれば 1，そうでなければ 0 の値をとる)を用いると

$$T_\nu = \sum_{k=1}^K \mathrm{I}[M_k = \nu], \quad S_\nu = \sum_{k=1}^K \mathrm{I}[X_k = \nu]. \tag{2.2}$$

$T = (T_0, \ldots, T_N), S = (S_0, \ldots, S_n)$ をそれぞれ $M = (M_1, \ldots, M_K), X = (X_1, \ldots, X_K)$ の“寸法指標(size index)”と呼ぶ．頻度の頻度(frequency of frequencies)あるいは partition vector, frequency spectrum と呼ぶ人もいる．$T_\nu$ は，母集団の特性をあらわす定数であるのにたいして，$S_\nu$ は確率変数である．

$K$ が非常に大きければ，個人の母数 $M_k, k = 1, \ldots, K$ に関心はなく，それを要約した $T_\nu, \nu = 1, 2, \ldots$ あるいは，さらに要約した量(パラメータ関数)に関心がある．3.4 節「種々の推測問題」でさらに議論する．本論文で，標本でも母集団でも孤立している個体の数，つまり

$$U := \sum_{k=1}^K \mathrm{I}[M_k = X_k = 1], \tag{2.3}$$

を予測すること，あるいは $E(U)$ を推定することを問題とする．後で $E(U) = T_1 n/N$ を示す．

利用できるデータは標本 $(X_1, \ldots, X_K)$ であるが，$(S_1, S_2, \ldots)$ が十分統計量となる．

寸法指標は

$$T_0 + T_1 + \cdots = S_0 + S_1 + \cdots = K, \quad \sum_\nu \nu T_\nu = N, \quad \text{and} \quad \sum_\nu \nu S_\nu = n, \tag{2.4}$$

の制約条件を満たしている．$T_1$ および $S_1$ はそれぞれ，母集団および標本における孤立個体(solitons，unique individuals)の数である．

母集団についての知識が不完全で，$K$ あるいは $T_0$ が未知のこともある．つまり正の確率をもつカテゴリーの数が不明の場合もある．$\max\{X_1,\dots,X_k\}$ がとる値の最大値を $\mu$ とする．つまり $\mu=\max\{\nu:\ P\{S_\nu>0\}\}$ とする．明らかに，$\mu=\min(n,\max_j M_j)=\min(n,\max\{\nu:T_\nu>0\})$ である．

もしも $n$ が $N$ に比べて大きければ，$\{\nu:E(S_\nu)>0\}$ の下限 $\max(0,\min_j(n+M_j-N))$ が存在するが，通常の標本調査では $n\ll N$ である．

### 2.3.2 素朴な推定量

定義式 (2.2) から，

$$
\begin{aligned}
(2.5)\qquad E(S_\nu)&=\sum_{k=1}^{K}P\{X_k=\nu\}=\sum_{k=1}^{K}\binom{M_k}{\nu}\binom{N-M_k}{n-\nu}\Big/\binom{N}{n}\\
&=\sum_{\lambda=1}^{n}T_\lambda\binom{\lambda}{\nu}\binom{N-\lambda}{n-\nu}\Big/\binom{N}{n}=\sum_{\lambda=1}^{n}T_\lambda\binom{n}{\nu}\binom{N-n}{\lambda-\nu}\Big/\binom{N}{\lambda}\\
&=\sum_{\lambda=1}^{n}T_\lambda\binom{\lambda}{\nu}\frac{n^{\underline{\nu}}(N-n)^{\underline{\lambda-\nu}}}{N^{\underline{\lambda}}},\quad n^{\underline{\nu}}=n(n-1)\cdots(n-\nu+1),\quad \nu=0,1,\dots
\end{aligned}
$$

これは和の計算順序の変更でしかない．結局 $(E(S_0),E(S_1),\dots)$ は $(T_0,T_1,\dots)$ の一次変換であり，$\mu$ が $S_\nu$ および $T_\nu$ の添字 $\nu$ の上限であったから，

$$
(2.6)\qquad \begin{bmatrix}E(S_0)\\E(S_1)\\\vdots\\E(S_\mu)\end{bmatrix}=W\begin{bmatrix}T_0\\T_1\\\vdots\\T_\mu\end{bmatrix},
$$

ただし $W$ は次のような $\mu+1$ 次の正則行列である．

$$
W=\begin{bmatrix}
1 & (N-n)/N & (N-n)^{\underline{2}}/N^{\underline{2}} & (N-n)^{\underline{3}}/N^{\underline{3}} & \dots\\
0 & n/N & 2n(N-n)/N^{\underline{2}} & 3n(N-n)^{\underline{2}}/N^{\underline{3}} & \dots\\
0 & 0 & n^{\underline{2}}/N^{\underline{2}} & 3n^{\underline{2}}(N-n)/N^{\underline{3}} & \dots\\
0 & 0 & 0 & n^{\underline{3}}/N^{\underline{3}} & \dots\\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}.
$$

$W$ の $(k+1,\nu+1)$ 要素が (2.5) の $T_\nu$ の係数である．方程式 (2.6) を解けば，$(S_\nu,\dots,S_\mu)$ の 1 次式で $(T_\nu,\dots,T_\mu)$ が定まる．$\nu=1$ として素朴な推定量，$(\hat{T_1},\hat{T_2},\dots,\hat{T_\mu})$ が得られる．これは $(S_1,\dots,S_\mu)$ の一次式で，不偏推定量である．$W$ の正則性から，一意的な線形不偏推定量である．

**命題 2.1.** 逆行列 $W^{-1}$ の $(k+1,\nu+1)$ 要素は

$$
\binom{\nu}{k}\frac{N^{\underline{\nu}}(-1)^{\nu-k}(N-n)^{\overline{\nu-k}}}{n^{\underline{k}}},\quad N^{\overline{\nu}}=N(N+1)\cdots(N+\nu-1)
$$

である．

$W$ の $(k+1,\nu+1)$ 要素と比較すると，比 $n^{\underline{k}}/N^{\underline{\nu}}$ が逆転しており，負符号が網目状に入っている．証明は 2 項係数の反転公式を用いる．

次の $W$ は $N=100$, $n=20$, $\mu=6$ の場合の $7\times 7$ 変換行列である．$W_1^{-1}$ はその第 1 行，1 列を除いた $6\times 6$ 行列の逆行列で推定量の係数行列である．

$$W=\begin{bmatrix} 1 & 0.8 & 0.63838 & 0.5081 & 0.40334 & 0.31931 & 0.25209 \\ 0 & 0.2 & 0.32323 & 0.39085 & 0.41905 & 0.42014 & 0.40334 \\ 0 & 0 & 0.038384 & 0.094001 & 0.15312 & 0.20734 & 0.25209 \\ 0 & 0 & 0 & 0.0070501 & 0.023258 & 0.047849 & 0.078572 \\ 0 & 0 & 0 & 0 & 0.0012356 & 0.0051483 & 0.012844 \\ 0 & 0 & 0 & 0 & 0 & 0.00020593 & 0.0010405 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.2515e-005 \end{bmatrix}$$

$$W_1^{-1}=\begin{bmatrix} 5 & -42.105 & 284.21 & -1827.9 & 11853 & -79649 \\ 0 & 26.053 & -347.37 & 3310.2 & -28275 & 234680 \\ 0 & 0 & 141.84 & -2670 & 33792 & -369460 \\ 0 & 0 & 0 & 809.33 & -20233 & 327780 \\ 0 & 0 & 0 & 0 & 4856 & -155390 \\ 0 & 0 & 0 & 0 & 0 & 30755 \end{bmatrix}$$

分散共分散の陽な形は複雑となるので省略する．

統計データ保護の場合には，標本 $X=(X_1,\dots,X_K)$ より，

$$U:=\sum_{k=1}^{K}\mathrm{I}[M_k=X_k=1],$$

を予測する，あるいは

$$E(U)=\sum_{k=1}^{K}E(\mathrm{I}[M_k=X_k=1])=\sum_{k=1}^{K}P\{X_k=1|M_k=1\}=T_1 n/N,$$

を推定することが一つの課題である．命題 2.1 から

$$\widehat{U}=\frac{1}{N}\sum_{\nu}\nu N^{\underline{\nu}}(-1)^{\nu-1}(N-n)^{\overline{\nu-1}}S_\nu, \tag{2.7}$$

が一つの素朴な plug-in 推定量であることが分かる．この和を何項とるか，どのように平滑化するかなどの問題があるが，いずれにせよ推定量は良くない．

この推測問題の困難は直感的に次のように説明できる．ある数 $\nu$ が小さく $\nu n\ll N$ であると，抽出率 $n/N$ のサンプリングでは，個体数 $\nu$ のカテゴリーからはほとんど個体が抽出されず，されても 1 個である．個体数 $\nu$ のカテゴリーが $T_\nu$ 個あるとこれから平均 $\nu T_\nu n/N$ 個の個体が抽出されるが，これらはほとんど孤立個体である．したがって $\sum_{\nu\ll N/n}\nu T_\nu$ 個の個体のうちの平均 $n/N$ が標本で孤立個体となり，この数はそれぞれの $T_\nu,\nu=1,2,\dots$ の値には依存しない．したがって逆に $S_1$ から $T_\nu,\nu=1,2,\dots$ についての情報は得られない．$S_\nu,\nu=2,3,\dots$ を加えても，小さな $\nu$ にたいする $T_\nu$ についての情報は増えない．これは一種の逆問題，不適切問題(ill-posed problem)である．

$\sum_\nu \nu T_\nu/N$ は $M_k$, $k=1,\dots,K$; の経験分布関数である．これが確率的により小さいほど推測はより困難になる．

表 1. Population size index.

| $\nu$ | Pop. $T_\nu$ | A $\nu T_\nu$ | Pop. $T_\nu$ | B $\nu T_\nu$ | Pop. $T_\nu$ | C $\nu T_\nu$ |
|---|---|---|---|---|---|---|
| 1 | 23 | 23 | 14 | 14 | 1 | 1 |
| 2 | 11 | 22 | 10 | 20 | 2 | 4 |
| 3 | 8 | 24 | 10 | 30 | 2 | 6 |
| 4 | 6 | 24 | 9 | 36 | 2 | 8 |
| 5 | 4 | 20 | 6 | 30 | 3 | 15 |
| 6 | 4 | 24 | 4 | 24 | 3 | 18 |
| 7 | 3 | 21 | 3 | 21 | 4 | 28 |
| 8 | 3 | 24 | 2 | 16 | 6 | 48 |
| 9 | 2 | 18 | 1 | 9 | 8 | 72 |
| | 64 | 200 | 59 | 200 | 31 | 200 |

表 2. Sample size index $S_\nu$ (3 samples of size 50 from 3 populations).

| $\nu$ | | Pop. | A | | Pop. | B | | Pop. | C |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 37 | 32 | 35 | 27 | 27 | 33 | 8 | 8 | 6 |
| 1 | 14 | 18 | 16 | 21 | 20 | 10 | 7 | 7 | 12 |
| 2 | 6 | 11 | 7 | 6 | 7 | 11 | 9 | 9 | 5 |
| 3 | 4 | 2 | 5 | 3 | 4 | 2 | 5 | 4 | 5 |
| 4 | 3 | 1 | 0 | 2 | 1 | 3 | 0 | 2 | 2 |
| 5 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 1 |

### 2.3.3 数値例

第 1 表のような寸法指数をもつ，3 種の母集団から，大きさ 50 の標本をそれぞれ 3 回とると，その寸法指数が第 2 表のようになる．標本寸法指標から $A, B$ を区別することは困難である．

このような標本より，母集団寸法指数を推定した結果は非常に悪い．

### 2.3.4 ポアソン過程モデル

多変量超幾何モデル(2.1)で $M_k/N \to p_k,\ k = 1,\ \dots, K,\ (M_1, \dots, M_K, N \to \infty)$ のとき, 確率 $p_1, \dots,\ p_K$ の多項分布により近似できる．$M_1 = \cdots = M_K$ で等確率 $1/K$ の多項分布で近似できれば，標本寸法指標について，陽な結果を得られる．しかしわれわれの課題にたいする知見には乏しい．渋谷政昭(1997)参照.

多変量超幾何分布モデルの多項モデルによる近似は，非復元抽出の復元抽出による近似ともみなせる．そうすると，多項標本からの副標本と，直接の標本との区別はなくなる．有限母集団の場合に，大きさ $n$ の標本と，標本に含まれなかった大きさ $N - n$ の部分とを区別する $2K$ 分布表が，多項モデルでは超母集団からの，大きさ $n,\ N - n$ の独立な標本となる．さらに $nM_k/N \to \rho\lambda_k$ (したがって $M_k(1 - n/N) \to (1 - \rho)\lambda_k; K \to \infty), k = 1, \dots,\ K$ ならば,$X_k,$ $M_k - X_k$ をそれぞれ平均 $\rho\lambda_k, (1 - \rho)\lambda_k, k = 1, \dots, K$ の独立なポアソン分布で近似できる．4.1.1 節「モデルの分類」で再び議論する．

独立な，強度 $\lambda_k, k = 1, \dots, K,$ のポアソン過程があり，それぞれの出現度数を観測できるとする(marked Poisson process)．時間間隔 $(-1, 0]$ での，各過程の出現度数を $(X_1, \dots, X_K)$, $(0, t], 0 < t < \infty,$ での出現度数を $(Y_1, \dots, Y_K)$ とする．$t = (N - n)/n$ とすると，$X_k$ が標本の個体数を，$Y_k$ が標本に入らなかった母集団の個体数を表わすものとみなせる．

多項モデルと同様に，標本，母集団での孤立個体数は $U = \sum_{k=1}^{K} \mathrm{I}[X_k = 1]\,\mathrm{I}[Y_k = 0]$ である．このモデルでは

$$P\{X_k = x\} = e^{-\lambda_k}\lambda_k^x/x!, \quad x = 0, 1, \dots,$$

$$P\{Y_k = y\} = e^{-\lambda_k t}(\lambda_k t)^y / y!, \quad y = 0, 1, \ldots,$$

であるから,

$$P\{X_k = 1 \ \& \ Y_k = 0\} = E(\,\mathrm{I}[X_k = 1]\,\mathrm{I}[Y_k = 0]) = \lambda_k e^{-\lambda_k} e^{-\lambda_k t}$$
$$= \lambda_k e^{-\lambda_k} \sum_{n=0}^{\infty} \frac{(-\lambda_k t)^n}{n!} = \sum_{n=0}^{\infty} (-1)^n (n+1) t^n \frac{(\lambda_k)^{n+1}}{(n+1)!} e^{-\lambda_k}.$$

これから

$$E(U) = \sum_{j=1}^{K} E(\,\mathrm{I}[X_k = 1]\,\mathrm{I}[Y_k = 0]) \sum_{j=1}^{K} (-1)^{j-1} j t^{j-1} E(S_j)$$

となる.

