超母集団モデルによる
寸法指標推測について

星野伸明

目 次

第1章　導入

1.1　寸法指標推測問題とは

「寸法指標」または「サイズインデクス (Size indices) 」(Sibuya[166]) は、「頻度の頻度 (Frequencies of frequencies)」(Good[62]) とも言う。動機付けの為に、壷のモデルを考える。N 個のボールが J 個の壷の中に分布しているとしよう。以下では N を非負整数、J を正の整数とする。ここで j 番目 $(j = 1, 2, \ldots, J)$ の壷にはボールが F_j 個入っている。ただし F_j は非負整数である。サイズインデクス S_i は、ボールの数が i の壷の数を示す。インディケータを

$$I(F_j = i) = \begin{cases} 1 & \text{if } F_j = i \\ 0 & \text{if } F_j \neq i \end{cases}$$

のように定義する。この時サイズインデクスを

$$S_i = \sum_{j=1}^{J} I(F_j = i), \quad i = 0, 1, 2, \ldots,$$

と定義する。もちろん $N + 1$ 以上の i について、$S_i = 0$ でなければならない。なお以下では、$\boldsymbol{S} = (S_1, \ldots, S_N)$ と書く。ボールの総数について

$$N = \sum_{j=1}^{J} F_j = \sum_{i=1}^{\infty} iS_i \tag{1.1}$$

が成立する。また空でない壷の総数を

$$U = \sum_{i=1}^{\infty} S_i \tag{1.2}$$

で表す。

$$J = U + S_0 = \sum_{i=0}^{\infty} S_i \tag{1.3}$$

である。

これらの壷から n 個のボールを標本として抽出する事を考える。ただし n は $n < N$ となる非負整数としよう。j 番目の壷からとられる標本ボールの数を f_j と書く。標本でのサイズインデクスを同様に

$$s_i = \sum_{j=1}^{J} I(f_j = i), \quad i = 0, 1, 2, \ldots.$$

のように定義する。標本数について

$$n = \sum_{j=1}^{J} f_j = \sum_{i=1}^{\infty} i s_i \tag{1.4}$$

となる。

$$u = \sum_{i=1}^{\infty} s_i \tag{1.5}$$

と書き、

$$J = u + s_0 \tag{1.6}$$

が成立する。

典型的な寸法指標推測問題は、データ $\boldsymbol{s} = (s_1, \ldots, s_n)$ から適当な $S_i, i = 1, 2, \ldots,$ を推測するというものである。なお U や複数の寸法指標が同時に推測対象になるかもしれない。本稿では母集団サイズが既知として、これらの寸法指標推測問題を考察する。以下で具体例を挙げるが、多くの状況では母集団サイズが分かっている。また母集団サイズが推測対象だとしても、推測された値を（所与の）外生変数とすれば本稿の方法論を用いる事が出来る。

以下の例 1.1 から 1.4 では、寸法指標推測問題の応用を手短に紹介する。このように本論文の主題は実際問題として重要性を持っており、5 章で特に重要な分野での応用について、詳しく述べる。

例 1.1 統計的生態学の群集モデル (Stochastic Abundance Model)

ある海域に住むプランクトンの種類と数を知りたいとしよう。しかし海洋調査などでは、全数調査はほぼ不可能である。故に標本抽出による、第 j 種の個体数が f_j という結果から推定せざるを得ない。一般に種数が非常に多いので、特定の種の個体数 F_j よりも個体数の少ない種と多い種の構成割合 $\boldsymbol{S}$ を扱うのが実際的である。このような場合、データ $\boldsymbol{s}$ 所与で寸法指標が推測される。—

例 1.2 語彙数の推測

シェイクスピアの語彙数はどのようにして知る事が出来るだろうか？作品の一部を抜き出し、単語の使用頻度を数える。ここで j 番目の単語が f_j 回使われているとしよう。ただ j 番目という順序には明らかに意味が無く、文章の単位長あたりの使用語数のみ分かれば良い。従って所与のデータを $\boldsymbol{s}$ と考えて、N を無限大に近づけた場合の U が語彙数として推測対象となろう。—

例 1.3 重複を持つデータベースのマージ

懸賞などで、応募の葉書が複数枚出される事は珍しくない。しかし多くの場合、興味の対象は重複を除いた応募者の数である。このような場合応募葉書の山から標本を取り、$\boldsymbol{s}$ をデータとして得る（葦番号 j は当然無意味である）。そして真の応募者数 U を推測すれば良い。一般に、重複を持つデータベースをマージする際、i 重に現れるレコードの数が S_i と考えれば良い。—

例 1.4 個票公開のプライバシー侵害リスク推定

ある個票データセットが、第j社会集団に属するf_j個体の標本を含んでいるとする$(j=1,2,\ldots,J)$。ここで適当な個人を特定する場合、F_jが小さい社会集団から特定する方が容易である。従って個票データセットを公開した場合に個人が特定される危険性は、w_iをiが小さい方が大きくなるウェイトとして、$\sum_i w_i S_i$で計るのが普通である。つまりそのようなリスク評価では、$\boldsymbol{S}$を推測する必要がある。—

なぜこの様に単純に見えるかもしれない問題が議論されるのか？初等的・古典的な母集団推測問題との違いについては、Engen[49](Chap. 2)のこなれた説明を参照されたい。ここでは導入としてしばらく、ナイーブな方法—母集団に関する仮定を用いない超幾何分布の下での推定を検討する。つまり大きさNが既知の有限母集団から非復元単純無作為抽出（2.2節を見よ）で取られたサイズnの標本より、寸法指標の推測を試みる。この場合の不偏推定（証明については付録Aを見よ）を確認しておこう。Goodman[63]によれば、もし$q:=\max_j F_j \le n$ならばUの唯一の不偏推定量が存在して、

$$\hat{U}_{Goodman}=\sum_{i=1}^{n} a_i s_i, \tag{1.7}$$

ただし

$$a_i = 1-(-1)^i \frac{(N-n+i-1)^{(i)}}{n^{(i)}},$$

ここで

$$n^{(i)} = n(n-1)\cdots(n-i+1)$$

は階乗積であり、この記法は本稿を通して使われる。そしてもし$q>n$またはq未知の場合、不偏推定量は存在しない。母集団におけるグループの大きさは通常未知（つまりqも未知）であり、不偏推定を重視するならば事態は単純でない。またもしqが既知だったとしても、nが小さい時、どうしたら良いだろうか。小標本での推定は応用上重要だが、不偏推定が出来ない場合も多い。加えてGoodmanは、(1.7)が極めて非合理な推定値を与えうると指摘する。つまり多くの状況で、$\mathrm{V}(\hat{U}_{Goodman})$が大きすぎるという事である。

またGoodman[63]が示すように

$$\mathrm{E}(s_i)=\sum_{l=i}^{q}\frac{\binom{l}{i}\binom{N-l}{n-i}}{\binom{N}{n}}S_l, \quad i=1\ldots,n,$$

ただし$q=\max_j F_j$である。ここから$q\le n$の時S_iの不偏推定量は、線型同時方程式

$$\begin{bmatrix} \mathrm{E}(s_1)\\ \mathrm{E}(s_2)\\ \mathrm{E}(s_3)\\ \vdots\\ \mathrm{E}(s_n)\end{bmatrix} = \frac{1}{\binom{N}{n}}\begin{bmatrix} \binom{1}{1}\binom{N-1}{n-1} & \binom{2}{1}\binom{N-2}{n-1} & \binom{3}{1}\binom{N-3}{n-1} & \cdots & \binom{q}{1}\binom{N-q}{n-1}\\ 0 & \binom{2}{2}\binom{N-2}{n-2} & \binom{3}{2}\binom{N-3}{n-2} & \cdots & \binom{q}{2}\binom{N-q}{n-2}\\ 0 & 0 & \binom{3}{3}\binom{N-3}{n-3} & \cdots & \binom{q}{3}\binom{N-q}{n-3}\\ \vdots & & \ddots & & \vdots\\ 0 & & \cdots & & \binom{q}{q}\binom{N-q}{n-q}\end{bmatrix}\begin{bmatrix} S_1\\ S_2\\ S_3\\ \vdots\\ S_q\end{bmatrix} \tag{1.8}$$