素朴な推定量は再び, $E(S_j)$ を $S$ に変えたものである. 最後の交項級数の収束をよくする方法を考えるとか, 部分和の項数をいくつにするかという問題が残るが, いずれにしろ推定量の性質はよくない.

## 3. 多数カテゴリーの多様性モデル

多数個のカテゴリーにたいする度数が数えられており, これを順序統計量つまり寸法指標 size index にまとめる. 寸法指標について, 経験的な Zipf 法則およびその拡張, 修正, 解釈がある. 記述統計量としての寸法指標を, 背後の実体(entity)も含め, 生態学の用語を採って, 多様性統計 abundance statistics と呼ぶことにする. 以下多様性統計の諸モデルに関する議論である.

### 3.1 Zipf 法則

Zipf (1949)は種々の社会現象を集めて 2 種類のややあいまいな経験法則を述べた. たとえば, アメリカの都市の人口を多い順に並べる. 世界の国を面積の広い順に並べる. ある著作物中の単語をその出現度数の多い順に並べる, 等々. このとき, 比較している量と順位を両対数目盛りでプロットすると直線上に並ぶ. これが "順位と大きさの関係(rank-size relation)"の法則である. 対象とする量はもちろん正の数である. 有限の量のカテゴリーへのランダムな分割であるのか, 多くの個体がもつ属性を, 確率標本の順序統計量とみなすのか. その場合に連続な確率変数であるのか離散確率変数であるのか, そのような区別に Zipf はむとんちゃくであった.

もうひとつは, 逆に小さな量を考える. 上記の著作物中の単語を数えると, 出現度数が 1, 2, ... と小さい単語の種類が非常に多く, しかも小さいほど多い. 連続な正の量の場合には一定長区間に級別して, それぞれの級に入る個体, カテゴリーをかぞえる. このときに, 出現度数(あるいは級番号)とそれに対応するカテゴリー数(あるいは個体数)とを両対数目盛りでプロットすると, やはり直線上に並ぶ. これが "大きさと頻度の関係(size-frequency relation)"の法則である. 非常に粗い議論をすると, 順位と大きさの関係から, 大きさと頻度の関係が導かれる. Zipf 法則の簡単な紹介として, たとえば Read (1988)を参照.

この経験法則が到る所で再発見され, 未だに言語学, 情報学, 物理学の分野で発見の論文が現われている. 厳密な議論をすれば法則の意味も多様になる. 本号の別論文で議論する確率分割も Zipf 法則に含めることができる. 大きさ頻度の関係を示す諸種の分布が提案されている(付録の一般 Zipf 分布参照). 以下の第 2 節, 第 3 節では本稿の主題にたいして重要な結果を紹介する.

### 3.1.1 Zipf 分布

上記のことを形式的に述べよう．$n$ 個一組のデータを降順に並べたものを

$$x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(n)},$$

とする．そのとき，$n$ が大きければ，だいたい

$$r^{\alpha} x_{(r)} = \text{constant}, \quad \alpha > 0, \tag{3.1}$$

となり，多くの場合に，$\alpha$ は 1 に近い．これが順位と大きさの法則である．

離散と連続の区別を曖昧にしたまま，あるオブジェクトの大きさ・規模を $x$, その相対頻度を $f(x)$, $\int_0^{\infty} f(x)\,dx = 1$, とする．データセットの中の $x$ 以上のオブジェクトの数を $N(x)$ とすると

$$N(x) = n \int_x^{\infty} f(u)\,du = \text{大きさ } x \text{ の一つのオブジェクトの順位}$$

となる．(3.1) より $N(x) = K/x^{\alpha}$ だから

$$f(x) = -n^{-1} N'(x) = A\, x^{-(1+\alpha)}, \tag{3.2}$$

である．これは大きさと頻度の法則である．

多くの著者はこれを確率分割，カテゴリーが多いときの寸法指数のばらつき，と考えている．$N$ 個の個体を $K$ 個の個体に分けたときの寸法指数 $(S_1, \ldots, S_N)$ の分布，特に小さな度数にたいする $(S_1, S_2, \ldots)$ の分布を大きさと頻度の法則とみなす．あるいは 頻度の順序統計量 $X_{(1)} \leq X_{(2)} \leq \cdots$, の分布を大きさ頻度関係とみなす．あるいは，どちらを考えているか明記しない．順位と大きさの法則を問題にするならば，むしろ単位区間(有限の資源)のランダムな分割，random spacing，のモデルのほうが適切と思われるが，ここでは議論しない．寸法指数にたいする比率の漸近論を考えれば random spacing である．

$\mathcal{N} = (1, 2, \ldots)$ 上の離散確率分布としての Zipf 法則は,

$$f(x) = x^{-(1+a)}/A, \quad x = 1, 2, \ldots;\ a > 0, \quad A = \zeta(1+a) = \sum_{r=1}^{\infty} r^{-(1+a)}, \tag{3.3}$$

$\zeta$ はツェータ(zeta)関数，とみなされている．この確率分布は Zipf 分布，あるいは ツェータ分布とよばれている．Pareto 分布の離散版である．

この小節以下の部分では B.M.Hill たちが導いた Zipf 法則的な極限定理をまとめる．事前分布を導入する点では，現在の研究方向に近いが，Bose-Einstein 統計(格子単体上の一様分布)から出発すること，事前分布の導入が技巧的なこと，歴史的に古いことなどから最初に述べる．

(A)Hill (1974)

$N$ 個の個体が $K$ 個のカテゴリーに分類されるとし，第 $k$ カテゴリーの個体の数を $X_k$ とする．Bose-Einstein 統計では

$$P\{(X_1, \ldots, X_M) = (x_1, \ldots, x_M)\} = \binom{N-1}{K-1}^{-1}, \quad \forall (x_1, \ldots, x_M),\ x_k > 0,\ 1 \leq k \leq K,$$

となる．$K, N$ が確率変数で，$F_N(y) = P\{K/N \leq y \,|\, N\}$ が $N \to \infty\,(\text{in } P)$ のときに proper な分布 $F(y)$ に収束することを仮定する．このとき度数 $\nu$ のセルの割合 $S_\nu/M$ が $\Theta(1-\Theta)^{\nu-1}$ の分布に収束する．ただし，$\Theta$ は $F$ に従う確率変数である．特に $F$ がベータ分布 $\mathsf{Be}\,(a, b)$ に従うならば

$$E\,(\Theta(1-\Theta)^{\nu}) \sim a\,\Gamma(a+b)\,(\Gamma(b))^{-1}\,\nu^{-(1+a)}.$$

これは Zipf 法則である．

期待値の収束を分布収束とするために，各カテゴリーを細分化し，3 種の統計量を考える．

(i) 各カテゴリー $k$ に属する $N_k$ 個の個体が $K_k$ の小カテゴリーに分かれるとし，その最大度数を $L_k,\ k=1,\dots,\ K$ とする．$K_k/N_k$ が独立で，$N_k \to \infty\,(\text{in}\,P)$ のときにある分布 $F$ に収束することを仮定し，$(L_1,\dots,\ L_K)$ の順序統計量を考えると，これが Zipf 法則を示す．

(ii) (i) と同様であるが，各カテゴリー内の最大度数ではなく，ランダムに選んだひとつの $X_k^*,\ k=1,\dots,\ K$, の順序統計量が Zipf 法則を示す．

(iii) 全体の $\sum K_k$ 個の小カテゴリの中の度数の順序統計量が Zipf 法則を示す．

(B)Chen (1980)

$(X_1,\dots,X_K)$ が対称なディリクレ多項分布(多変量負の超幾何分布) MNgHg $(N,K,\beta)$ にしたがうとする．さらに $K$ が $N$ に依存する確率変数で $F_N(y)=P\{K/N\le y\,|\,N,\,\beta\}$ について上と同じ仮定をする．$F(y)\sim cy^{\alpha}\ (y\to 0),\ \alpha>0$, ならば

$$\lim_{N\to\infty} E\left(K^{-1}S_\nu\,|\,N,\,\beta\right)=\int_0^1 h(\nu;\beta,\theta)\,dF(\theta)=\phi(\nu),$$

$\phi(\nu)=A\nu^{-(1+a)},\ \nu\to\infty$. ただし

$$h(\nu;\beta,\theta)=\frac{\Gamma(\nu+\beta-1)}{\Gamma(\beta)\,(\nu-1)!}\left(\frac{\theta\beta}{1-\theta+\theta\beta}\right)^{\beta}\left(\frac{1-\theta}{1-\theta+\theta\beta}\right)^{\nu-1},\quad \nu=1,2,\dots,$$

さらに $S_\nu$ の漸近正規性を示せる．

(C)Hill and Woodroofe (1975)

(A)と同じ二重階層モデルで，さらに条件を加えることにより $S_\nu/M$ の Zipf 分布への分布収束を示している．

最後に extreme process から Zipf 法則が導けることを示す．

(D)Khmaladze et al. (1997)

新記録

$$(X_i)_{i=1}^{\infty}\quad \mathcal{N}_+\text{上の iid},\quad M_n:=\max_{i\le n}X_i,\quad \tau_n=\inf\{t:\ X_t=M_n\},$$

$M_n$ は時刻 $n$ における記録値，あるいは単に新記録(record)という．$\tau_n$ は時刻 $n$ の記録値が生じた時刻である．

$$S_n=\sum_{\tau_n\le i\le n}\mathrm{I}[\,M_{\tau_n-1}<X_i\le M_n\,]$$

とすると

$$P\{S_n=k\}=\frac{1}{k(k+1)}+\frac{1}{n}\mathrm{I}[\,k=1\,].$$

$\sum_{i=1}^{n}\mathrm{I}[\,X_i=M_n\,]$ は収束しない．$\sum_{\tau_n\le i\le n}\mathrm{I}[\,M_n-\epsilon<X_i\le M_n\,]$ について興味ある結果がある．

### 3.2 Karlin-Rouault 理論

確率の小さなカテゴリーが多数存在して，標本数を大きくすれば，あるいは観測時間を長くすればそれらが現われてくると考える．当然確率の小さなカテゴリーの個数についてのモデルが必要である．Karlin (1967)は，アーベル型理論を適用するために確率の系列が regular varying

であることを仮定して，中心極限定理，大数の強法則を導いた．この分野でもっとも強い結果である．これらの定理の条件は強過ぎるが，このような仮定なしに議論することは，複雑過ぎる，と Karlin は述べている．

Rouault (1978)は数理的言語理論で用いられるモデルを用いる．文を話す人，書く人の頭脳に，統語法則(syntax)に従う単語がランダムに現われ，マルコフ連鎖に従って単語が継続して文を作る，という生成文法モデルである．このようなモデルで現われる単語の出現確率が Karlin の条件を満たすことを示した．彼はまた，Karlin の理論で，度数 1，2，... の寸法指標の比率について大数法則が成り立ち，それが簡単な確率関数であることを注意した．

### 3.2.1 準備

可算集合 $\mathcal{N}=\{1,2,\dots\}$ の上の確率分布 $p=(p_n)_{n=1}^{\infty}$, $p_n\geq p_{n+1}>0$, $\sum_{n=1}^{\infty}p_n=1$, にたいして

$$\alpha(x):=\max\{j|p_j\geq 1/x\}=\sum_{n=1}^{\infty}\mathrm{I}[p_n\geq 1/x],\quad 1<x<\infty, \tag{3.4}$$

とする．言わば $p$ の上側確率 $1/x$ の確率点が $\alpha(x)$ である．これについて

$$\text{Condition 1:}\qquad \alpha(x)=x^{\gamma}L(x),\quad 0\leq\gamma\leq 1, \tag{3.5}$$

ただし $L:(0,\infty)\to\mathcal{R}_+$ は緩変化関数 (slowly varying function)，を仮定する：

$$\lim_{x\to\infty}L(cx)/L(x)=1,\quad \forall c>0.$$

一般性を失うことなく，$L(x)$ は連続で $L(0)<\infty$ とする．

Condition 1 は，生存関数 $\beta(x):=\sum_{n=1}^{\infty}\mathrm{I}[p_n\geq x]$, $0<x<1$，について

$$\beta(x)=x^{-\gamma}L(x),\quad 0\leq\gamma\leq 1,\quad L(cx)/L(x)=1,\quad x\to 0,$$

を仮定することと同等である．

$(X_N)_{N=1}^{\infty}$ を $p$ に従う独立な確率変数の系列とし，これより導かれる確率変数列 $(X_N^k)_{N=1}^{\infty}$, $(Z_N^r)_{N=1}^{\infty}$，$(Z_N^*)_{N=1}^{\infty}$ を

$X_N^k:=\sum_{m=1}^{N}\mathrm{I}[X_m=k]$; $k=1,2,\dots$,　変数値 $k$ の出現度数(無限カテゴリー数の多項確率変数),

$Z_N^r:=\sum_{k=1}^{N}\mathrm{I}[X_N^k=r]$; $r=1,2,\dots$,　$r$ 回出現した変数の数(寸法指標),

$Z_N^*:=\sum_{r=1}^{N}Z_n^r$;　出現した変数値の数,

により定義する．あるいは強度 1 のポアソン過程 $\{N(t),0\leq t<\infty\}$ を用い，互いに独立な，可算個の，強度 $p_n$, $n=1,2,\dots$, のポアソン過程を $(X_N)_{N=1}^{\infty}$ より

$$X_{N(t)}^k,\quad 0\leq t<\infty:\quad \text{時間間隔 } (0,t) \text{ における変数値 } k \text{ の出現度数},$$

とし，これより $Z_{N(t)}^r$, $Z_{N(t)}^*$, などを定義する．

**Condition 1 について**

**補題 3.1.** 確率母関数 $P(\xi)=\sum_{k=1}^{\infty}p_k\xi^k$ の収束半径が 1 より大ならば(Condition 2)$\alpha(x)$ は緩変化である(つまり $\gamma=0$)．

**Remark 3.1.** $\overline{\lim}_{n\to\infty} p_{n+1}/p_n = \rho < 1$ であれば $P(\xi)$ の収束半径は $1/\rho$ 以上である．つまり Condition 2 を満たす．このとき，$(p_n)_n$ の分布関数を $F(u) = \sum_{n\le u} p_n$ として，$A(x)$ を

$$1 - F(A(x)) \le 1/x \le 1 - F(A(x)-)$$

で定義すると $A(x) \sim \alpha(x)$, $x \to \infty$, である．

**例 1．** 幾何分布: $p_n = \lambda(1-\lambda)^{n-1}$, $0 < \lambda < 1$, $n = 1, 2, \ldots$ とすると，$\gamma = 0$:

$$\alpha(x) \sim \log x \,/\, (-\log(1-\lambda)), \quad x \to \infty.$$

**例 2．** ポアソン分布: $p_n = e^{-\lambda}\lambda^{n-1}/(n-1)!$, $n = 1, 2, \ldots$ とすると，$\gamma = 0$:

$$\alpha(x) \sim \log x \,/ \log(\log x)\,, \quad x \to \infty.$$

**例 3．** Condition 2 は $\alpha(x)$ が緩変化となる十分条件で，必要条件ではない．

$$p_n = c2^{-n^\beta}, \quad 0 < \beta < 1,$$

とすると $P(\xi)$ の収束半径は 1（Condition 3）だが，

$$\alpha(x) \sim (\log x/\log 2)^{1/\beta}$$

は緩変化，$\gamma = 0$.

**例 4．** ツェータ分布: $p_n \sim cn^{-\beta}$, $\beta > 1$, $n \to \infty$ ならば，$0 < \gamma = 1/\beta < 1$:

$$\alpha(x) \sim c^{1/\beta} x^{1/\beta}.$$

**例 5．** $p_n = b\,/(n+1)(\log(n+1))^{\beta+1}$, $\beta > 0$, ならば，$\gamma = 1$:

$$\alpha(x) \sim x/(b(\log x)^{1+\beta}).$$

**補題 3.2.** $p_{n+1}/p_n \to 1$，したがって $P(\xi)$ の収束半径が 1 ならば（Condition 3）

$$\alpha((1+c)x) - \alpha(x) \to \infty, \quad x \to \infty, \quad \forall c > 0.$$

Condition 1 で $\alpha(x)$ の条件を与えたが，定義そのものから次の条件を満たしている．

**補題 3.3.** $\alpha(x)$ は $\alpha(x)/x \to 0$ $(x \to \infty)$ および $\int_1^\infty (\alpha(x)/x^2)\,dx \le 1$ を満たす．

### 3.2.2 期待値の増大

$$M(t) := E(Z^*_{N(t)}) = \sum_{n=1}^{\infty}(1 - e^{-tp_k}) = \int_0^\infty (1 - e^{-t/x})\,d\alpha(x) = \int_0^\infty \frac{t}{x^2} e^{-t/x}\alpha(x)\,dx$$

$$= \int_0^\infty \frac{1}{y^2} e^{-1/y}\alpha(ty)\,dy \;\sim\; \alpha(t)\int_0^\infty \frac{y^\gamma}{y^2} e^{-1/y}\,dy = \alpha(t)\Gamma(1-\gamma), \quad 0 \le \gamma < 1, \quad t \to \infty.$$

$\alpha$ を含む積分の漸近評価は regular varying に関する Karamata 理論による．（たとえば Bingham et al.（1989）に詳しく説明されている．）

$\gamma = 1$ の場合は別に扱わなければならない．このとき補題 3.3 から $L(t) \to 0$ $(t \to \infty)$ であり，十分大きな $t$ に関して $L(t)$ は有界である．

**補題 3.4.**

$$\alpha(t) = tL(t)$$

で $L(t)$ が緩変化関数であれば$(\gamma = 1)$

$$L^*(t) = \int_0^\infty \frac{e^{-1/y}}{y} L(ty)\, dy \tag{3.6}$$

も $t \to \infty$ で緩変化である．

**命題 3.1.** $\alpha(x)$ が Condition 1 を満たすとき，$M(t) = E(Z^*_{N(t)})$ は

$$M(t) \quad \sim \quad \begin{cases} \Gamma(1-\gamma)t^\gamma L(t), & 0 \le \gamma < 1, \\ tL^*(t), & \gamma = 1; \end{cases} \qquad t \to \infty,$$

を満たす．$L^*(t)$ は 補題 3.4 で導入した関数である．

**Corollary.**

$$M(t;r) := E(Z^r_{N(t)}) = \frac{1}{r!}\int_0^\infty e^{-t/x}\frac{t^r}{x^r}\, d\alpha(x),$$

とすると，命題と同様の計算により，

$$M(t;r) \quad \sim \quad \begin{cases} \gamma\frac{\Gamma(r-\gamma)}{\Gamma(r+1)} t^\gamma L(t), & 0 < \gamma < 1, \quad r \ge 1 \quad \text{or} \quad \gamma = 1, \quad r \ge 2, \\ tL^*(t), & \gamma = 1,\ r = 1; \end{cases} \qquad t \to \infty. \tag{3.7}$$

$\gamma = 0$ のとき (3.7) の右辺は無意味で，$M(t;1)$ の行動は erratic である．たとえば $\overline{\lim}_n p_{n+1}/p_n < 1$ のとき $\gamma = 0$．下記の Remark 3.3 とその反例にしたがって，$M(t;1)$ が有界だが，振動する可能性があることを確かめられる．

**Remark 3.2.** 一般に $L^*(t)/L(t) \to 0$ であるが，さらに詳しく調べる．

$$L(t) \quad \sim \quad 1/(\log t)^\rho, \quad \rho > 1 \quad \Rightarrow \quad L^*(t) \quad \sim \quad 1\big/\big((\rho-1)(\log t)^{\rho-1}\big)\,, \quad t \to \infty,$$

$$L(t) \quad \sim \quad 1\,/((\log t)(\log\log t)^\rho)\,, \rho > 1 \quad \Rightarrow \quad L^*(t) \quad \sim \quad 1\big/\big((\rho-1)(\log\log t)^{\rho-1}\big)\,, \quad t \to \infty.$$

命題 3.1 は $t$ を $N$ に変えれば，$M_N = E(Z^*_N)$ の漸近定理となる．

### 3.2.3 分散の増大

$Z^r_{N(t)}$ が独立な 2 進確率変数の和であるから，

$$V(t) := \mathrm{Var}(Z^*_{N(t)}) = \sum_{r=1}^\infty \mathrm{Var}(Z^r_{N(t)}) = \sum_{n=1}^\infty (e^{-p_n t} - e^{-2p_n t}) = M(2t) - M(t).$$

したがって

$$V(t) \quad \sim \quad \begin{cases} \Gamma(1-\gamma)(2^\gamma - 1)L(t)t^\gamma, & 0 < \gamma < 1, \\ tL^*(t), & \gamma = 1, \end{cases} \qquad t \to \infty.$$

$M(t)$ の場合と違い，上式には $\gamma = 0$ したがって $\alpha(x) = L(x)$ が緩変化の場合が含まれておらず，別に扱わねばならない．pdf の収束範囲が 1 で，したがって $V(t) \sim L(2t) - L(t) \to \infty (t \to \infty)$ となる場合が典型的である．しかしいろいろな場合がある．

1. **Remark 3.3.** Remark 3.1 の条件が満たされる場合
$$\overline{\lim}_{n\to\infty} p_{n+1}/p_n < 1 \quad \Rightarrow \quad V(t) \quad \text{bounded.}$$
2. 上の条件の下で，$V(t)$ が収束するとは限らない．たとえば次のような例がある．
**反例．** $p_n = C2^{-2^n}, n = 1, 2, \ldots,$ とすると $t_l = 2^{2^l}, l = 1, 2, \ldots$ にたいして $V(t_l) \geq C > 0.$ しかし $\tau_l = 2^{2^l+l}$ にたいして $V(\tau_l) \to 0$ $(l \to \infty)$. この場合 $V(t)$ は有界.
3. また $V(t)$ が発振しながら $+\infty$ になることもある．
**反例．** $p_n = C2^{-2r}$, $(r-1)r/2 + 1 \leq n \leq r(r+1)/2$, $r = 1, 2, \ldots$ とすると $t_l = 2^{2^l}$ にたいして $V(t_l) \geq l(e^{-1} - e^{-2}) \to \infty$ $(l \to \infty)$, $\tau_l = 2^{2^l+l}$ にたいして $V(\tau_l) \to 0\,(l \to \infty)$.
4. 収束する場合もある．
**命題 3.2.**
$$\overline{\lim}_{n\to\infty} p_{n+1}/p_n < 1, \quad \lim_{x\to\infty} \frac{1}{x} \int_0^x [\alpha(2\xi) - \alpha(\xi)]\, d\xi = \gamma_0, \quad 0 < \gamma_0 < \infty,$$
$$\Rightarrow \quad \lim_{t\to\infty} V(t) = \gamma_0.$$
5. 収束する具体例として次の場合がある．
**例 6．**（例 1 と同じ幾何分布の場合）
$$\lim_{x\to\infty} \frac{1}{x} \int_0^x (\alpha(2u) - \alpha(u))\, du = \log 2 \,/\, (-\log(1-\rho)).$$

これらをまとめると，以下のようになる．

$$\mathrm{Var}(Z^r_{N(t)}) \sim \begin{cases} \frac{\gamma}{\Gamma(r+1)} t^\gamma L(t) \left( \Gamma(r-\gamma) - \frac{2^\gamma}{2^{2r}} \frac{\Gamma(2r-\gamma)}{\Gamma(r+1)} \right), & 0 < \gamma < 1, \quad r \geq 1 \quad \text{or} \quad \gamma = 1, \quad r \geq 2, \\ tL^*(t), & \gamma = 1,\ r = 1; \qquad t \to \infty. \end{cases}$$

$$\mathrm{Var}(Z^*_{N(t)}) \sim \begin{cases} \Gamma(1-\gamma)(2^\gamma - 1) N^\gamma L(N), & 0 < \gamma < 1, \\ NL^*(N), & \gamma = 1; \qquad t \to \infty. \end{cases}$$

$$\mathrm{E}(Z^*_{N(t)} - M(t))^{2m} \sim d_{\gamma,m} (M(t))^m, \quad m = 1, 2, \ldots$$

### 3.2.4 漸近正規性

これまでの結果に基いて $Z^*_{N(t)}$，$Z^r_{N(t)}$ の漸近正規性を示す．

**主要な結果**

**命題 3.3.** Condition 1，$0 < \gamma \leq 1$ の下で
$$(Z^*_N - E(Z^*_N))/B_N^{1/2} \quad \xrightarrow{\mathcal{D}} \quad N(0,1),$$
$$B_N = \begin{cases} \Gamma(1-\gamma)(2^\gamma - 1) N^\gamma L(N), & 0 < \gamma < 1, \\ NL^*(N), & \gamma = 1, \quad N \to \infty. \end{cases}$$

**命題 3.4.** Condition 1, $0<\gamma<1$, の下で正整数 $r_1<\cdots<r_\nu$ を固定すると

$$\left(Z_N^{r_j}-E(Z_N^{r_j})\right)/\left(N^\gamma L(N)\right)^{1/2},\quad j=1,\ldots,\nu$$

の同時分布は $N(0,\Sigma)$ に分布収束する．$\Sigma$ の要素は次の通りである．

$$\sigma_{ij}=-\frac{\gamma\Gamma(r_i+r_j-\gamma)}{r_i!r_j!}2^{\gamma-r_i-r_j},\quad i\neq j,$$
$$\sigma_i^2=\frac{\gamma}{\Gamma(r_i+1)}\left(\Gamma(r_i-\gamma)-2^{-2r_i+\gamma}\frac{\Gamma(2r_i-\gamma)}{\Gamma(r_i+1)}\right),\quad i=1,\ldots,\nu.$$

$\gamma=1$ のとき,

$$(Z_N^1-E(Z_N^1))/(NL^*(N))^{1/2}\quad\xrightarrow{\mathcal{D}}\quad N(0,1),$$
$$(Z_N^r-E(Z_N^r))/(b_rNL(N))^{1/2}\quad\xrightarrow{\mathcal{D}}\quad N(0,1),$$
$$b_r=\frac{\Gamma(r-1)}{\Gamma(r+1)}-2^{1-2r}\frac{\Gamma(2r-1)}{(\Gamma(r+1))^2},\quad r\geq 2.$$

### 3.2.5 強法則

**命題 3.5.** $(p_n)_n$ に関して何の制約もなく,

$$Z_N^*/E(Z_N^*)\quad\xrightarrow{\text{a.s.}}\quad 1,\quad N\to\infty,$$
$$Z_N^{r*}/E(Z_N^{r*})\quad\xrightarrow{\text{a.s.}}\quad 1,\quad N\to\infty.$$

しかし $Z_N^r$ については条件が必要である．

**命題 3.6.** Condition 1, $0<\gamma\leq 1$, の下で

$$Z_N^r/E(Z_N^r)\quad\xrightarrow{\text{a.s.}}\quad 1,\quad N\to\infty.$$

Condition 1 の条件の下では，Rouault (1978)が注意したように,

$$Z_N^r/Z_N^*\quad\xrightarrow{\text{a.s.}}\quad\frac{\gamma}{\Gamma(1-\gamma)}\frac{\Gamma(r-\gamma)}{\Gamma(r+1)}=\frac{\gamma(1-\gamma)^{[r-1]}}{r!},\quad r=1,2,\ldots,\quad N\to\infty.$$

これは Karlin-Rouault-Sibuya 分布である．

Rouault (1978)は Markov 連鎖モデルから Condition 1 を導いた．

## 3.3 多数出現の希少事象 LNRE

### 3.3.1 LNRE

$c$ 個のカテゴリーの確率の系列 $\boldsymbol{p}_n=(p_{1n},\ldots,p_{cn})$ にたいする多項確率の系列を $\mathsf{Mn}\,(n,\boldsymbol{p}_n)$, これからの標本を $\boldsymbol{X}_n=(X_{1n},\ldots,X_{cn}),\sum_{i=1}^c X_{in}=n$, とする．一般に $c=c(n)$ が $n$ とともに増加する 3 角配列を考える．

$$\mu_n(m)=E\left(\sum_{i=1}^c \mathrm{I}[X_{in}=m]\right),\quad m=1,\ldots,c;\ \mu_n=\sum_{m=1}^c\mu_n(m),$$

と記す．$\mu_n(m)$ は $\boldsymbol{X}_n$ の寸法指標の期待値，$\mu_n$ はその総和である．$\boldsymbol{p}_n$ が $n$ によらず固定した確率であると $X_{in}\to\infty$(a.s.) であるから

$$\lim_{n\to\infty}\mu_n(m)=0,\ m<\infty;\quad\lim_{n\to\infty}\mu_n=\infty,$$

というつまらない結果となる．語彙調査で見られ興味あるのは次の現象である．

**条件.**

(d.1) $$\liminf_{n\to\infty} E(\mu_n(1))/n > 0,$$

(d.2) $$\lim_{n\to\infty} E(\mu_n) = \infty \text{ and } \lim_{n\to\infty} \frac{E(\mu_n(1))}{E(\mu_n)} > 0,$$

(d.1)は，度数 1 のカテゴリーが全体に占める割合がいつまでも消滅しない条件,(d.2)は，新しいカテゴリーの出現が無限に続き，出現カテゴリー中で度数 1 だけの新しいカテゴリーの割合が消えない，という条件である．Khmaladze たちは(d.1),(d.2)の条件を満たす $(\boldsymbol{p}_n)_{n=1}^{\infty}$ を a sequence with large number of rare events (LNRE)と名付け，研究した．Khmaladze (1987), Khmaladze and Chitashvili (1989)参照.