を解いて得られる。つまり (1.8) の右辺の行列の逆行列を、標本サイズインデクス $\boldsymbol{s}$ に掛ければ良い。なお Engen[49] の定理 2.1 によれば、もし $i \leq n$ ならば（標本寸法指標に基づく）唯一の不偏推定量が存在して、

$$\hat{S}_{iEngen} = \sum_{k=i}^{n} \sum_{j=i}^{k} \frac{\binom{N}{j}\binom{k}{j}\binom{j}{i}}{\binom{n}{j}} (-1)^{j-i} s_k \tag{1.9}$$

と書ける。しかし $S_i, i > n$, について、不偏推定量は存在しない。つまり $\hat{U}_{Goodman}$ の場合と同様、小標本での S_i の推定問題は単純ではない事がわかる。加えて (1.9) は分散が極めて大きい事が知られている（例えば渋谷 [168] の数値例を見よ）。なぜ不偏推定 (1.7),(1.9) の挙動は不安定なのだろうか。

まず (1.8) の右辺の行列がほとんど特異に近い事が、Shlosser[162] により指摘されている。また渋谷 [167](p.159) は、以下のように説明する。抽出率 n/N のサンプリングでは、$F_j n/N << 1$ のセル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 i のセルが S_i 個有ると、これから平均 $iS_i n/N$ 個の個体が抽出されるが、これらはほとんど標本で一意な個体（s_1 に含まれるという事）である。従って $\sum_{i<<N/n} iS_i$ 個の個体のうち平均 n/N が標本で一意な個体となり、この数はそれぞれの $S_i, i = 1, 2, \ldots,$ の値には依存しない。逆に s_1 の $S_i, i = 1, 2, \ldots,$ についての情報は少ない。$s_i, i = 2, 3, \ldots,$ を考慮しても、小さな i の S_i について情報は増えない。

そもそも以下の例 1.5 に見られるように、F_j が小さい時、F_j の推定は難しい。そして我々の応用領域では、多くの小さな F_j を取り扱う事になる。Khmaladze[100] はこのような特性を、LNRE (Large Number of Rare Events) と呼ぶ。3 章で後述される khmaladze の問題意識は我々のそれと少し異なるが、稀な事象の解析における困難を示唆する点で参考になる。Baayen[9] はその 2.4 節において LNRE 概念を用い、観測された相対頻度 f_j/n が F_j/N といかに異なるか、数値例を用いて説明している。要するに稀な事象が支配的な場合、大数の法則に多くを期待できないという事である。S_1 等の不偏推定は、特に標本抽出率が小さい時、不安定と考えられる。

例 1.5 ある j について $F_j = 1$ としよう。ここから標本が取られる確率は n/N である。つまり $P(f_j = 1) = n/N, P(f_j = 0) = (N-n)/N$ である。不偏推定量 $\hat{F}_j = f_j N/n$ の分散は $N/n - 1$ となる。標本抽出率 n/N が小さい時、$\hat{F}_j$ の分散は極めて大きい。—

また不偏推定量を導出する際 $\mathrm{E}(s_i)$ を s_i で置き換えているが、標本 s_i は整数なので $\mathrm{E}(s_i)$ からの偏差は避けられない。そしてこの差が推定量の分散を大きくするかもしれない。例えばこのような問題意識から、Good[61] は観測された s_i に関して平滑化を考えている。

ここまでの議論をまとめよう。寸法指標の推定は、特に小標本の場合問題をはらんでいる。また不偏推定が存在する時でも、多くの場合実用的な目安を与えない。だとすると寸法指標の推測では、何らかの工夫を考えざるを得ない。実際様々な著者が、先の例 1.1 から 1.4 のような多様な文脈で、広範な議論をしている。例えば統計的生態学の群集モデルはその代表であり、早くから研究されている。具体的には、Engen[49] が初期の結果をまとめている。また近年では U の推定について、Bunge and Fitzpatrick[24] のレビューが参考になる。なお同レビューによれば、満足できる結果は存在していない。

さて、このように寸法指標推測問題が議論を要するものだとして、工夫の糸口はどこであろうか。我々は問題の存在を確認するため、超幾何分布を利用した不偏推定が不十分な事を指摘した。だとしたらまず、不偏制約を捨てる事を考えるべきだろう。不偏でなくてもバイアスが小さい推定

量は、実用的でありうる。そもそも (1.9) 式による不偏推定は負の値を（頻繁に）とる為、負の推定値を 0 に切り上げるだけでも平均的なバイアスが減少する。しかし明らかにこれだけでは不十分であり、何らかの追加的制約が必要だろう。例えば Good[61] の提案した平滑化は、一般に不偏性を失わせる事になる。そして Good が示唆したように、平滑化は間接的に寸法指標の構造を仮定する事とも考えられる。平滑化の要求する構造制約が、母集団の真の構造と一致する場合のみ、不偏となる。そのかわり推定は、構造の要求する範囲に定まる。

このような、構造の仮定による一般性の喪失と推定の安定に関するトレードオフは、統計学における典型的なジレンマである（例えばノンパラメトリック法に関わる議論をあげる事が出来る）。そのため如何に構造を分析におりこむかについては、既存の文脈が大いに参考になるはずである。我々は、超幾何分布に依存した有限母集団解析の意味を探る所から始める事にしよう。

1.2 有限母集団解析

統計学の源流の一つは、国情記述の学問である。これは別の源流である古典的確率論とは、興味の焦点を異にする。つまり現実の母集団に対する問題意識の強さが、前者を特徴付けるように思われる。有限母集団解析は、このような実社会の必要性に応える事を目的としている。故に古くから研究されながらも依然として統計学の重要な領域であり、例えば近年も Levy and Lemeshow[110], Rao[141] 等、教科書が出版され続けている。なお Smith[189] が言うように、古典的確率論に依拠する大標本理論では、仮想的な無限母集団を司る法則が直接的興味の対象である。そのため調査の実務的要請との関係は明かではない。従って有限母集団解析は、出自において大標本理論とは異なる着想を持ちこんだように思われる。しかしながら確率の援用という枠組みが有る以上、相互に無関係ではあり得ない。本節ではこのような関連を見る事で、有限母集団解析の意味を考察する。

有限母集団とは有限個の要素（個体）の集合であり、各要素は一意に整数を用いて特定出来る。この整数を「ラベル」と呼ぶ。有限母集団を構成する個体のラベルの集合 $\Lambda = \{1, 2, \ldots, N\}$ を考えよう。標本は、母集団から適当に抽出された n 個体のラベル集合 $\tilde{s}_n \subset \Lambda$ で表される。変数 x について、ラベル λ に対応する個体の値を x_λ で表す。有限母集団解析の問題は、観測値 $\{(\lambda, x_\lambda) : \lambda \in \tilde{s}_n\}$ から適当な関数 $\Phi(\cdot)$ について、母集団の値 $\boldsymbol{x} = (x_1, \ldots, x_N)$ に関する量 $\Phi(\boldsymbol{x})$ の推測を行う事である。なお有限母集団からの標本抽出が無限母集団と異なるのは、抽出部分を所与のラベルにて特定できる事である。そのため抽出に関し、ラベルに依存した確率を導入できる。本稿で「標本設計」とは、q 番目にラベル λ のついた個体を抽出する確率 $\mathrm{P}(q, \tilde{s}_{q-1}, \lambda), q = 1, 2, \ldots,$ を事前に定める事を言う。ただし $\tilde{s}_0 = \emptyset$ である。なお本稿では観測された x に依存する標本設計は考察しない。このような設計は実用の観点から重要でないが、情報量の評価について問題になる。