(d.1)$\Rightarrow$(d.2)であるが逆は成り立たないことを後で反例により示す．$\boldsymbol{p}_n$，$c(n)$ をどのように動かせば(d.1),(d.2)が成り立つかをこの節で議論する．寸法指標そのものでなく，期待値を扱う制約は最後に議論する．

次の 2 つの関数が本質的となる．これらを $G$ 関数，$Q$ 関数と呼ぶ.

(3.8) $$G_n(z) = \sum_{i=1}^{c} \mathrm{I}[p_{in} > z], \quad Q_n(z) = \sum_{i=1}^{c} p_{in}\,\mathrm{I}[p_{in} \le z],$$

$G_n$ は $G_n(0) = c, G_n(1) = 0,$ を満たす減少関数で，$\boldsymbol{p}_n$ を降順に並べたときの順位である．前節の Karlin の関数 $\alpha$ を用いると $Q_n(z) = \alpha(1/z)$ である．$Q_n(z)$ は離散確率変数 $Z_n$ を $P\{Z_n = p_{in}\} = p_{in},\ i = 1, 2, \ldots$ により定義すると，その分布関数である．これらの関数によって条件(d.1),(d.2)を書き換えることができる．

**命題 3.7.** (i) 条件(d.1)は次の条件と同値である．(c.1) ある $z < \infty$ にたいして

$$\liminf_{n\to\infty} Q_n(z/n) > 0,$$

(ii) 条件(d.2)は次の条件と同値である．(c.2) ある $z < \infty$ にたいして

$$\lim_{n\to\infty} nQ_n(z/n) = \infty \quad \text{and} \quad \limsup_{n\to\infty} \frac{G_n(z/n)}{nQ_n(z/n)} = \infty.$$

### 3.3.2 $G$ 関数，$Q$ 関数

離散分布，離散関数は扱い難いので $\boldsymbol{p}_n$ に対応する次の関数を考える.

$$p_n(t) = \sum_{i=1}^{c} p_{in}\,\mathrm{I}[i-1 \le t < i], \quad 0 < t < c,$$

$$f_n(t) = \sum_{i=1}^{c} np_{in}\,\mathrm{I}[(i-1)/n \le t < i/n] = np_n(nt), \quad 0 < t < c/n.$$

$p_n(t)$ は $\boldsymbol{p}_n$ をヒストグラム型確率密度に対応させたものであり，$f_n(t)$ はその尺度パラメータを $1/n$ にしたものである.

連続な確率密度関数 $f$ にたいする $G$ 関数，$Q$ 関数を (3.8) に対応して

(3.9) $$G_f(z) = \int \mathrm{I}[f(t) > z]dt, \quad Q_f(z) = \int \mathrm{I}[f(t) \le z]f(t)dt,$$

とする．$f$ が減少関数ならば $G_f$ も減少関数で互に逆関数である．このとき $zG_f(z) \le 1$ である．またこのとき $Q_f$ は $f$ の生存関数を $z$ に変数変換したものである．両者の間には

$$Q_f(z) = -\int_0^z x G_f(dx)$$

の関係がある．

以上を使うと $\boldsymbol{p}_n$ の $G$ 関数，$Q$ 関数 (3.8) を

$$G_{f_n}(z) = n^{-1} G_{p_n}(z/n), \quad G_{p_n}(z) = G_n(z), \quad Q_{f_n}(z) = Q_n(z/n),$$

と表わせる．これらを用いて(c.1), (c.2)の例を構成できる．

**例 1．** $Z_1, \dots, Z_n$ を pdf $f$ をもつ iid 確率変数列

$$p_{in} = \int_{(i-1)/c}^{i/c} f(t)dt, \quad f_n(t) = cp_{in} \quad (i-1)/c \le t < 2/c,\ 1 \le i \le c,$$

とすると，$c(n) \to \infty$ のとき $f_n(t) \to f(t)$ a.e. したがって $c = c(n) \to \infty$ のとき，$G_{f_n}$ は $G_f$ に $Q_{f_n}$ は $Q_f$ に弱収束する．$\alpha$ を正定数，$c(n) = \alpha n$ とすると $Q_n(z/n) = Q_{f_n}(\alpha z)$, となり(c.1)が満たされる．

**例 2．** $p$ が減少関数とすると $p = G_p^{-1}$ である．$Z_1, \dots, Z_n$ を $p$ からの iid 確率変数列,

$$p_{in} = p_i = \int_{i-1}^{i} p(t)dt, \quad X_{in} = \sum_{j=1}^{n} \mathrm{I}[i-1 \le Z_j < i]$$

とすると，任意に固定した $p$ にたいして $(X_{1n}, \dots, X_{nn})$ は(d.1)を満足しない．

**例 3．**

$$p(t) = t^{-\gamma} L(t),\ 0 < \gamma \le 1, \quad L(tc)/L(t) \to 1,\ t \to \infty,\ \forall c > 0, \tag{c.3}$$

とする．つまり $L(t)$ は slowly varying である．$p_i, X_{i,n}$ を例 1 と同じように定義すると(d.2)を満たす．

### 3.3.3 収束定理

最後に確率標本にたいする $G$ 関数，$Q$ 関数を定義する．

$$\hat{p}_n(t) = \sum_{i=1}^{c} n^{-1} X_{in}\, \mathrm{I}[i-1 \le t < i], \quad \hat{f}_n(t) = \sum_{i=1}^{c} X_{in}\, \mathrm{I}[(i-1)/n \le t < i/n].$$

これらは単にヒストグラムである．これに対応して,

$$G_{\hat{f}_n}(z) = n^{-1} \sum_{i=1}^{c} \mathrm{I}[X_{in} > z], \quad Q_{\hat{f}_n}(t) = n^{-1} \sum_{i=1}^{c} X_{in}\, \mathrm{I}[X_{in} \le z], \tag{3.10}$$

とする．$G$ 関数は度数 $z$ 以上の出現度数の割合，$Q$ 関数は $Q_{p_n}$ に対応する，度数 $z$ 以下のカテゴリーの経験分布関数である．次の事実が重要である．

**命題 3.8.** 次の条件は(c.1), (d.1)と同等である．

$$\liminf_{n\to\infty} \| \hat{p}_n - p_n \| > 0. \tag{c.4}$$

つまり LNRE では相対度数が非一致推定量である.

しかし $Q_{f_n}$ が $Q_f$ に弱収束するならば $G_{\hat{f}_n}$ は

$$C(z) := \int_0^\infty \Lambda(z,x)x^{-1}Q_f(dx) = -\int_0^\infty \Lambda(z,x)G_f(dx), \quad \Lambda(z,x) = \sum_{k>z} e^{-z}x^k/k!,$$

に一様に確率収束する．この命題の示すことは，もしも

$$\frac{E(\hat{G}_{f_n}(z))}{E(\hat{G}_{f_n}(0+))} = \frac{1}{z} \Leftrightarrow \frac{E(\mu_n(z))}{E(\mu_n)} = \frac{1}{z(z+1)},$$

のような正則性が成り立っていても，$p_n, f_n$ についての対応する正則性

$$G_{f_n}(z)/G_{f_n}(0+) \approx 1/z,$$

などは成り立たない.

上記の収束条件を書き直すと次の命題が得られる.

**命題 3.9.** $E(\mu_n) \to \infty$ とし

$$L_n(z) := G_n(z/n) \int_0^\infty (1-e^{-z})\, dG_n(z/n),$$

が

$$\limsup_{\varepsilon\to 0} \int_0^\varepsilon z\, dL_n(z) = 0,$$

を満たすとすると

$$\frac{E(\mu_n(m))}{E(\mu_n)} \to \frac{1}{m(m+1)} \quad \Leftrightarrow \quad L_n(z) \xrightarrow{d} \int_0^\infty \frac{e^{-zx}}{1+x}\, dx.$$

**例 4.** 任意の $\varepsilon > 0$ にたいして $z_i^\varepsilon$，$1 \le i \le c_\varepsilon$ を

$$L(z_1^\varepsilon) = \varepsilon, \quad L(z_i^\varepsilon) - L(z_{i-1}^\varepsilon) = \varepsilon, \quad 1 < i \le c_\varepsilon;$$
$$c_\varepsilon = \min\left\{ c : \sum_{i=1}^{c} (1 - e^{-z_i^\varepsilon}) \ge 1/\varepsilon \right\}$$

を満たすように定める．$i$ の上限 $n_\varepsilon$ を固定し，$z_i^\varepsilon$ を正規化し，$p_{in}$ を求める.

### 3.4 種々の推測問題

本論文で議論している “母集団と標本で孤立している個体の数”以外にも，関連する多様性統計学の “種々の”課題があり，同様に困難である．この小節で短く触れておく.

まず母集団カテゴリー数の推定である．観測を続けると新しいカテゴリーが現われる．その総数が有限として，標本寸法指標から母集団カテゴリー数を推定したい．シェイクスピア全作品の corpus から “シェイクスピアは単語をいくつ知っていたか”を推定する．計算言語のプログラムでは一定期間にバグを発見した後で，“未だ何個残っているか”を推定する.

調査データでは諸属性のカテゴリーが定まっており，組合せ数も分かってはいるものの，その中に論理的，経験的にあり得ない，組合せが生ずる．既婚の少年少女や，老人の出産などで

表 3. 基本モデルの分類.

| | dependent, $\sum_j X_j = n$. | independent, $n$: mean of sum |
|---|---|---|
| | multivariate hypergeometric | independent binomial |
| absolute | $\mathsf{MvHg}(c, n, \boldsymbol{M}),\ \boldsymbol{X} \in \prod_j \mathcal{N}_{M_j}$<br>$\mathrm{E}[X_j] = n\,M_j/M =: n\,\xi_j,$<br>$\xi_j = M_j/M,\ M = \sum_{j=1}^{c} M_j.$ | $\prod_{j=1}^{c} \mathsf{Bn}(M_j, p),\ \boldsymbol{X} \in \prod_j \mathcal{N}_{M_j}$<br>$\mathrm{E}[X_j] = M_j\,p =: n\,\xi_j,\ n = pM,$<br>$\xi_j = M_j/M,\ M = \sum_{j=1}^{c} M_j.$ |
| | multinomial | independent Poisson |
| relative | $\mathsf{Mn}(c, n, \boldsymbol{\xi}),\ \boldsymbol{X} \in \Delta(c, n)$<br>$\mathrm{E}[X_j] = n\,\xi_j.$ | $\prod_{j=1}^{c} \mathsf{Po}(\lambda_j),\ \boldsymbol{X} \in \mathcal{N}_\infty^c$<br>$\mathrm{E}[X_j] = \lambda_j =: n\,\xi_j,$<br>$n = \sum_{j=1}^{c} \lambda_j, \quad \xi_j = \lambda_j/n.$ |

$c$ categories: $c \le \infty$, $\mathcal{N}_k = \{0, 1, \ldots, k\}$; $\Delta(c, n)$: lattice simplex.

ある．これらは “構造的零 structural zero” と呼ばれている．何が構造的零か曖昧な組み合わせもあるために，セル数の上限が既知だとしても，実際の数は未知で推定することになる．

しかしカテゴリー数が有限であることと非常に小さな確率が存在することは区別できない．有限が明確なときに推定すべきである．

上限が不確定であれば，むしろ未観測カテゴリーの母集団での割合に意味がある．生態学の種の多様性研究における一つの課題である．その変形として，観測数をさらに増やしたときに新しいカテゴリーがどれだけ増えるか，予測する課題がある．

以上の課題の推測法，その他の課題については Bunge and Fitzpatrick (1993)参照.

## 4. 事前分布の導入

### 4.1 モデルの分類と事前分布の役割

#### 4.1.1 モデルの分類

問題が本質的に困難であり，素朴な推定法がよく働かないときに利用されているのは経験ベイズ法である．ここの文脈では超母集団 super population を仮定し，母集団をそこからの標本とみなす．事前分布を導入する前に，第 1 章で議論した基本モデルを表 3 のように整理しておく．

行方向 2 分類の絶対モデルでは，母集団からサンプリングにより標本を得る．相対モデルは超母集団を想定しており，$n$ が母集団のサイズか，標本のサイズかの違いとなる．より正しくは，標本のサイズと，標本に含まれない母集団のサイズの違いで，標本から他の標本を予測することになる．“absolute abundance”，“relative abundance” は Engen (1978)の用語である．

列方向はサンプリング法の違いでもある．母集団リストから予め定めた大きさの標本をとるか，ランダムに選んだ小集団全部を観測するかによる．

4 種のモデルでは分布範囲が異なるが，$c, n \to \infty$ のとき漸近的に同等であり，独立ポアソン分布で代表することができる．

基本モデルのそれぞれにたいする，経験ベイズアプローチが提案されている．ただしこれ

らを確率過程としてみることにより，さらに視野が広がる．離散的な壺のモデル，連続的な Lévy 過程，単位区間を分割する residual allocation model（RAM）などがあるが，ここでは議論しない．

### 4.1.2 事前分布

4 種のモデルのそれぞれにたいする事前分布を考える．

1. absolute and dependent

$\boldsymbol{M} = n\boldsymbol{\xi}$ の従属多変量同時分布でなければ固有の事前分布とはならない．ひとつのアプローチがカテゴリーを一項目とせず，他項目の組合わせとする方法である．それが $S = (s_1, s_2, \dots)$ の分布に及ぼす影響は明らかでない．

2. relative and dependent

$\boldsymbol{\xi}\,(c = \infty)$ が $\Delta$（無限次元単位単体）上の GEM 分布，2 パラメータ GEM 分布に従うならば Ewens 確率分割，Pitman 確率分割となる．Ewens-Pitman 確率分割では「母集団」全体からの非復元抽出による標本の確率分割は $n$ の違いだけとなる．

$\boldsymbol{\xi}\,(c < \infty)$ が Dirichlet 分布 $\mathsf{Dir}\,(c, \boldsymbol{\alpha})$ に従うならば $\boldsymbol{X}$ は多変量負の超幾何分布 $\mathsf{MvNgHg}\,(c, n, \boldsymbol{\alpha})$ に従う．対称 Dirichlet 分布 $\mathsf{Dir}\,(\gamma\mathbf{1})$ に従うならば Pitman 確率分割 $(\theta = c\gamma,\ \alpha = -\gamma < 0)$ となる．$\theta$ を固定し $\gamma \to 0, c \to \infty$ とすれば Ewens 確率分割となる．

**注意．** 多変量負の超幾何分布 $\mathsf{MvNgHg}\,(c, n, \boldsymbol{\alpha}) \sim X$ から，$\nu$ 個の個体を非復元抽出した標本は $\mathsf{MvNgHg}\,(c, \nu, \boldsymbol{\alpha}) \sim Y$ である．$X - Y | Y = \boldsymbol{y} \sim \mathsf{MvNgHg}\,(c, n - \nu, \boldsymbol{\alpha} + \boldsymbol{y})$ となり，両部分は独立ではない．

3. absolute and independent

現存の人間をランダムに分割するのではなく，過去に運命づけられた人間がランダムに生まれたと見ることに相当する．

**命題 4.1.** $(M_j)_{j=1}^{c}$ が iid その pmf（probability mass function = pf）を $p_M(x) = p_M(x;\theta)$, $x = 0, 1, \dots,$ とする．$(M_j)_{j=1}^{c}$ の 寸法指標を $(T_1, T_2, \dots)$ とする．$\mathrm{E}(M_j) = m(\theta) = \sum_{x=0}^{\infty} x p_M(x;\theta) \to 0$ $(\theta \to 0)$ となるようにパラメータ $\theta$ を持つ確率分布族を選び，$c$ を $c p_M(x;\theta(c)) \to \lambda(x)$ $(c \to \infty, \theta \to 0)$ を満たすように選べば

$$(T_1, \dots, T_l) \xrightarrow{d} \prod_{i=1}^{l} \mathsf{Po}(\lambda(i)), \quad c \to \infty.$$

**証明．** $T_k$ の周辺分布は

$$T_k \sim \mathsf{Bn}(c, p_M(k)), \quad k = 0, 1, \dots,$$

であり，同時分布は “サイズをカテゴリーとする無限多項分布”で，

$$\Pr\{(T_0, T_1, \dots) = (t_0, t_1, \dots)\} = c! \prod_{i=0}^{\infty} \frac{p_M(i)^{t_i}}{t_i!}, \quad \sum_{i=0}^{\infty} t_i = c. \tag{4.1}$$

このとき $T = \sum_{j=1}^{c} M_j = \sum_{i=1}^{\infty} i\, T_i$ は確率変数で，その pgf は $G^c(z)$ であり $E(T) = c\, E(M_j)$ である．

その周辺分布は非退化多項分布である: $(T_1, \dots, T_l) \sim \mathsf{Mn}\,(l, c; (p_M(1), \dots, p_M(l)))$. その pgf は $(1 + \sum_{i=1}^{l} p_M(i)(z_i - 1))^c$ である．したがって $(T_1/c, \dots, T_l/c) \xrightarrow{\text{a.s.}} (p_M(1), \dots, p_M(l))$, $c \to \infty$. また各 $T_i$ をポアソン近似できる．□

命題 4.1 は母集団分布が超母集団からの標本であるとみなしている．母集団 $(M_j)_{j=1}^c$ からの標本 $(X_j)_{j=1}^c$ の大きさは $X_j \sim \mathsf{Bn}(M_j, p)$ ($p$ は抽出率)つまり

$$P\{X_j = x\} = \sum_{i=x}^{\infty} p_M(i) \binom{i}{x} p^x (1-p)^{i-x}.$$

したがって $(X_i)_{i=1}^c$ の寸法指標 $(S_1, S_2, \dots)$ の分布は，一般に $(T_1, T_2, \dots)$ の分布と同じではない．

4. relative and independent

ポアソン分布パラメータ $\lambda_j$ の事前分布として以下のものを想定する．$\boldsymbol{\xi} \in \Delta$ にたいして $\mathrm{P}\{Z = \xi_j\} = \xi_j,\ j = 1, 2, \dots,$ という "characteristic random variable" $Z = Z_{\boldsymbol{\xi}}$ を想定し，その確率分布を事前分布とする．母集団比率 $(Z_j)_{j=1}^c$ の，$\xi_j$ の相対度数を，$\hat{\xi_j}$ とする．標本比率の分布は $P\{X_j = \xi_j\} = \hat{\xi_j}$ であり，$\hat{\xi_j} \xrightarrow{\text{a.s.}} \xi_j\ (c \to \infty)$ より，$\boldsymbol{X} \xrightarrow{d} \boldsymbol{Z}(c \to \infty)$. 応用よりは理論的な興味であるが，母集団と標本が同じ分布であることに意味がある．

$X_j \sim \mathsf{Po}\ (\lambda\eta_j), j = 1, \dots, c, \sum_j \eta_j = 1$ が独立，$\lambda$ もこれらと独立で，$\lambda \sim \mathsf{Ga}\ (\gamma, a)$ ならば，$\boldsymbol{X} \sim \mathsf{NgMn}\ (\gamma, \boldsymbol{\xi}), \boldsymbol{\xi} = ((1+a)^{-1}, a(1+a)^{-1}\boldsymbol{\eta})$. 一般に独立な確率変数が mixing により従属となる．

$n\,\xi_j$ が Gamma 分布など無限分解可能連続確率分布に従う必要十分条件は，abs. ind. $n\xi_j$ が無限分解可能離散確率分布に従うことである．

**注意と議論．** ポアソン・モデル(rel. ind.)では母集団と標本の区別が $n$ の違いであるが，被混合分布 mixture では一般に両分布は違う．負の 2 項分布であっても $X \sim \mathsf{NgBn}(\xi, k)$, $Y \sim \mathsf{Bn}(X, \rho)$ ならば

$$Y \sim \mathsf{NgBn}(\xi/(1-(1-\xi)(1-\rho)), k), \quad E(Y) = \rho E(X),$$
$$X - Y | Y = y \sim \mathsf{NgBn}(1-(1-\xi)(1-\rho), k+y)$$

である．標本と残りは独立でないし，パラメータの変化も注意を要する．$X \sim \mathsf{Po}(v)$ のとき，$v$ が確率変数で $v \sim \mathsf{Ga}(k, \alpha)$ であることを表わす Gurland の記号を用いると，

$$\mathsf{Po}\ (v) \bigwedge_v \mathsf{Ga}\ (k, a) \sim \mathsf{NgBn}\ (1/(1+a), k),$$

だから，$\xi = 1/(1+a)$ を $1/(1+\rho a) = \xi/(\xi + (1-\xi)\rho)$ に変えることになる．

負の多項分布からの 2 項サンプリングでも平行した議論となる．

もっとも考えやすいのは，$c$ 個の独立なポアソン過程の混合において，あい交わらない時間間隔での観察で，標本と，標本から残された母集団を想定することである．このモデルは，rel.dep. において $n$ が確率変数で，標本が $\mathsf{Po}\ (\rho M)$ 母集団の残りが $\mathsf{Po}\ ((1-\rho)M)$ であることに相当する．

カテゴリーの確率に事前分布を導入することにより，カテゴリーが消滅して，寸法指標の議論は，離散確率変数の標本度数の順序統計量の議論に帰着することを命題 4.1 で見た．ポアソン・モデルに事前分布を導入すると，より多様なモデルとなる．緩やかな条件の下で，寸法指標について次の命題が成り立つ．

**命題 4.2.** 非負整数値をとる pmf $f(x), x = 0, 1, \dots,$ からの確率標本を $X = (X_1, \dots, X_c)$ とし，その寸法指標を $S = (S_0, S_1, \dots), S_\nu = \sum_{k=1}^{c} \mathrm{I}[X_k = \nu]$ とする．$X$ の同時 pmf

$$X \sim \prod_{k=1}^{c} f(x_k),$$

を変形すれば，$S$ の同時 pmf は，

$$S \sim c! \prod_{\nu=1}^{\infty} (f(\nu))^{s_\nu}/s_\nu!, \quad \sum_\nu s_\nu = c, \sum_\nu \nu s_\nu = \sum_k x_k.$$

$S$ の同時階乗キュミュラントは，

$$E\left(\prod_\nu S_\nu^{\underline{r_\nu}}\right) = \begin{cases} c^{\underline{r}} \prod_\nu (f(\nu))^{r_\nu}, \ r = \sum_\nu \le c, \\ 0, \quad r > c, \end{cases} \quad \to \prod_\nu (cf(\nu))^{r_\nu}, \ c \to \infty.$$

つまり寸法指標 $S_\nu, \nu = 0, 1, \dots$ を平均 $cf(\nu)$ の独立なポアソン分布で近似できる．

### 4.2 無限分解可能離散分布の役割

R. A. Fisher はマレーシアの蝶の種類を議論し対数級数分布を導入した．それはポアソン分布の強度パラメータがガンマ分布に従うことを仮定し，負の 2 項分布を導く．さらに零を打切り，ガンマ分布の尺度母数を零に近づけることにより対数級数分布を導いた．Fisher et al. (1943)を参照．混合する分布を変える試みの結果，それが無限分解可能確率母関数をもつと扱いやすいことが分かった．

それは，上記の rel.ind. の場合に述べたように混合する分布が無限分解であれば混合されたポアソン分布(Poisson mixtures)およびその極限も無限分解可能となる．さらに次の命題で表わされる分布，確率母関数，に限ることにより議論が容易となる．

**命題 4.3.** (Steutel and van Harn (1979))非負整数値をとる rv $X$ の pgf を $G(z) = E(z^X)$ とし，$P\{X = 0\} = G(0) > 0$ を仮定する．$G(z)$ が無限分解可能な pgf である必要十分条件は

$$G(z) = G(z; \theta) = \exp(\theta(g(z) - 1)), \ \theta > 0$$

と表わせることである．$g(z)$ も pgf であり，$g(0) = 0$ と制約すれば一意に定まる．(下記の注意参照)

つまり pgf が $g(z)$ である rv をクラスターの大きさ(cluster size)とするポアソン中断和(stopped sum)として表わされる．

**注意．** 一般に $\mathrm{pgf}\, G(z) = \exp(\theta(g(z) - 1)), \ Z \sim g(z) = \sum_{k=0}^{\infty} p_k z^k$ において $P\{Z = 0\} = p_0 = g(0) = 0$ と仮定して一般性を失わない．実際

$$G(z) = \exp\left(\theta\left(\sum_{k=1}^{\infty} p_k z^k - (1 - p_0)\right)\right) = \exp((1 - p_0)\theta(g^*(z) - 1)),$$

$$g^*(z) = (1 - p_0)^{-1} \sum_{k=1}^{\infty} p_k z^k,$$

で $g^*(z)$ は $Z|Z>0$ の pgf である．逆に $G(z)=\exp(\theta(g(z)-1)), g(z)=\sum_{k=1}^{\infty}p_kz^k$ のときに，任意の $\rho(0<\rho<1)$ にたいして，

$$G(z)=\exp(\rho^{-1}\theta(g^*(z)-1)),\quad g^*(z)=\sum_{k=0}^{\infty}p_k^*z^k,\ p_0^*=1-\rho,\ p_k^*=\rho p_k, k>0,$$

と表現できる．つまり $\theta$ の変化は $p_0$ に影響し $(p_k, k>0)$ の相対的な大きさに影響しない．

**命題 4.4.** $g(0)=p_0=0$ のとき，

$$\frac{G(z;\theta)-G(0;\theta)}{1-G(0;\theta)}\ \to\ g(z)\quad(\theta\ \to\ 0). \tag{4.2}$$

**証明．** 一般に pgf $G(z)$ の 0 打切り分布 の pgf は $(G(z)-G(0))/(1-G(0))$ である．ポアソン分布 $\mathsf{Po}\,(\theta)$ の 0 打切り分布 $\mathsf{ZtPo}\,(\theta)$ は $\theta\to 0$ のとき値が 1 の分布に分布収束する．ポアソン中断和では，クラスター 1 個の場合，つまりクラスター分布そのものとなる．

**命題 4.5.**

$$X\sim G(z)=\exp(\theta(g(z)-1)),\quad g(z)=\sum_{k=1}^{\infty}p_kz^k,$$

とすると

$$P\{X=x\}=e^{-\theta}\sum\theta^t\prod_{i=1}^{x}\frac{p_i^{s_i}}{s_i!},\quad t=\sum_{i=1}^{x}s_i, \tag{4.3}$$

ただし $\sum$ は，$\sum_{i=1}^{x}is_i=x$ を満たす，$x$ のすべての分割 $(s_1,\ldots,s_x)$ に関する和である．

**証明．** $P\{X=x\}$ は $G(z)$ を展開したときの $z^x$ の係数である．

**補題．**

$$Y_n\sim g^n(z),\ g(z)=\sum_{k=0}^{\infty}p_kz^k,$$

とすると

$$P\{Y_n=y\}=n!\sum\prod_{i=0}^{y}p_i^{s_i}/s_i!,$$

ただし和は $\sum_{i=1}^{y}is_i=y,\ \sum_{i=0}^{y}s_i=n$，を満たすすべての分割 $(s_0,\ s_1,\ldots,s_y)$ に渡る．$g(0)=p_0=0$ のときは $s_0=0,\ \sum_{i=1}^{y}s_i=n$ に限られる．

**証明．** $P\{Y_n=y\}$ は $g^n(z)$ を展開したときの $z^y$ の係数であるが，$y$ より高次の項は関係しないから，多項展開の項を整理して，

$$\left(\sum_{k=0}^{y}p_kz^k\right)^n=n!\sum_{\sum n_k=n}\prod_{k=0}^{y}\frac{1}{n_k!}(p_kz^k)^{n_k}=n!\sum_{w=0}^{ny}z^w\sum_{\substack{\sum n_k=n\\ \sum kn_k=w}}\prod_{k=0}^{y}\frac{p_k^{n_k}}{n_k!}.$$

$p_0=0$ ならば $p_0$ のべき乗の項が消える．

**命題 4.5 の証明.** 補題において $n$ が平均 $\theta$ のポアソン分布に従う確率変数 $N$ ならば,

$$P\{X=x\}=P\{Y_N=x\}=e^{-\theta}\sum_{n=0}^{\infty}\theta^n\sum_{\substack{\sum s_i=n\\ \sum is_i=x}}\prod_{i=1}^{x}\frac{p_i^{s_i}}{s_i!}.$$

二つの和を合わせれば (4.3) が得られる.