前節の Goodman 等の議論では、標本設計として非復元単純無作為抽出を考えている。そしてある推定量が、設計に依存して不偏と主張された。有限母集団解析では、このような標本設計を利用した推測が古典的である。これは少なくとも、Neyman[123] までさかのぼる事が出来る。彼は有為抽出と確率抽出を、標本平均の誤差の大小という観点から比較した。有為抽出は「代表的標本」に関する何らかの標本選別基準を持っているはずだが、その基準が母集団構造に適合している場合、確率抽出より望ましいかもしれない。しかし通常、そのような基準の適合度を評価する事は不可能である。逆に確率抽出ならば、未知の母集団構造に依存せず信頼区間を構成できる。Neyman は未知の母集団に関する標本の挙動の予測可能性に、確率抽出の有為抽出に対する優越の根拠を求

めた。

このような古典的枠組みでの興味は (a) 推定量を固定して設計を選択、(b) 所与の設計の下で推定量を選択、(c) 設計と推定量の組を選択、のいずれかという事になる。Cassel et al.[29] がここでの選択基準等に関する包括的議論を与えており、(b),(c) について今は議論しない。(a) の場合、例えば単純無作為抽出と系統抽出では、いずれが小さな標本平均の分散を与えるだろうか？系統抽出の場合、標本平均の分散はラベルのつけ方に依存する。しかし個体とラベルの対応関係については、特に知識を持っていないのが普通であろう。このような状況は、ラベルに関して交換可能なモデルとして記述できる。すなわち、個体とラベルの全ての対応関係（$N!$ 通り）が同様に確からしいと考えればよい。(Ericson[50] は主観的ベイジアンの立場から、ラベルが個体について情報を持っていない状態をこのようにモデル化する。) このような場合、ラベルと個体の対応関係が変動する事で生ずる標本平均の分散を評価出来る。Madow and Madow[114] はこのように考えて、系統抽出と単純無作為抽出が平均的に等しい標本平均の分散を与えると述べた。

ラベルと個体の対応を確率的とみなすという事は、母集団を確率的とみなす事である。Cochran[38] はより一般的に考えて、所与の母集団を母数が未知の無限母集団からの確率標本とみなす事を提案した。無限母集団から母集団を確率抽出すると考えた時、様々な母集団が実現可能である。そして所与の母集団は、その実現可能な母集団の一つとなる。故に実現可能な母集団に共通する性質が分かれば、所与の母集団についても性質が明らかになる。つまり母集団の確率分布を仮定するという事は、所与の母集団を含むような一群の母集団が分析の対象になるという事である。Cochran はここで仮定される無限母集団を、「超母集団 (Superpopulation)」と呼んだ。超母集団モデルについては、インド学派の成果を踏まえた整理が分かりやすい（例えば Thompson[202] を参照の事)。超母集団の仮定とは、母集団の性質 $\boldsymbol{x}$ を確率ベクトル $\boldsymbol{X} = (X_1, \ldots, X_N)$ の実現値とみなす事である。超母集団モデルとは、$\boldsymbol{X}$ の同時分布 ξ の集合 $\{\xi\}$ である。例えば $\boldsymbol{X}$ の分布を定義する関数 $p(\boldsymbol{X};\theta)$ の集合

$$\{p(\boldsymbol{X};\theta)|\,\theta \in \Theta\}$$

が、超母集団モデルという事になる。

例 1.6 回帰モデル等では

$$X_h = \alpha + \beta y_h + \epsilon_h, \quad h = 1, \ldots, N,$$

但し ϵ_h は各 h 独立に平均 0・分散 σ^2 の正規分布 $\mathbf{N}$ に従い、y_h は個体 h の変数 y の値、α, β, σ^2 は未知母数と考える。これは超母集団モデルとして

$$\{\prod_{h=1}^{N} \mathbf{N}(\alpha + \beta y_h, \sigma^2)|\alpha, \beta, \sigma^2\}$$

を仮定したという事である。—

そもそも Cochran[37] が考えていたのは、系統抽出という標本設計が母集団の性質にどのように依存するかという事である。もしいかなる母集団についても望ましいような設計・推定量が存在しないなら、母集団への依存を評価するのは自然である。実際（設計に依存する）不偏推定量のクラスを考えて、どのような母集団についても最小の分散を与える不偏推定量は存在しない。この事について、Basu[13] は簡単な証明を与えている。また Godambe[58] は広範な線形不偏推定量のクラ

スについて、最良不偏推定量を考察した。しかしそこで提案されたクラスでは、いかなる母集団についても望ましい標本設計・推定量は存在しない事が示される。その為母集団をモデルで限り、特定の設計・推定量が良いと示した。Cochran 以後、Raj[140] 等が超母集団を明示して標本設計の比較評価をしている。当初「超母集団」が専門用語として定着していったのは、このような文脈においてであった。なお Smith[189] の整理によれば、これらは Neyman 流解析の前提が一般的に過ぎる事を示唆する最も初期の例という事になる。つまり少なくともある種の問題については、母集団の性質を仮定しない限り、十分に鋭い結果を得る事が出来ないように見える。

このように超母集団は、ある有限母集団群の平均的性質を評価する為に導入された。そしてこれは結果的に、有限母集団分析から有限性を取り去るように働く。言いかえれば、超母集団モデルを仮定する事で有限母集団解析は、形式的に無限母集団解析と同等になりうる。例えば $\{\xi\}$ が数個の母数で定められる分布の集合ならば、まずその母数を推定する。そして ξ が定まれば、興味の対象の推測値 $\Phi(\boldsymbol{x})$ 等が得られる。つまり ξ を推定しそれを用いて興味の対象を計算する、このような方法を Neyman 流と区別して「予測アプローチ (Prediction Approach)」という。この立場で書かれた文献として、Cassel et al.[29], Bolfarine and Zacks[19], Valliant et al.[209] を挙げておく。

予測アプローチでは、標本は所与とみなされる。すなわち観測されなかった部分は、観測された値が所与の条件付分布を用いて「予測」される。これに対し Neyman 流では、標本と抽出されなかった残りの部分をつなぐ論理は、抽出されたかもしれないという可能性である。一度標本が観測されてしまうと、古典的推測すなわち「推定」は成立しない。

例 1.7 (Basu[13]) 母集団の値 $x_1+\cdots+x_N$ を推定するものとする。標本として $(1,x_1),\ldots,(n,x_n)$ が観測されたとしよう。条件付に考えて $x_1+\cdots+x_n$ が既知なので、$x_{n+1}+\cdots+x_N$ の推測問題になる。観測された標本は、観測されなかった部分とどのような関係が有るのか？—

Godambe[59] 等の議論によれば、例 1.7 において Neyman 流に考えた場合、母数は $(x_1,\ldots,x_N)$ とみなす事ができる。ここで確率は、ラベルについて付与されている事に注意すべきである。そして標本設計は、母数の値に依存しない。すなわち母集団に対する仮定を置かない限り、全ての $\boldsymbol{x}\in\mathbf{R}^N$ について $\mathrm{P}(q,\tilde{s}_{q-1},\lambda)$ は定数である。この意味でラベルは、母数に関する情報を持っていない。