**注意.** 命題の意味について後で議論する. 命題で, $p_0>0$ の項を含めても, 結果は変わらない. また, 上記の証明は,

$$P\{X=k\}=(1/k!)\,(d/dz)^k\,G(z)|_{z=0}$$

であることに注意して, 合成関数の高階微分を求める次のファ・ディ・ブルノの公式を用いることと同じである.

$$\frac{d^\nu}{dz^\nu}f(g(z))=\nu!\sum\frac{d^k}{dy^k}f(y)\big|_{y=g(z)}\prod_{i=1}^{\nu}\frac{1}{s_i!}\left(\frac{1}{i!}\frac{d^ig(z)}{dz^i}\right)^{s_i},\quad k=\sum_{i=1}^{\nu}s_i,$$

ただし $\sum$ は, $\sum_{i=1}^{\nu}is_i=\nu$ を満たす, すべての $(s_1,\ldots,s_\nu)$ に関する和である.

この節の議論をまとめる. 命題 4.2 で, 事前分布を導入したとき寸法指標のポアソン近似について述べた. ところが混合されたポアソン分布は, 少数の例を除くと複雑な形となり, 応用を妨げている.

ところで命題 4.5 において, $X=(X_1,\ldots,X_c)$ が $\mathrm{pgf}\,G(Z;\theta)=\exp(\theta(g(z)-1))$ の確率標本であれば $Z_c:=\sum_{k=1}^{c}X_k\sim G(z;c\theta)$ であり, その pmf が (4.3) の形となる. これは $Z_c=z$ の条件のもとで $z$ の確率的分割を示している. $X$ の寸法指標 $S=(S_1,\ldots,S_z),\sum_{\nu=1}^{z}\nu S_\nu=z,\sum_{\nu=1}^{z}S_\nu=c$ とし, 命題 4.2 と比較すると, $\sum_{\nu=1}^{z}\nu S_\nu=z$ の条件の下で, $S$ は $g(z)$ からの確率標本の寸法指数となっている. したがって次の命題が成り立つ. 精確な近似評価を求めることが必要である.

**命題 4.6.**

$$(X_i)_{i=1}^{\infty}\quad\text{iid}\quad X_i\sim G(\lambda(g(z;\theta)-1)),g(z;\theta)=\sum_{k=0}^{\infty}p_k(z\theta)^k,$$

$$m(\theta)=\sum_{k=1}^{\infty}kp_k\theta^k,\quad\theta<\theta_0,\quad\theta_0=\sup\{\theta:m(\theta)<\infty\},$$

とする. $(X_1,\ldots,X_c)$ の寸法指標を

$$S_{c,\nu}=\sum_{i=1}^{c}\mathrm{I}[X_i=\nu]$$

と記す. 任意の正整数 $l$ にたいして

$$(S_{c,1},\ldots,S_{c,l})\xrightarrow{d}\prod_{j=1}^{l}\mathsf{Po}\,(mp_j),\quad m=cm(\theta),\ c\to\infty,\ \theta\to0.$$

### 4.3 新しい研究方向

これまでどのような研究が行われてきたかを概観した．いずれの方法も完全ではないが，有望な方向を示している．本特集号の諸論文が現在の新しい研究，特に具体的な推測の方法を扱っている．

付録の一般 Zipf 分布の多くは壺に玉を入れる過程として導ける．これらと別に集団遺伝学で発展した確率分割の議論がある．Ewens と Pitman による確率分割の族は基礎概念で，確率過程論といろいろ結び付いている．Zipf 法則の新しい代表である．

モデルに関する仮定をなるべく少なくし，セミパラメトリックなモデルにより推測できれば非常に都合がよい．応用統計学としては標本の大きさが非常に大きい例であるが，推測精度の議論が単純ではない．

この節で議論したベイズ法は諸困難を回避している．観測データ類別による標本寸法指標について事前分布を想定する意味が明確でない．しかも任意のクラスター分布を考えられるので，モデルの自由性は基本モデルの自由性とあまり変わらない．LNRE 理論は期待値についての法則である．この結果から見ると，経験ベイズの尤度の意味も問い直すことになる．

一方 Karlin の定理は，条件が限定されているものの明確な強い定理である．当然 1 パラメータ分布族では実際データに当てはまらないから，その変形として一般 Zipf 分布族を利用することも考えられる．

当面はこれらの方途 (approach) の間の相互関係の探求が必要であろう．

#### 4.3.1 無限分解可能分布に基づくモデル

以下ポアソン過程のベイズ法を議論する．Gurland の記号法で形式化すると，

$$\text{mixture}: \quad \text{Poisson}(\lambda) \bigwedge_{\lambda} F(\xi) \quad F: \text{infinitely divisible}$$

$$\text{stopped sum}: \quad \text{Poisson}(\lambda) \bigvee F(\eta) \quad F: \text{positive integer r.v.}$$

このスキームで混合する分布，混合結果の分布が陽に表わせるものを探すと，表 4 のようになる．混合する分布としては他に多くの提案がされているが，一般にその結果が複雑になり利用しにくい．離散安定分布，離散 Linnik 分布，KRS については付録 KRS 分布を参照．

表 4 の中の 2 つの分布を調べる．

1. Neyman の A 型分布

次の 3 つのモデルが同等である．ただし $Z_t$ は零打切りを表わす．

$$G(z) = \exp(\lambda(\exp(\phi(z-1)) - 1)).$$

$$\mathsf{Po}\ (\phi j) \bigwedge_{j} \mathsf{Po}\ (\lambda), \quad \mathsf{Po}\ (\lambda) \bigvee \mathsf{Po}\ (\phi), \quad \mathsf{Po}\ (\lambda(1-e^{-\phi})) \bigvee \mathsf{ZtPo}\ (\phi).$$

表 4. 陽な混合ポアソン分布.

| Mixture-Generalized | Mixing | Summed |
|---|---|---|
| Negative Binomial | Gamma | Logarithmic series |
| Neymann Type A | Poisson | Poisson |
| Hermite | | shifted-binomial |
| PIG | Inverse Gauss | 0-Trunc. neg. binom. |
| Discrete stable | Gamma*Stable | KRS |
| Discrete Linnik | Stable | KRS |

$$p(x)=\frac{e^{-\lambda}\phi^x}{x!}\sum_{j=0}^{\infty}\frac{(\lambda e^{-\phi})^j j^x}{j!}=\frac{\exp(-\lambda(1-e^{-\phi}))\phi^x}{x!}\sum_{k=1}^{x}\left\{{x\atop k}\right\}(\lambda e^{-\phi})^k,$$

$$p(0)=\exp(-\lambda(1-e^{-\phi})).$$

ただし $\left\{{x\atop k}\right\}$ は第 2 種スターリング数である．

2. 離散安定分布（5.2.3 節参照）

$$G(z)=\exp(-\lambda(1-z)^{\gamma}).\quad \mathsf{Po}\,(\lambda)\bigvee \mathsf{KRS}\,(\gamma),$$

$$p(x)=(-1)^x\sum_{j=0}^{\infty}\binom{\gamma j}{x}\frac{(-\lambda)^j}{j!}=(-1)^x e^{-\lambda}\sum_{m=0}^{x}\sum_{j=0}^{m}\binom{m}{j}\binom{\gamma j}{x}\frac{\lambda^m}{m!}$$

$$=e^{-\lambda}\sum\lambda^k\prod_{i=1}^{x}\frac{1}{s_i!}\left(\frac{\gamma(1-\gamma)^{\bar{i}}}{i!}\right)^{s_i},\quad k=\sum_{i=1}^{x}s_i,$$

$\sum$ は $\sum_{i=1}^{x} is_i=x$ を満たす分割についての和である．最後の式は (4.3) 式による．

3. Zipf 法則へのアプローチ

Zipf 法則を導く種々の方法を第 3 章で紹介したが，いずれも複雑で統計モデルの構築には適さない．もうひとつのアプローチとして，ポアソン分布の混合により，寸法指数が Zipf 法則に近いものを探したい．しかし特殊な例を除くと，混合された分布の陽な表現を得られず，混合分布を選び出すことが難しい．

発想を変えて混合分布でなく，クラスター分布を想定すれば扱いはやさしくなることを，命題 4.2，4.4 が示している．たとえばクラスター分布として一般 Zipf 分布を選ぶことが考えられる．このときのポアソン中断和がどうなるか，最も簡単な Zipf 分布で調べる．一般 Zipf 分布を選びパラメータ推測を行うことは別稿で議論する．

なお査読者から Zipf 分布をシフトした $p_0(1+x)$，$x=0,1,2,\ldots$ が無限分解可能であることを示す，あるいは否定する問題を提示されたが，未解決である．

Zipf 分布

$$p_0(x)=1/x(x+1),\quad x=1,2,\ldots$$

の確率母関数は

$$G_0(z)=1-(1-z^{-1})\log(1-z),$$

これをクラスター分布とする無限分解可能分布の確率母関数は

$$G(z;\lambda)=\exp(\lambda(G_0(z)-1))=\exp\left(\lambda\left(\sum_{k=1}^{\infty}\frac{z^k}{k(k+1)}-1\right)\right)=(1-z)^{-\lambda(1-1/z)}$$

である．これを展開すると，

$$G(z;\lambda)=e^{-\lambda}\left(1+\frac{\lambda}{2}z+\left(\frac{\lambda}{6}+\frac{\lambda^2}{8}\right)z^2+\left(\frac{\lambda}{12}+\frac{\lambda^2}{12}+\frac{\lambda^3}{48}\right)z^3\right.$$
$$\left.+\left(\frac{\lambda}{20}+\frac{\lambda^2}{18}+\frac{\lambda^3}{48}\right)z^4+\left(\frac{\lambda}{30}+\frac{7\lambda^2}{180}+\frac{5\lambda^3}{288}\right)z^5+\cdots\right),$$

$$G(z;1)=e^{-1}\left(1+\frac{1}{2}z+\frac{7}{24}z^2+\frac{9}{48}z^3+\frac{91}{720}z^4+\frac{129}{1440}z^5+\cdots\right).$$

$G(z,\lambda)$, $\lambda \to 0$, を考えると，$z^k$ の係数である $\lambda$ の多項式で 1 次項が支配的であり，その係数は $1/k(k+1)$ となることに注意せよ．

### 4.3.2 多数希少現象との関係

数の可算順合 $\boldsymbol{\lambda} = \{\lambda_i\}_{i=1}^{\infty}$, $\lambda_i \geq 0$, $\lambda = \sum_{i=1}^{\infty} \lambda_i \leq \infty$ にたいして系列

$$\boldsymbol{\mu} = (\mu_j)_{j=0}^{\infty}, \quad \mu_j = \mu_j(\boldsymbol{\lambda}) := \sum_{i=1}^{\infty} e^{-\lambda_i} \lambda_i^j \,/\, j!, \tag{4.4}$$

を定める．変換 $\boldsymbol{\lambda} \mapsto \boldsymbol{\mu}$ が多数希少現象 LNRE の主要課題である．特に $\boldsymbol{\lambda} = n\boldsymbol{p}$, $p_j \geq 0$, $\sum_{j=1}^{\infty} p_j = 1$ の場合に，パラメータ $n$ にたいする $\boldsymbol{\mu}(n\boldsymbol{p})$, $n \to \infty$, の挙動を議論する．

$(\lambda_i)_{i=1}^{L}$ を等確率でとる確率分布の分布関数(経験分布関数)を $Q_L(\lambda)$ とすると，

$$\mu_j \,/\, L = \int_0^{\infty} \left( e^{-\lambda} \lambda^j \,/\, j! \right) dQ_L(\lambda), \quad j = 0, 1, \ldots$$

であり $L^{-1}\boldsymbol{\mu}$ は，成分の順序を適当に変えれば，ポアソン混合分布 $\mathsf{Po}(\lambda) \bigwedge_{\nu} Q_L$ の確率関数である．

$Q_L$ を連続確率分布で近似し(事前分布)，さらに無限分解可能分布で近似すればポアソン中断和モデルで表わされる．無限分解可能の節の初めで注意したように，ポアソン強度パラメータの変化はクラスター要素数 0 の確率に影響するだけで，$(\mu_1, \mu_2, \ldots)$ の相対的な大きさには影響しない．したがって寸法指標の期待値を問題にする限り，LNRE の一般理論はポアソン中断和の導入により一段落し，そこから新しい課題が始まる．

### 4.3.3 Pitman 確率分割と関連する分布

1. 一様多変量負の超幾何分布と Pitman 確率分割

多変量負の超幾何分布 $\mathsf{MvNgHg}\ (m, n; \boldsymbol{\nu})$

$$p(\boldsymbol{x}) = \frac{n!}{\nu^{\overline{n}}} \prod_{j=1}^{m} \frac{\nu_j^{\overline{x}_j}}{x_j!}, \quad \nu = \sum_{j=1}^{m} \nu_j \tag{4.5}$$

において $\nu_1 = \cdots = \nu_m = \gamma$ とすると，

$$p(\boldsymbol{x}) = \frac{n!}{(m\gamma)^{\overline{n}}} \prod_{j=1}^{m} \frac{\gamma(1+\gamma)\cdots(x_j - 1 + \gamma)}{x_j!},$$

$(x_1, \ldots, x_n)$ の寸法指標を $(s_0, \ldots, s_n)$ とすると，

$$\begin{aligned} p(s_1, \ldots, s_n) &= \frac{n!}{(m\gamma)^{\overline{n}}} \sum \frac{m!}{\prod_{i=0}^{n} s_i!} \prod_{i=1}^{n} \left( \frac{\gamma(1+\gamma)\cdots(i-1+\gamma)}{i!} \right)^{s_i} \\ &= \frac{n! \gamma^k m^{\underline{k}}}{(m\gamma)^{\overline{n}}} \prod_{i=1}^{n} \left( \frac{(1+\gamma)^{\overline{i-1}}}{i!} \right)^{s_i} \frac{1}{s_i!}, \quad k = \sum_{i=1}^{n} s_i = m - s_0. \end{aligned} \tag{4.6}$$