このような典型的状況で、Basu[12] は最小十分統計量が（繰り返しを除いた）観測値、すなわちラベルと x の組である事を指摘している。一般に、母数に関する推測で差異を産まない情報は、十分性の観点から等価である。つまり Basu の結果を解釈すると、非復元で取る以上の標本設計の工夫は情報縮約の観点から望めないという事になる。また尤度原理（Birnbaum[16], Barnard et al.[10] 等を見よ）の立場からは、全ての標本設計は同等である。何故なら母数に関して情報を持たない設計では、尤度関数が平らになる。そして定数倍の尤度関数は「統計的根拠」において同一視されるので、尤度原理の支持者は標本設計に重大な価値を認めないのである。

ここまで、構造に関する仮定を用いない標本設計に対する経験的・原理的異論を見てきた。実際効果的とされる標本設計は、母集団に関する既知の補助情報を利用する。例えば Smith[189] によれば、母集団に関する情報を明示的に取りこんだ悪いデザインの推定量は、母集団に関する仮定を用いない良いデザインの推定量を経験的に優越する。代表的な標本設計として層化抽出が挙げられるが、これは層内の同質性に関する既知の情報を利用する。この場合 Ericson[50] が示唆するように、観測する値と未知母数の間で、尤度として表現されていない類似性をあてにしている。尤度原

理の含意は、いかなる補助情報も尤度に反映される形、すなわち確率モデルとして明示的に表現するべきだという事だろう。超母集団モデルは、方法としてこのような要請も満たす事が出来る。そしてRoyall and Herson[144]曰く、多くの標本調査での問題は便利かつ現実的に、適当な超母集団モデル下の予測問題として分析できる。

なお主観的ベイジアンの立場からは、超母集団モデルは自然な仮定である。母集団を所与とした標本抽出に関して主観の介在する余地は無いので、問題は母集団の事前分布となる。Ericson[50]の定式化では、$\boldsymbol{X}$ の事前分布を次のようにして求める。

$$p(\boldsymbol{X}) = \int_{\theta\in\Theta} p(\boldsymbol{X}|\theta)dF(\theta),$$

ただし $F(\theta)$ は（主観的な）未知母数の事前分布である。ここで $p(\boldsymbol{X}|\theta)$ が、超母集団に対応する。つまり今までの議論との違いは、未知であった超母集団の母数に主観確率が付与される点である。超母集団モデルそのものを事前分布とみなして良いのではないかとも思うが、Ericsonの議論ではラベルに関する交換可能性を母数 θ 所与で各 X_i が独立同一分布に従う事と考えている。すなわち

$$p(\boldsymbol{X}|\theta) = \prod_{i=1}^{N} p(X_i|\theta)$$

とするので、超母集団モデルを入れる意味がある。そしてこの事前分布の下で、（尤度原理を持ち出すまでも無く）標本から母集団に関する事後分布の導出が可能となると主張されている。

Neymanは母集団に関する偏見からの自衛として、先見情報に依存しない方法を導入した。しかしその後の研究は、先見情報を利用する方向に進んでいる。問題は偏見と補助情報の区別を如何に行うかという事になる。これは「代表的標本」が何かという問題と同じ困難を抱えている。とはいえ各論のレベルで、使用する補助情報が真実から乖離した場合の頑健性研究が行われている（例えばRoyall and Herson[144][145], Scott et al.[156], Tam[201]を見よ）。またおそらく補助情報の形態の多様性から、各論をせざるを得ない。そして超母集団モデルもまた、補助情報の形態の一つとして考える必要がある。

ここまでの議論を受けて、超母集団モデル ξ の解釈を考えよう。原理的問題に触れた事からも分かるように、唯一の答は存在しない。確率に対する態度の違いによって、大きく三分類する。(a) 分布 ξ は、実世界の確率的機構・過程を頻度的に記述する。(b) 分布 ξ は主観的信念を表した（ベイズ的）事前分布である。(c) 分布 ξ は、議論を明確にする為の単なる数学的道具にすぎない。—すなわちこの問題をつきつめると、統計学の基礎論に到達する。そして基礎論は、本稿の視野を超えるものである。これらの一般的議論に代えて、次節で本稿における解釈を述べる事にする。

1.3 本論文について

構造を仮定しない古典的アプローチが、寸法指標推測問題について実用に耐えない事を1.1節で説明した。さらに前節の認識を踏まえて、我々は寸法指標推測問題についてどのようなアプローチを取るべきか。思うにまず構造として導入したい先見知識を明らかにし、それを定式化するにふさわしい方法を考えるべきだろう。従って、応用分野に依存した議論は避けられない。ここで探索的データ解析がグラフの解釈を重視するように、我々の一義的認識は寸法指標のプロットからもたらされる。

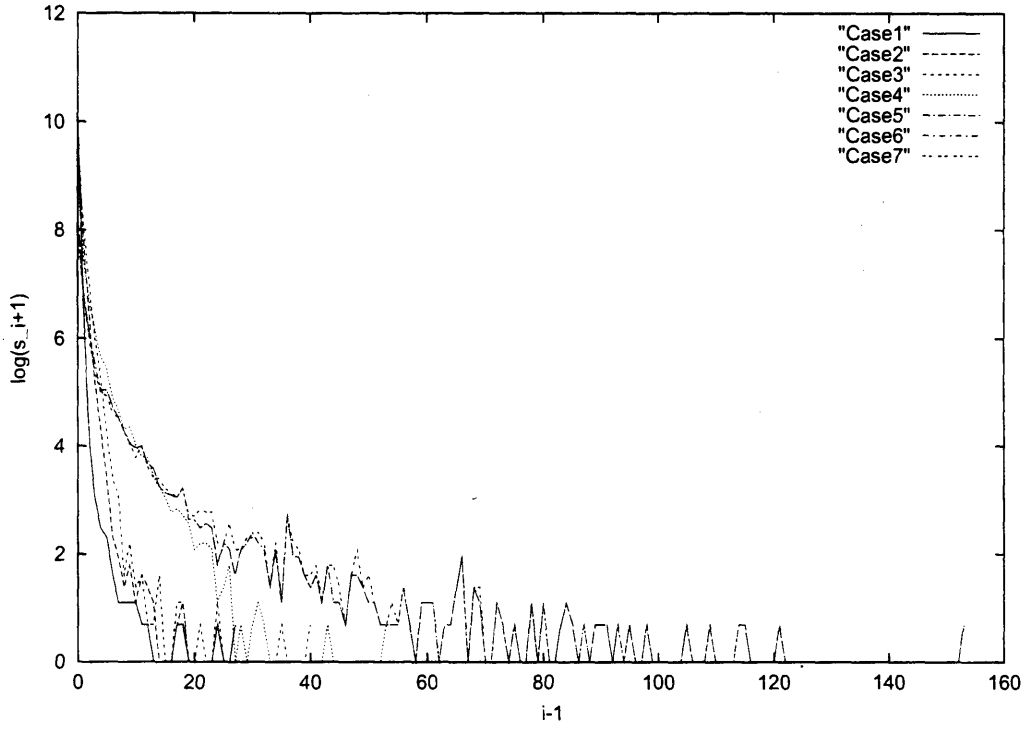

図 1.1: 労働力調査より (例 1.8)

例 1.8 労働力調査 (竹村 [198])

関東周辺の 9 都県における労働力調査（1995 年 1 月分）の個票データから、竹村 [198] は匿名化の程度を変えて 7 種類のデータセットを作成した。レコード数 n は 27230 であり、それぞれ寸法指標が提供されている。—

例 1.8 で説明されているデータセットについて、縦軸を $\log(s_i+1)$ で横軸を $i-1$ として描いたのが、図 1.1 である。Case 1 から 7 の全ての場合で、長い右裾が観測できる。そして Zipf[234], Mandelbrot[115] 等が指摘するように、様々な分野の寸法指標は例 1.8 と同様の傾向を示す（他に 5.4 節の数値例を参照のこと）のである。

例えば単語の使用頻度、個人所得や都市の人口を、大きさ順に並べるとする。一般に、順序統計量 $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(n)}$ で考えよう。ここで $r = 1, 2, \ldots, n$ について、ある定数 c が存在して

$$r \cdot x_{(r)} \doteq c \tag{1.10}$$

のように書けると Zipf は主張した。様々な分野でこのようなデータが観測される為、これは「Zipf の法則」と呼ばれている。この経験則は寸法指標に関して、以下のような含意を持つ (Urzúa[208])。大きさ x の事象の相対頻度を $f(x)$ と書く。この時大きさ x の事象の順位 $R(x)$ について、

$$R(x) = n \int_x^\infty f(z)dz$$

である。(1.10) が成立するならば、$f(x) = c/n \cdot 1/x^2$ となる。従って、寸法指標に関して

$$S_i \propto \frac{1}{i^2}$$

と主張されている事になる。同様に Pareto[131] は、所得分布の研究から

$$S_i \propto \frac{1}{i^a},$$

ただし a は適当な定数、としている。

残念ながら Zipf の法則のあてはまりは常に良好とは言えず、改善の試みが続けられている（Johnson et al.[90] の 11.20 節を見よ）。従って $\log S_i$ が $\log i$ に比例するという仮定は強すぎるかもしれない。故に、より大きな分布のクラスを探求の対象とすべきだろう。例えば Zipf の法則を分布の凸性として理解すれば、これが可能となる。

定義 寸法指標が

$$S_{i+1} - S_i \geq S_i - S_{i-1}, \quad i = 1, 2, \ldots, \tag{1.11}$$

を満たすとき、凸であるという。

定義 寸法指標が

$$\frac{S_{i+1}}{S_i} \geq \frac{S_i}{S_{i-1}}, \quad i = 1, 2, \ldots \tag{1.12}$$

を満たすとき、対数凸であるという。

例えば（本来寸法指標は非負の整数だが）$S_i = 1/i^2$ であるならば、寸法指標は対数凸である事を示すことが出来る。なお対数凸ならば凸である。しかし凸性は、一つの可能性に過ぎない。ここでは凸性も含めて経験的に現れる分布の形状を、厳密ではないが「L 字型」と形容する。（なお Simon[182] はこれを “J-shaped” と呼び、Chen and Keller-McNulty[36] は “inverse-J shape” と呼ぶ。）応用局面では、このような先見知識が利用できるだろう。

とはいえ視野を応用の文脈に限る事で、本稿の主題は一般的な寸法指標推測問題ではなくなる。この事は明記されるべきだろう。1.1 節で述べたように、一般的な寸法指標推測では不十分な精度しか確保できていない。そのため議論をする余地が有るという事であった。これは直感的な意味で情報の欠如が原因であり、構造追加の巧拙が議論の焦点になる。本稿では、応用分野の先見情報を構造として追加する事に、一般性の喪失以上の意義を見ている事になる。そして本論文の目標は、寸法指標推測応用問題で用いられる超母集団モデルの考察である。なおその他の寸法指標推測問題へのアプローチは、ここでは追求しない。しかし特に s 以外の情報が利用できる場合、複雑さを厭わなければ推定を改善できるだろう。

個別応用分野の研究史は 5 章でそれぞれ簡潔なサーベイを与えるが、近年の研究は盛んであるとは言えない。文脈の主流をなす統計的生態学モデルの理論的研究は、1980 年代には峠を超えたように見える。しかし現在新しい研究分野が、立ち上がる過程にある。Bethlehem et al.[15] は、個票開示リスクの超母集団モデルによる評価という問題を提唱した。そしてこれが寸法指標推測問題の一応用例だという事は、それほど広く知られていない。故に個票開示リスク評価の立場から包括的に文脈を解釈した文献は存在せず、他応用分野との関連も十分明かにされていない。後に紹介さ

れる近年の著者等の研究は、これらの疑問に答えようとするものである。本稿では更に各分野の文献調査結果を、寸法指標推測問題として統合する。特に知られていない最新の成果が各分野と関連している事を明らかにする事で、研究の活性化を計りたい。

以下では私見を交え、分布形を限った超母集団モデルを応用する意味を更に考察する。ここまでの議論では、分布形（L 字）に関する先見知識を補助情報として使用するという事であった。そしてこの先見知識が過去に繰り返し起きた事象に基づいていると考えれば、母集団を生成する確率メカニズムが同定の対象になるのは自然である。この場合、Neyman 流の分析とも自然に接続する。しかし Lehmann and Casella[108] の 4 章で指摘されているように、過去の構造が現在の問題と共通するかどうかの判断は "extrastatistical judgment" である。そのため補助情報が主観の記述だという解釈との差異は、極めて微妙である。だとすると、ベイズ的解釈で良いかもしれない。

しかし漠然とした先見知識が事前分布（この場合 ξ）として記述できるとは限らない。今回の問題について言えば、我々の立脚点は「L 字」という漠然とした情報である。決して特定の確率分布だという確信が有るわけではない。確かに事前分布は経験と利便性を勘案して決めるものであり、恣意的に選択しても良いかもしれない。しかし少なくとも、その母数に関する不確実性は分析の枠組みに取り入れるべきではないか。このように考えた場合、二つ方法が知られている。一つ目がいわゆる「経験ベイズ」法で、母数をデータから推定する。この場合仮定される分布が頻度的裏付けを持たないと、論理的整合性に疑問が残る。言いかえれば、主観を推定するのは奇妙である。なお経験ベイズ法はモデルの誤特定について比較的頑健 (Lehmann and Casella[108], Chap. 4) と考えられており、この点で望ましい。そしてもう一つの方法は、「階層的ベイズ」である。これは不確実な母数が、主観によって定められる分布に従うと考える。前節で言及した Ericson による超母集団モデルの解釈が、まさにこれに該当する。そこでは事前分布を主観的な（母数に関する）混合分布と考えていて、通常のベイズ法と本質的に変わらない。

結局ベイズか非ベイズかという「踏絵」を踏まざるを得ないと思われる。本稿では、非ベイズの枠組を採用する。すなわち、経験ベイズ法を用いる。なお Berger[14] が指摘するように、経験ベイズ法は母数の推定誤差を考慮していない（この点は後に再考する）。また階層ベイズはそのような誤差を構造に取りこんでいる事になるが、その分計算が複雑である。著者は、利便性の観点から前者を選択したという事である。そもそも Berger も言うように、事前分布の母数の誤特定の影響は、事前分布の形状の誤特定に比べれば小さい。なお Berger は、他にもベイズ的方法論に関する興味深い考察を与えている。

超母集団モデルアプローチの採用については、もう一つ別の論点がある。どのような補助情報が使えるかという判断は、標本を見てから行う事ではないか。少なくとも「L 字」という情報が使えるかどうかは、標本の寸法指標構造 $\boldsymbol{s}$ から見当をつける事が出来る（もし標本が L 字とかけ離れていれば、本稿の方法は役立たずである）。このように考えれば、予測アプローチが論理的に必然となる。

著者の考える方法論では、L 字型のモデル集合を漠然とした先見情報に対応するものとして提供、データを見る事で L 字で良いと判断できるなら、情報量基準等で集合の中からモデルを選択する。そして、選択されたモデルによる推測を採用するというものである。なお「推測」の詳細については 3 章で考察される。