一方，Pitman の確率分割

$$p(s_1, \ldots, s_m) = \frac{n! \theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)}{\theta^{\overline{n}}} \prod_{i=1}^{n} \left( \frac{(1-\alpha)^{\overline{i-1}}}{i!} \right)^{s_i} \frac{1}{s_i!}, \quad k = \sum_{i=1}^{n} s_i,$$

で $-\alpha = \gamma > 0$, $\theta = m\gamma$, $m = 1, 2, \ldots$ とすると

$$p(s_1, \ldots, s_n) = \frac{n! \gamma^k m^{\underline{k}}}{(m\gamma)^{\overline{n}}} \prod_{i=1}^{n} \left( \frac{(1+\gamma)^{\overline{i-1}}}{i!} \right)^{s_i} \frac{1}{s_i!}$$

となり，両者は一致する．

多変量負の超幾何分布は，多項分布の Dirichlet 分布による混合として，あるいは負の 2 項確率変数の和を与えたときの条件付分布として得られる．(4.6)で $\gamma = -\alpha < 0$ とすると Pitman 確率分割が得られそうに見えるが，それほど簡単ではない．

2. 和の条件付分布

Pitman 確率分割で $n$ を任意に固定したとき $\sum_{i=1}^{n} S_i = k$ の条件の下での $(S_1, \ldots, S_n)$ の同時確率関数は

$$\frac{n!}{c(n,k,\alpha)} \prod_{i=1}^{n} \left( \frac{\alpha(1-\alpha)^{\overline{i-1}}}{i!} \right)^{s_i} \frac{1}{s_i!}, \quad \sum_{i=1}^{n} i s_i = n, \quad \sum_{i=1}^{n} s_i = k,$$

である．ただし $c(n,k,\alpha)$ は $t$ の恒等多項式

$$(t\alpha)^{\overline{n}} = \sum_{k=1}^{n} c(n,k,\alpha) t^{\overline{k}}$$

で定義される $\alpha$ の多項式である．これは $k$ を任意に固定し，

$$(Y_1, \ldots, Y_k) \quad \text{iid} \quad Y_1 \sim \text{KRS}(\alpha)$$

の $\sum_{j=1}^{k} Y_j = n$ の条件の下での寸法指標 $(S_1, \ldots, S_n)$ の同時確率関数に等しい．条件の与え方が違うことに注意．

より一般的には Engen's generalized negative binomial に従う確率変数の，和(十分統計量)が与えられた条件の下での分布でもある．

ちなみに KRS 変数の和 $\sum_{j=1}^{k} Y_j$ の pgf は次のようになる．

$$G^k(z) = (1-(1-z)^{\alpha})^k = \sum_{\nu=0}^{\infty} (-z)^{\nu} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \binom{\alpha j}{\nu} \tag{4.7}$$

$$= \sum_{\nu=k}^{\infty} \frac{\alpha}{\nu!} \sum_{i=1}^{k} \binom{k}{i} (-1)^{i-1} z^i (1-i\alpha)^{\overline{\nu-1}}.$$

## 5. 付録

### 5.1 一般 Zipf 分布

Zipf 法則の定式化は多様であるが，典型的な発想は，中心極限定理のような一般的法則があり，その極限分布として得られる，というものである．一般的法則はともかく，極限分布の候補となるものを模索する仕事が行われている．その中にグルジア共和国の首都トビリシの人々の成果がある．本文の LNRE の理論もこれらと結び付いている．

Zipf 法則に関連して提案されてきた諸確率分布が，次の確率関数により統一的に表わされる．Orlov and Chitashvili (1983a, 1983b), Baayen (2001)を参照．

$$p(x;\alpha,\beta,\gamma) = \int_0^{\infty} \frac{(\ln(1+t))^{\gamma-1} t^{\alpha}}{(1+t)^{x+\beta+1}} dt \Big/ \int_0^{\infty} \frac{(\ln(1+t))^{\gamma-1} t^{\alpha-1}}{(1+t)^{\beta+1}} dt, \tag{5.1}$$

$$x = 1, 2, \ldots; \quad \alpha, \beta, \gamma > 0.$$

これが確率分布であることは $\sum_{x=1}^{\infty} 1/(1+t)^x = 1/t$ から確かめられる．明らかに $x$ の減少関数である．次の特別の場合には陽に表わすことができる．

$$p(x) = p(x;\alpha,\beta,1) = \frac{\alpha\Gamma(\beta-\alpha+x)\,\Gamma(\beta+1)}{\Gamma(\beta-\alpha+1)\,\Gamma(\beta+1+x)} = \frac{\alpha(\beta-\alpha+1)^{\overline{x-1}}}{(\beta+1)^{\overline{x}}}$$

$$=\frac{(\beta-\alpha+1)^{\overline{x-1}}}{(\beta+1)^{\overline{x-1}}}-\frac{(\beta-\alpha+1)^{\overline{x}}}{(\beta+1)^{\overline{x}}},\quad x=1,2,\dots,\quad 1+\beta>\alpha>0.$$

これを Waring-Herdan-Muller 分布と呼ぶ．この分布の確率母関数は

$$G(z)=(z-1)\,{}_2F_1(\beta-\alpha+1,1;\beta+1;z)-1,$$

と表わせる．さらにこの 2 パラメータ分布族は，特別な場合として次の諸分布のうちの 4 種を含んでいる．

| | | |
|---|---|---|
| $(1,1,1)$ | Zipf 分布 | $p(x)=1\,/\,x(x+1)$, |
| $(\alpha,\alpha,1)$ | Yule 分布 | $p(x)=\alpha(x-1)!\,/\,(\alpha+1)^{\overline{x}}$, |
| $(1,\beta,1)$ | Yule-Simon 分布 | $p(x)=\beta\,/\,((\beta+x-1)(\beta+x))$, |
| $(\alpha,0,1)$ | Karlin-Rouault-Sibuya 分布 | $p(x)=\alpha(1-\alpha)^{\overline{x}}\,/\,x!$, |
| $(1,1,\gamma)$ | Zipf-Mandelbrot 分布 | $p(x)=1\,/\,x^{\gamma}-1\,/\,(x+1)^{\gamma}$. |

Waring-Herdan-Muller 分布，およびこれに含まれる 4 種類の分布は

$$p(x+1)\,p(x-1)-p^2(x)\geq 0,\quad x=2,3,\dots$$

という意味で対数凸であり，したがって凸である．Zipf-Mandelbrot 分布は，対数凸ではないが凸である．$p(x;\alpha,\beta,1),\ x=1,2,\dots,$ をずらした $y=x-1=0,1,2,\dots$ の pmf は

$$p(y)=\frac{\Gamma(\alpha+1)\Gamma(\beta+1)\Gamma(\beta-\alpha+1+y)\Gamma(1+y)}{\Gamma(\alpha)\Gamma(\beta-\alpha+1)\Gamma(\beta+2+y)y!}$$

であるが，これは一般超幾何分布 B3 型 GHgB3（ベータ負の二項分布）

$$q(y)=\frac{\Gamma(a+c)\Gamma(b+c)\Gamma(a+y)\Gamma(b+y)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(a+b+c+y)y!},\quad y=0,\ 1,\dots;a,\ b,\ c>0,$$

で $a=1,\ b=\beta-\alpha+1,\ c=\alpha$ の場合である．$q(y)$ については Sibuya (1979)，Sibuya and Shimizu (1981)を参照.

また $p(x;\alpha,\beta,1)$ は幾何分布 $(1-u)u^{x-1},\quad x=1,\ 2,\dots,$ で $u$ がベータ分布 $\mathsf{Be}(\beta-\alpha+1,\ \alpha)$ に従うときの混合分布である.

### 5.2 Karlin-Rouault-Sibuya 分布

Karlin-Rouault-Sibuya 分布は，本稿に関連する 3 つの分野で独立に現われる．事前分布を導入するモデルでも利用できそうである．これらの分野で現われることの本質ははっきりしていない．ここではいくつかの特徴を記しておく.

#### 5.2.1 分布の定義

KRS($\alpha$) により次の正整数上の分布を表わす.

$$p(x)=\frac{\alpha(1-\alpha)^{\overline{x-1}}}{x!}=(-1)^{x+1}\binom{\alpha}{x},\quad x=1,2,\dots;\ \alpha\in(0,1]. \tag{5.2}$$

$$\sum_{y>x}p(y)=\frac{(1-\alpha)^{\overline{x}}}{x!},\quad \frac{p(x)}{\sum_{y\geq x}p(y)}=\frac{\alpha}{x}, \tag{5.3}$$

$$\text{pgf}\quad G(z)=1-(1-z)^{\alpha},\quad \mathrm{E}(X)=\infty \text{ unless } \alpha=1.$$

$G(z)=(z-1)\,{}_2F_1(1-\alpha,1;1;z)-1$ とも表わせる.

KRS(1) の場合は，　$p(1)=1,\ p(x)=0,\ x>1.$

(5.2) は次のように表現することができる.

$$p(x) = \pi^{-1}\sin(\alpha\pi)\,B(\alpha+1, x-\alpha) = \pi^{-1}\,\Gamma(\alpha+1)\,\sin(\alpha\pi)\,x^{-\alpha-1} + O(x^{-\alpha-2}).$$

KRS(1/2) は,

$$p(x) = \frac{1}{2x-1}\binom{2x}{x}2^{-2x} = \frac{(2x-3)!!}{x!2^x} = 2C_{x-1}2^{-2x}, \tag{5.4}$$

$$x = 1, 2, \ldots; \quad (-1)!! = 1, \quad C_x \text{ は Catalan 数.}$$

これは原点から出発し，一定間隔で左右に等確率で動く酔歩が，$2x$ 回目に初めて原点に戻る確率である(random-walk distribution)．Feller (1968)，Chapter 3 参照．Catalan 数についての詳しい説明は Stanley (2001)参照.

Devroye (1993)は “Sibuya distribution” の名前を提唱し，乱数の生成を目的に，3 種の離散分布(離散安定分布，離散 Linnik 分布，Sibuya 分布)の関連を “triptych (三枚絵)” の表題で議論した．第 5.2.3 節参照.

### 5.2.2 分布の生成

1. ベルヌーイ試行列で $k$ 回目の成功確率が $\alpha/k$ のとき $k = 1, 2, \ldots$，$x$ 回目に初めて成功する確率が KRS($\alpha$) である．(5.3) のハザード関数参照.

2.
$$p(x) = \frac{\Gamma(c-a)\Gamma(c-b)}{\Gamma(c-a-b)\Gamma(a)\Gamma(b)}\frac{\Gamma(x+a)\Gamma(x+b)}{\Gamma(x+c)\Gamma(x+1)},$$

を一般超幾何分布と呼び $F(a, b; c)$ で表わす．特に $F(\alpha, \beta; \alpha+\beta+\gamma)$, $\alpha,\ \beta,\ \gamma > 0$, を B3 型一般超幾何関数(ベーター負の 2 項分布)と呼ぶ.

$$p(x) = \frac{\Gamma(\alpha+\gamma)\Gamma(\beta+\gamma)}{\Gamma(\alpha+\beta+\gamma)\Gamma(\gamma)}\frac{\alpha^{\overline{x}}\beta^{\overline{x}}}{(\alpha+\beta+\gamma)^{\overline{x}}x!}, \quad x = 0, 1, \ldots$$

$Y \sim F(1, 1-\alpha; 2)$ のとき $X = Y + 1 \sim$ KRS($\alpha$) である Sibuya (1979).

3. digamma distribution

$$p(x; \alpha, \gamma) = \frac{1}{\psi(\alpha+\gamma) - \psi(\gamma)}\frac{\alpha^{\overline{x}}}{x(\alpha+\gamma)^{\overline{x}}}, \quad x = 1, 2, \ldots;\ \alpha > 0, \gamma > 0,$$

において $\alpha + \gamma \to 0$ とすれば KRS($\gamma$)，Sibuya (1979).

4. Pitman 確率分割

自然数 $n$ の順序のない Pitman 確率分割における寸法指標を $S = (S_1, \ldots, S_n)$ とする.

$$P\{S = s; \theta, \alpha\} = \frac{n!\theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)}{\theta^{[n]}}\prod_{j=1}^{n}\left(\frac{(1-\alpha)^{[j-1]}}{j!}\right)^{s_j}\frac{1}{s_j!},$$

$$s = (s_1, \ldots, s_n), s_j \geq 0, j = 1, \ldots, n;\ k = \sum_{j=1}^{n} s_j, \quad \sum_{j=1}^{n} j s_j = n.$$

このとき

$$S_j \Big/ \sum_{j=1}^{\infty} S_j \quad \xrightarrow{\text{a.s.}} \quad \frac{\alpha(1-\alpha)^{\overline{j-1}}}{j!}, \quad j = 1, 2, \ldots$$

Pitman (1997)，Yamato and Sibuya (2000).

5. カテゴリー数無限の多項分布の寸法指標に関する中心極限定理(本文参照).