我々の前提である L 字という曖昧な認識は、特定のモデルというよりモデルの集合によってより正確に表されるのではないか。L 字という条件を満たすモデルを全て書き出す事は出来ないが、そ

の中で使われた・使えるモデルのリストを示す事は出来る。著者のすべき事はリストの整備だと考える。なお実務家はリストから適当にモデルを取捨選択して、集合を構成すれば良い。もちろんこのモデル集合の要素の数は、一つでも許されるだろう（この場合モデル選択の議論は不要となる）。しかし直感的にはこの要素の数が多い方が、母集団に関する偏見について頑健と思われる。なお著者自身は、モデルは便宜的思考である以上、利便性・柔軟性の観点からモデル集合を構成する。しかしモデルは実機構の記述だと考えるので有れば、そのように解釈できるモデルのみ用いれば良い。そして選択されたモデルの構造から、実機構について何らかの洞察を得られるかもしれない。

さて、問題は固定した集合からのモデル選択の方法である。選択のルールが妥当ならば、頑健な枠組みになるはずである。なおモデル選択は統計学の重要な主題であり、他の文献で慎重に考察されている。例えば Atkinson[7], Konishi and Kitagawa[106] を参照の事。Hoshino[80] では現実の個票データから寸法指標を推測する際、赤池情報量基準 (AIC) をモデル選択基準として使用したが、これは一例でしかない。他に、ピアーソンの χ^2 統計量の使用例なども統計的生態学で散見される。思うにモデル選択基準は、データ分析者の責任において選択されるしかない。

そもそも各応用分野での理論は、現実を抽象化したものと考えられる。従って理論と現実の差に関するいくつかの判断は、分析者の良識に任せざるをえない。本稿の問題について注意を喚起しておこう。まずL字のような構造が繰り返されるとしても、せいぜい経験則でしかない。私見を述べれば、母集団の真の分布はそもそも特定出来ない。言いかえれば、実母集団における寸法指標の構造がどのように生成されるかは、不可知である。表面的には中心極限定理を用いて説明する議論があるが、それは例えばどのように種が資源を分割するかという実構造の説明ではない。モデルを実構造に対応しているように解釈する事は出来る（そしてかなりの人々が解釈を重視しているように見える）かもしれないが、解釈が正しいかどうかは誰にも分らない。

例えば後述されるガンマ＝ポアソン分布（負の二項分布）は、多くの応用例を持つ基本的モデルである。にも関わらず、Cassie[30] の主張するガンマ＝ポアソン分布の賢明な解釈は、経験的かつ表面的な度数の記述である。何故なら Anscombe[6] が示すように、いくつかの異なる構造仮説からガンマ＝ポアソン分布モデルを導く事が出来る。そして観測値 $\boldsymbol{s}$ からは、同じ分布につながるような構造仮説は識別出来ない。特に Bliss and Fisher[18] が指摘するように、相反する二つの構造仮説から同じ $\mathrm{E}(S_i)$ の系列が生成される例さえ有る。また Feller[55] の 2.4 節における、ロジスティック分布関数に関する警告を見よ。多くの相容れない理論モデルが同じデータによって支持されてきた実例が、ここにも紹介されている。もう一点注意しておこう。分布のあてはまりが検定されるとして、帰無仮説の受容は帰無仮説が正しい事を必ずしも意味しない。

要するに我々は、度数の表面的挙動から真の構造を特定できない。このように考えると、モデルの解釈にこだわる意味は薄れる。そして寸法指標の推測が一義的問題の場合、モデルに要求される資質は、寸法指標の表面的挙動を柔軟に記述出来る事であって、実構造に関する解釈が出来る事ではない。もちろん実構造に対応しているモデルは柔軟な記述が出来るはずだが、柔軟なモデルを解釈が分らないからといって捨てる事はない。単に経験的に妥当であれば、考察する価値が有ると考える。我々にとってモデルの価値は、標本と母集団の関係が明かになる事である。なお生態学分野におけるモデルの記述統計学的使用を例に挙げたが、このような目的意識の下での分析手法は、必ずしも寸法指標推測という目的に適合しない事も指摘しておく。

結局、経験則を手もとのデータ、今回の事象に適用するか否かは統計学内の議論では決められない。どのようなモデル集合を作り、どのようにモデルを選択するか。利便性が個人的なものである

以上、恣意性は必ず残ってしまう。そして結果の解釈こそ万人の恣意に委ねられている。このような中で我々の出来る事は、仮定から演繹の過程を明示する事である。超母集団モデルという形で仮定を明示して議論をすれば、最善が尽くされたと考える。

第2章　推測の方法論

本章では、ある超母集団モデル所与で寸法指標を推測する方法を考察しよう。2.1節では経験ベイズ法の定式化、及び推測精度を議論する。2.2節では本稿における標本分布の意味を明らかにする。また2.3節では、頻度のモデルと頻度の頻度（寸法指標）モデルの関係について考察する。

2.1　推測の精度について

まず1章で議論した方法論を、より具体的に記述しよう。標本データとして $\boldsymbol{s} = (s_1, \ldots, s_n)$ が与えられたとする。また、このデータに適合するような超母集団モデルを選択したとする。興味の対象が適当な S_i の場合、推測値とその精度についてどのように考えれば良いか。

大きさ N が既知の母集団に関する予測アプローチを試みる。この場合 n 個体が観測されたものとして、残りの $N-n$ 個体を推定対象と考える。例えば観測された個体のラベルの集合

$$\tilde{s}_n = \{i_1, i_2, \ldots, i_n\}$$

が得られたものとする。観測されなかった部分のラベルの集合を

$$\tilde{r}_n := \{1, 2, \ldots, N\} \setminus \tilde{s}_n$$

と書く。つまり推測は壷番号 $x_l, l \in \tilde{r}_n$, について行われる。実現している寸法指標は

$$S_i = \sum_{j=1}^{J} I((\sum_{l=1}^{N} I(x_l = j)) = i)$$

と書けるので、予測アプローチを用いた寸法指標推測は

$$\hat{S}_{i\,Predic} = \sum_{j=1}^{J} I((\sum_{l \in \tilde{s}_n} I(x_l = j) + \sum_{l \in \tilde{r}_n} I(\hat{x}_l = j)) = i), \tag{2.1}$$

ただし $\hat{x}_l$ は l 番目の個体の推測された壷番号、という事になる。ここで $x_l, l \in \tilde{r}_n$, は、観測値 $x_l, l \in \tilde{s}_n$ 所与の条件付分布 $\mathrm{P}(\cap_{l \in \tilde{r}_n} x_l | \cap_{l \in \tilde{s}_n} x_l)$ から予測される。

しかし問題は、標本データ $\boldsymbol{s}$ において壷番号を識別できない事である。つまり (2.1) 式の $I(x_l = j)$ の部分を評価出来ない。従って、標準的な予測アプローチは適用不可能という事になる。故に我々は、予測量を

$$\hat{S}_{i\,Aprox} = \sum_{j=1}^{J} I((\sum_{l \in \tilde{s}_n} I(\hat{x}_l = j) + \sum_{l \in \tilde{r}_n} I(\hat{x}_l = j)) = i) = \sum_{j=1}^{J} I((\sum_{l=1}^{N} I(\hat{x}_l = j)) = i) \tag{2.2}$$

でおきかえる事にする。このような推測について意味を検討しておこう。(2.2) 式は、「新たに大きさ N の母集団を（同じ超母集団から）抽出した場合の寸法指標 S_i」の推測式と形式的に同等である。つまり既に観測済みの値を、推測値でおきかえている。故にこのような場合、x_l を条件付分布 $\mathrm{P}(\cap_{l\in\tilde{r}_n} x_l | \cap_{l\in\tilde{s}_n} x_l)$ から推測するのは論理的におかしい。全く新たに母集団を抽出するとして、分布 $\mathrm{P}(x_1,\ldots,x_N;\theta)$ から $\hat{x}_l$ を得るべきである。そして標本 $\boldsymbol{s}$ は、超母集団の母数の推定値 $\hat{\theta}$ を得る為に使われる。結局、興味の対象が既に抽出された母集団だとしても、寸法指標推測問題では便宜的に別問題を考えざるを得ないと思われる。もっとも $N >> n$ の場合、問題の置き換えによる影響は限られる（そもそも $N \doteq n$ ならば超母集団アプローチをとらなくて良い）。そして (2.2) 式のように考えれば、$\hat{F}_j = \sum_{l=1}^{N} I(\hat{x}_l = j)$ として

$$\hat{S}_{i\,Aprox} = \sum_{j=1}^{J} I(\hat{F}_j = i)$$

のように問題を単純化出来る。つまり、X_l の構造まで遡らなくて済む。本稿ではこのように考えて、頻度のモデル $\mathrm{P}(F_1,\ldots,F_J;\theta)$、ただし θ は母数（ベクトル）、を出発点とする。

予測尤度アプローチ（Bolfarine and Zacks[19] の 4 章を見よ）の観点から考え方を整理しよう。観測値 $\boldsymbol{s} = (s_1,\ldots,s_n)$ から未知母数 θ と未知数 S_i を推測すると考える。未知量の同時尤度関数を

$$l_{\boldsymbol{s}}(S_i,\theta) = \mathrm{P}_\theta(\boldsymbol{s}, S_i)$$

のように表す。真の興味の対象 S_i の予測尤度関数 $L(S_i|\boldsymbol{s})$ は、$l_{\boldsymbol{s}}$ から θ を除く事で得られる。ここで θ は局外母数であり、これを除く方法が異なれば異なる予測尤度関数を得る。ところが frequentist にとって、局外母数の問題は非常に厄介である。例えば Bjørnstad[17] では、局外母数を除く 14 通り (!) の考え方が紹介されている。このような状況では、利便性の観点から手法を選ぶのもやむを得ないと思われる。従って最も単純に考えると、θ を最尤推定値 $\hat{\theta}(\boldsymbol{s})$ で置き換える事になる。このような方法を「推定的（Estimative）アプローチ」と言う。すなわち、

$$L_{Estimative}(S_i|\boldsymbol{s}) = \mathrm{P}_{\hat{\theta}}(S_i|\boldsymbol{s}),$$

と考える。しかし Aitchison and Dunsmore[4] や Butler[25] が強調するように、このような予測尤度はミスリーディングなまでに正確になるだろう。経験ベイズ法が母数の推定誤差を取りこまないのを Berger[14] が批判した事は既に述べたが、これは軌を一にする指摘である。従って点予測はともかく、区間予測の際は注意が必要となる。

まず点予測から考える事にしよう。最大予測尤度予測量は、$L_{Estimative}(S_i|\boldsymbol{s})$ を最大化するような S_i である。また都合の良い事に、標本とは独立に大きさ N の母集団を新たに抽出するという事なので、$\mathrm{P}_{\hat{\theta}}(S_i)$ の S_i に関する最大化を考えれば良い。しかし必ずしも $\mathrm{P}(S_i)$ が容易に得られるわけではない。また最尤予測値が一意に決まるとも限らない。ここで我々は決定理論的に考えて、予測量 $\hat{S}_i$ のベイズリスクを最小化する事にしよう。すなわち、もし損失関数として二乗誤差 $(\hat{S}_i - S_i)^2$ を使用するならば、リスク $\mathrm{E}_\theta[(\hat{S}_i - S_i)^2]$ が最小になるような $\hat{S}_i{}^*$ が望ましいとする。つまり $\hat{S}_i{}^* = \mathrm{E}_\theta(S_i)$ とすべきである。この時最小化されたリスクは、分散 $\mathrm{V}(S_i) := \mathrm{E}({S_i}^2) - \mathrm{E}^2(S_i)$ である。なお $\mathrm{P}(F_j)$ を用いて表現した寸法指標の一次と二次のモメントは、以下の通りである。

$$\mathrm{E}(S_i) = \sum_{j=1}^{J} \mathrm{P}(F_j = i). \tag{2.3}$$

$$\begin{aligned} \mathrm{E}(S_i{}^2) &= \mathrm{E}[\sum_{j=1}^{J} I^2(F_j = i) + \sum_{j=1}^{J}\sum_{l\neq j} I(F_j = i)I(F_l = i)] \\ &= \mathrm{E}(S_i) + \sum_{j=1}^{J}\sum_{l\neq j} \mathrm{P}(F_j = i, F_l = i). \end{aligned} \tag{2.4}$$

故に我々は $\boldsymbol{s}$ から母数の推定値 $\hat{\theta}$ を得、

$$\hat{S}_{i\,LeastRisk} = \mathrm{E}_{\hat{\theta}}(S_i)$$

を予測値として用いる事にする。以下ではこのように考えて、分布に依存した S_i のモメントが示される。

問題は区間予測である。Bolfarine and Zacks[19] の 5 章を参考に、S_i の予測集合 $\mathcal{C}(\boldsymbol{s})$ の構成を考える。まず Θ を超母集団モデルの母数空間としよう。定義に素直に従うと、$0 < \alpha < 1$ について

$$\mathrm{P}_\theta(S_i \in \mathcal{C}(\boldsymbol{s})) \geq 1 - \alpha, \quad \forall\theta \in \Theta$$

を満たすような $\mathcal{C}$ が水準 $(1-\alpha)$ の信頼予測集合という事になる。これは $\boldsymbol{s}$ と S_i の同時分布を考えている。また区間予測については、許容 (tolerance) 予測集合という概念も有る。これは $\boldsymbol{s}$ 所与での S_i に関する条件付確率に関する言明と、この条件付確率の分布に関する言明を区別する。具体的には、$\boldsymbol{s}$ 所与で（θ に依存して）$\mathrm{P}_\theta(S_i \in \mathcal{C}_\theta(\boldsymbol{s})|\boldsymbol{s})$ を、$\mathcal{C}_\theta$ の「被度 (coverage)」という。そして被度 $(1-\gamma)$ の $\mathcal{C}_\theta$ について

$$\mathrm{P}_\theta(\mathcal{C}_\theta \subset \mathcal{T}(\boldsymbol{s})) \geq 1 - \alpha, \quad \forall\theta \in \Theta$$

を満たすような $\mathcal{T}$ を、信頼係数 $(1-\alpha)$、被度 $(1-\gamma)$ の許容予測集合という。しかし許容予測集合の概念は、一般的とは言い難い。故に実務をにらんだ場合、信頼予測集合の構成が重要と思われる。

適当な α を所与とする。$\mathrm{P}_{\hat{\theta}}(S_i)$ から、$\alpha_2 - \alpha_1 = 1 - \alpha$ となるように α_1, α_2 分位点 $\hat{z}(\alpha_1), \hat{z}(\alpha_2)$ を求める。なお $\mathrm{P}_{\hat{\theta}}(S_i)$ が解析的に求めがたいとしても中心極限定理が成立するとみなせば、S_i が平均 $\mathrm{E}_{\hat{\theta}}(S_i)$、分散 $\mathrm{V}_{\hat{\theta}}(S_i)$ の正規分布に従う。このようにして予測区間 $\mathcal{C}_{Estimative} = [\hat{z}(\alpha_1), \hat{z}(\alpha_2)]$ を構成できる。信頼水準を $Cl(\theta) = \mathrm{P}_\theta(\hat{z}(\alpha_1) \leq S_i \leq \hat{z}(\alpha_2))$ と書く。$Cl(\theta)$ が $1-\alpha$ に近い保証は無いが、推定的アプローチの枠組みではそのようにみなすしかない。しかし正確に見え過ぎるという指摘の通り、Bjørnstad[17] では信頼水準が大きく $1-\alpha$ を下回る例が示されている。正確な区間予測が必要な場合は、このようなアプローチを用いるべきではない。