### 5.2.3 無限分解可能確率母関数との関係

1. 確率変数列 $(X_n)_{n=1}^{\infty}$ が独立で同一の $\mathsf{KRS}(\alpha)$ に従うとし，$N$ が幾何分布に従うとする：$P\{N=k\}=pq^k,\ k=0,1,\dots$ このとき $X_1+\cdots+X_N$ は確率母関数

$$G(z)=1/(1+c(1-z)^{\alpha}),\quad c=p^{-1}-1,\quad \alpha>0,$$

となる．この分布を discrete Mittag-Leffler 分布と呼び $\mathsf{DML}(\alpha,c)$ と書く．これは次の分布の特別な場合である．

2. 確率変数列 $(X_n)_{n=1}^{\infty}$ が独立で同一の $\mathsf{KRS}(\alpha)$ に従い，$N$ がこれらと独立な負の二項分布に従うとする．$P\{N=k\}=(\Gamma(\beta+k)/\Gamma(\beta)k!)p^{\beta}q^k,\ k=0,1,\dots$ このとき $X_1+\cdots+X_N$ は確率母関数

$$G(z)=1/(1+c(1-z)^{\alpha})^{\beta},\quad c=p^{-1}-1,\quad \alpha>0,\quad \beta>0,$$

をもち，無限分解可能である．この分布を discrete Linnik 分布と呼び $\mathsf{DL}(\alpha,\beta c,\beta)$ と書く．$\mathsf{DML}(\alpha,c)=\mathsf{DL}(\alpha,c,1)$ である． $c=\lambda/\beta,\beta\to\infty$ の極限は次の離散安定分布となる．

3. 確率変数列 $(X_n)_{n=1}^{\infty}$ が独立で同一の $\mathsf{KRS}(\alpha)$ に従い，$N$ がこれらと独立な平均 $\lambda$ のポアソン分布に従うとき $X_1+X_2+\cdots+X_N$ は確率母関数 $\exp(-\lambda(1-z)^{\alpha})=\exp(\lambda(1-(1-z)^{\alpha}-1))$ をもつ．この分布は Steutel and van Harn (1979)の意味で離散安定分布である．上の記号法では $\mathsf{DL}(\alpha,\lambda,\infty)$ である．この確率母関数は当然無限分解可能である．

確率変数 $T_\gamma$ がラプラス変換 $E(e^{-sT_\gamma})=e^{-s^\gamma}$, $\Re(s)\ge 0$, $0<\gamma<1$, をもつものとする．つまり特性関数が

$$\begin{aligned}\psi(t)&=\exp(-(it)^{\gamma}),t>0,\\&=\exp\left(ct^{\gamma}\left(1+i\tan\left(\frac{\pi\gamma}{2}\right)\right)\right),\ c=\cos\left(\frac{\pi\gamma}{2}\right),\end{aligned}$$

である．これから $T_\gamma$ は連続な狭義安定分布である．ポアソン分布 $\mathsf{Po}(\theta)$ の $\theta$ が確率変数で $\lambda^{1/\gamma}T_\gamma$ と同じ分布に従うとき，混合分布が $\mathsf{DL}(\gamma,\lambda,\infty)$ となる．Devroye (1993)，Christoph and Schreiber (1998，2000)参照.

### 5.2.4 零打切り負の二項分布

負の 2 項分布 $\mathsf{NgBn}\ (\xi,k)$:

$$p_0(x)=\binom{k+x-1}{x}\xi^k(1-\xi)^x=\frac{\Gamma(k+x)}{\Gamma(k)x!}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^x,\quad x=0,1,2,\dots,$$

$$\mu=\frac{k(1-\xi)}{\xi},\ 0<\xi<1,\ 0<k<\infty,$$

の零打切り分布は

$$p(x)=\frac{p_0(x)}{1-p_0(0)}=\frac{\xi^k}{1-\xi^k}\binom{k+x-1}{x}(1-\xi)^x=\frac{-1}{1-\xi^{-k}}\binom{-k}{x}(-1+\xi)^x \tag{5.5}$$

である．これは “Engen’s extended (truncated)negative binomial distribution”, Engen (1978)，と呼ばれている．特に $-1<k<0$ の場合にマトモな確率分布であることを指摘したことが extended である．そのとき $\xi=0$ とすれば $\mathsf{KRS}(-k)$ である．

## 5.3 データ公有化の環境(調査データ公有化の政治)

### 5.3.1 統計法

国勢調査を始めとする，全国規模の調査の多くは中央官庁によって行われている．この制度の根拠となっているのは，現行憲法の直前(1957 年 3 月)に制定された “統計法”である．統計

法は “統計の真実性を確保する” ことを第一の目的としている．統計法が対象とする官庁統計にも，指定統計，届出統計，承認統計の別があるが，簡単のため国勢調査を含む指定統計を念頭にして述べる．法理論のことは知らないが，常識的に読めば “統計調査は社会，国家の実体を正しく把握し，それを誤りなく人々に伝えることを任務とする” という理念を表明しているものであろう．

第 1 条の目的のために被調査者は報告を命令され義務づけられている(第 5 条)．選挙に行かないのは権利の放棄であるが，国勢調査に協力しないのは脱税と同じような，全体に協力しない，犯罪である．違反にたいしては罰則も設けられているが(第 19 条)，適用された例はないそうである．

調査に応ずることを義務づけるために，調査によって知られた，個人，法人，団体の秘密は，統計の集計，公表，目的外使用のすべての段階を含めて保護される(第 14 条)．これは公務員が，その職務執行によって得た情報を洩らしてはならない守務義務を補うものである．

さらに第 15 条では，“何人も… 調査票を統計上の目的以外に使用してはならない．” つまり税務，検察，公安などに個人データを利用することはないことを宣言する．家計調査の所得と税務署への申告の食い違いを追及されることはない．行政のための届出と分離することにより，事実を報告して欲しいという期待である．しかし “目的外使用” の配慮はされ，実際には利用されている．その際は，非調査者を識別できない方法で使用するよう限定している．

### 5.3.2 統計の真実性

この基本法も半世紀後には再検討が必要である．第 1 に “統計の真実性” という理念は理想主義的であって，目標の設定，効率，国際協調などの意識が薄い(それは法律の外のことかもしれないが)．これが制定された頃は第 2 次大戦中に，統計を軍事機密とし戦争を破局的に拡大したことにたいする反省が強かったのであろう．それ以降，人々の生活，意識，国際環境，統計技術が変われば，統計の目的，利用法も変わる．統計作成における官庁の役割にも影響が及ぶ．調査するもの，されるもの，の二分法では，互いに規制を課すだけとなり柔軟性を損なう．

生活環境，生産，経済の急速な変化に対応する統計の必要は絶えず指摘されている．何が必要な統計であるか国家が考え国民に命令するという感覚では，必要に対応できない．何が重要であるか，そのためにどこまで調査し，誰にどのように公表するのか，人々が議論し同意し協力する過程が重要である．

統計の真実性は官庁のキャビネットに格納されているものではなく，人々によって利用され検討されるものである．集団の事実は，どのような統計量でも公表する努力が必要であろう．それによって，統計の真実性，信頼性も改善される．日本の統計の質が高いという国際評価があるようだが，コストと比較した品質であろうか．人々の生活と思考が急速に変化している中で，これまでの品質を保てるのであろうか．さらに日本の統計が国際的に十分流通していないのは大きな損失である．

### 5.3.3 副次的分析と個人の秘密

公表されているのは調査されたデータを集計した一側面であって，収集されたデータの中には多くの統計的事実が記録され，潜んでおり，公表されている型の要約以外の要約，集約が可能である．その結果は調査の再設計にも活かすことができる．非回答の偏りを正し，記入誤りを調整する(pruning)仕事を正当に評価すべきである．調査対象の活動・生活様式，調査組織の人員構成，記録・通信・計数・機器の能力，などの環境変化によって非抽出誤差は変化する．それを測定するには，詳細なデータの慎重な分析を必要とする．行政記録との比較なども双方の品質管理の手段として役立つであろう．

被調査個体の秘密にはいろいろな水準がある．個体に法人，家族，個人の別があり，個人の

問題が深刻である．個人の秘密について，人々の感情も変化している．過去には身体障害者，難病患者は隔離され，その家族は疎外された．今では困窮する人が声を出し，周囲の人々が助けようとし，個人家族の秘密ではなくなっている．しかしプライバシーはあくまでも各個人の主観，感情であり，秘密の基準を強制することは難しい．罰則により報告を義務付けても，その実質効果がないならば，説明し同意を得た範囲でデータを得るのと同じではないだろうか．

個票を公開し，なお匿名を保つ，つまり公開されたデータから，それが誰か再識別することが事実上不可能であることを保証するためには，副次解析の精度が落ちることは犠牲にしなければならない．解析方法も限定され，たとえば極値解析はあきらめることになる．それでも，目的外使用の計画を立てるには十分に有用なデータとなる．

#### 5.3.4 研究者の倫理

データを分析しているとき特異な個体が研究者の目に入ることは可能である．多重回帰における外れ値などが典型的である．研究者は極端な外れ値の特性を調べた上で，分析のために残すか，除去するかを決定する．外れ値を調べることにより新しい発見もあるだろう．そのときのプロジェクトから外れた解析を直ちに実行，発表することはできないだろうが，中間結果を探索する自由まで制限すると，新しい研究の萌芽が妨げられる．研究者が好奇心をもち過ぎて漏洩に到るシナリオを心配する人もいるだろうが，研究者は自ら倫理綱領を作り，研究の自由を広げなければならない．だからといって，個人が誓約した上でデータに触れ，しかも対話的に分析する機会をはばむ理由にはならない．すべての分析結果について，たとえ見込み違いで新しい知見が得られなくても報告書を提出してもらい，査読，出版されることが望ましい．そのような分析そのものも業績として評価する習慣が必要である．

### 追　記

本稿 4.2 節特に命題 4.3 について査読者は「無限分解可能確率母関数すなわち無限分解可能確率変数（分布）ではない」ので，論文では前者を扱うことを明記するよう示唆されそれに従った．しかし「非負整数値確率変数が無限分解可能である必要十分条件は，その母関数が命題 4.3 の型にあらわせることである．」

Steutel, F.W. and van Harn, K. (2003). *Infinite Divisibility of Probability Distributions on the Real Line*, Chapter II, Theorem 3.2, p.30, Marcel Dekker, New York.

## 参　考　文　献


Baayen, R. H. (2001). *Word Frequency Distributions*, Kluwer, Dordrecht.

Bingham, N. H., Goldie, C. M. and Tengels, J. L. (1989). *Regular Variation*, Cambridge University Press, Cambridge, U. K.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review, *J. Amer. Statist. Assoc.*, **88**, 364–373.

Chen, W. C. (1980). On the weak form of Zipf's law, *J. Appl. Probab.*, **17**, 611–622.

Christoph, G. and Schreiber, K. (1998). The generalized discrete Linnik distributions, *Advances in Stochastic Models for Reliability, Quality and Safety* (eds. W. Kahle, E. von Collani, J. Franz and U. Jensen), 3–18, Birkhäuser, Berlin.

Christoph, G. and Schreiber, K. (2000). Scaled Sibuya distribution and discrete self-decomposability, *Statist. Probab. Lett.*, **48**, 181–187.

Devroye, L. (1993). A triptych of discrete distributions related to the stable law, *Statist. Probab. Lett.*, **18**, 349–351.

Engen, S. (1978). *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversities*, Chapman and Hall, London.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd ed., Wiley, New York.

Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population, *Journal of Animal Ecology*, **12**, 42–58.

Hill, B. M. (1974). The rank-frequency form of Zipf's law, *J. Amer. Statist. Assoc.*, **69**, 1017–1026.

Hill, B. M. and Woodroofe, M. (1975). Stronger forms of Zipf's law, *J. Amer. Statist. Assoc.*, **70**, 212–219.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes, *Journal of Mathematics and Mechanics*, **17**, 373–401.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events, Tech. Report, MS-R8804, CWI, Center for Mathematics and Computer Science, Amsterdam.

Khmaladze, E. V. and Chitashvili, R. Ya. (1989). Statistical analysis of large number of rare events and related problems, *Transactions of the Tbilisi Mathematical Institute*, **91**, 196–245.

Khmaladze, E., Nadareishvili, M. and Nikabadze, A. (1997). Asymptotic behaviour of a number of repeated records, *Statist. Probab. Lett.*, **35**, 49–58.

Orlov, J. K. and Chitashvili, R. Y. (1983a). Generalized $Z$-distribution generating well-known "rank-distributions", *Bulletin of Academy of Science*, *Georgia*, **110**, 269–272.

Orlov, J. K. and Chitashvili, R. Y. (1983b). On the statistical interpretation of Zipf's law, *Bulletin of Academy of Science*, *Georgia*, **110**, 505–508.

Pitman, J. (1997). Partition structures derived from Brownian motion and stable subordinators, *Bernoulli*, **3**, 79–96.

Read, C. B. (1988). Zipf's law, *Encyclopedia of Statistical Sciences*, Vol. 9, 675–676, Wiley, New York.

Rouault, A. (1978). Lois de Zipf et sources markoviennes, *Ann. Inst. H. Poincaré Sect. B*, **14**, 169–188.

Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distributions, *Ann. Inst. Statist. Math.*, **31**, 373–390.

渋谷政昭(1997). 多項分布における度数 0,1 のセルの数──漏洩管理のための基礎事実──, 応用統計学, **26**, 161–170.

渋谷政昭(1999). ミクロデータの公有化と利用の技術的課題, 日本統計研究所(法政大学)研究所報, No.25, 100–113.

Sibuya, M. and Shimizu, R. (1981). The generalized hypergeometric family of distributions, *Ann. Inst. Statist. Math.*, **33**, 177–190.

Stanley, R. P. (2001). *Enumerative Combinatorics, Vol. 2,* Cambridge University Press, Cambridge, U.K.

Steutel, F. W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability, *Ann. Probab.*, **1**, 893–899.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*, Lecture Notes in Statist., Vol. 111, Springer, New York.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Lecture Notes in Statist., Vol. 155, Springer, New York.

Yamato, H. and Sibuya, M. (2000). Moments of some statistics of Pitman sampling formula, *Bull. Inform. Cybernet.*, **32**, 1–10.

Zipf, G. K. (1949). *Human Behavior and the Principle of the Least Effort, An Introduction to Human Ecology,* Hafner, New York.

Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 261–295 (2003) 295

# Number of Categories with a Singleton in Sample and Population

Masaaki Sibuya

(Department of Business Management, Takachiho University)

A classical statistical problem is the study of a population with many categories. The main concern is not the probabilities of each category but their behavior as a whole when the sample size is increased. Typical examples are the ecological abundance of species, vocabularies in statistical linguistics, and patterns in archaeological artifacts. One aspect of statistical disclosure control (SDC), estimation of individuals who are unique in both population and sample, is related to the problem.

This review discusses the problem of estimating the number of those categories that have a unique element in a sample and its population, based on the observed sample.

The motivation to solve the problem in SDC is summarized in the beginning sections. The problem is shown to be difficult because it has the inverse problem feature. It is related to some classical problems in statistical abundance models, and the main results in this field are surveyed. These results are Zipf's law, the central limit theorem by Karlin, and the Large Number of Rare Events by a Tbilisi school.

New approaches are discussed in other papers of this special issue of the journal, in particular the use of infinitely divisible probability generating function. Other approaches, an application of the Ewens-Pitman family of random partitions and a semi-parametric inference method are related to Poisson mixtures.

Key words: Abundance models, infinitely divisible probability generating functions, large number of rare events, sample and population uniqueness, statistical disclosure control, Zipf's law.