母数に関する不確実性を反映するように予測尤度を構成すれば、より正確な区間予測が出来る可能性が有る。Harris[73] による推定的アプローチの拡張を紹介しておく。$\hat{\theta}(\boldsymbol{s})$ の標本分布の密度を $f_\theta(\cdot)$ と書く。この場合 $\hat{\theta}$ の変動を考慮に入れた S_i の分布は

$$\int \mathrm{P}_t(S_i) f_\theta(t) dt$$

で表される。しかしこれは母数 θ に依存するので、標本から得た $\hat{\theta}$ で置き換えて利用する。これはいわゆるパラメトリックブートストラップ法（Efron[45]）である。S_i の「ブートストラップ予測分布」は

$$\int \mathrm{P}_t(S_i) f_{\hat{\theta}}(t) dt$$

のように書ける。もし $f_\theta(\hat{\theta})$ が解析的に評価出来ないとしても、所与の標本 $\boldsymbol{s}$ からブートストラップを用いて評価出来るだろう。また Hall et al.[71] は、推定的方法における被度の評価誤差 $|\mathrm{P}_{\hat{\theta}}(S_i \in \mathcal{C}) - \mathrm{P}_{\theta_0}(S_i \in \mathcal{C})|$ (ただし θ_0 は母数の真値) をブートストラップで修正 (calibration) する事を提案している。なお Hall et al. によれば、精緻な予測尤度概念を用いても被度の評価誤差は、ほとんど改善されない。彼らの主張では、ブートストラップで修正をした推定的方法は精度と利便性の点で現実的な選択肢である。モデルについて信頼予測区間は解析的に評価しづらいが、実際はブートストラップのような数値的方法で十分と考える。これ以上の議論は分布の仮定に依存するので、ここでは扱わない。

なお区間推定に関するここまでの議論は、モデル集合から特定のモデルを選択する過程を考慮していない。つまりモデルが一つ特定された後（またはモデル集合の大きさが一の場合）の考え方を示している。残念ながら本稿では、モデル選択の不確実性を取り入れた区間推定を、十分に議論する余力が無い。ただそのような推定では、重度の複雑化は免れないだろう。事実上は、ブートストラップのような数値的方法に頼らざるを得ないと思われる。この議論は今後の課題としたい。

4章ではモデルに依存して S_i の期待値と分散が示されるが、分散を区間予測に使う事は必ずしも前提とされていない。

2.2 母集団と標本

本節では以下の議論で使用される標本設計、及び母集団分布と標本分布の関係について一般的結果を示しておく。母集団が固定されている場合、「標本分布」は確率抽出の結果の変動である。しかし超母集団モデルアプローチでは、母集団が確率的に生成される。そして母集団が生成されたという条件付きで、標本設計を利用する。以下で「標本分布」という用語は、標本の周辺分布を指す。

まず標本設計について、本節の記法は一章と同じである。すなわち N 個体からなる母集団から n 個の標本を抽出する。q 番目までに抽出された個体のラベルの集合を $\tilde{s}_q$ と書く。q 番目の抽出でラベル λ のついた個体が抽出される確率を $\mathrm{P}(q, \tilde{s}_{q-1}, \lambda)$ と書く。設計一般の議論については、1.2節の冒頭で挙げた教科書などを参照の事。特に生態学分野の標本設計については、別に Seber[157] 等を見よ。本稿では、母集団所与で以下の標本設計のいずれかを用いる。

定義 「非復元単純無作為抽出」は $q = 1, \ldots, n$ について

$$\mathrm{P}(q, \tilde{s}_{q-1}, \lambda) = \begin{cases} 0 & \text{if } \lambda \in \tilde{s}_{q-1}. \\ 1/(N-q+1) & \text{otherwise.} \end{cases}$$

定義 「ベルヌーイ抽出」では、$q = 1, \ldots, N$ について

$$\mathrm{P}(q, \tilde{s}_{q-1}, \lambda) = \begin{cases} 0 & \text{if } q \neq \lambda. \\ n_0/N & \text{otherwise.} \end{cases}$$

但し n_0 は N より小さい正の数である。

社会調査のように名簿が利用可能な場合は、標本が非復元単純無作為抽出されたとみなすのは自然な仮定である。それに対しベルヌーイ抽出では、標本数 n が期待値 n_0 の確率変数になる。統

計的生態学では、単位時間内に光線などのしかけにはまった個体を観測する場合が有る。この時標本数は偶然変動とみなされるので、このような抽出方法を仮定するのが自然である。なおベルヌーイ抽出は、非復元単純無作為抽出の近似として良く用いられている。呼称については Särndal et al.[153] に倣った。

現実には、複数の層から各個体が異なる確率で取られる場合も多い。しかし本稿では、全ての個体が単層から等確率で取られるような設計を前提とする。これは結果の適用可能性を過度に制限するように見えるかもしれない。だが複層の場合でも各層の中では等確率で取られる設計ならば、層毎に単層等確率が前提の分析をして結果を統合すれば同じ事である。

次に母集団分布と標本分布の関係について考察する。確率変数 $X_i, i = 1, \ldots, N$, が、それぞれ壷番号 $(1, 2, \ldots, J)$ を表すと仮定しよう。すなわち実現値 $x_i \in \{1, \ldots, J\}, i = 1 \ldots, N$, である。この時寸法指標を評価するために、まず第 j 壷内のボールの数

$$F_j = \sum_{i=1}^{N} I(x_i = j)$$

から考える。また標本のラベル集合として、$\tilde{s}_n = \{i_1, \ldots, i_n\}$ が得られたとしよう。

$$f_j = \sum_{k=1}^{n} I(x_{i_k} = j)$$

となる。まず母集団が固定されている場合の標本分布を評価しよう。非復元単純無作為抽出の場合、

$$\mathrm{P}(f_1, \ldots, f_J | F_1, \ldots, F_J) = \frac{\binom{F_1}{f_1} \cdots \binom{F_J}{f_J}}{\binom{N}{n}}$$

である。抽出率 n_0/N_0 のベルヌーイ抽出の場合は

$$\mathrm{P}(f_1, f_2, \ldots, f_J | F_1, F_2, \ldots, F_J) = \prod_{j=1}^{J} \binom{F_j}{f_j} \left(\frac{n_0}{N_0}\right)^{f_j} \left(\frac{N_0 - n_0}{N_0}\right)^{F_j - f_j}$$

である。標本抽出について $(f_1, \ldots, f_J)$ は確率変数と考えられているが、記法が煩雑になるので実現値と区別せずに記す。このような省略記法は、本論文を通して用いられる。母集団が確率的だとして、標本の周辺分布は

$$\mathrm{P}(f_1, \ldots, f_J) = \sum_{F_1, \ldots, F_J} \mathrm{P}(f_1, \ldots, f_J | F_1, \ldots, F_J) \mathrm{P}(F_1, \ldots, F_J) \tag{2.5}$$

と書ける。本稿では条件付きでない $\mathrm{P}(f_1, \ldots, f_J)$ を、「標本分布」と呼ぶ。所与の超母集団モデル $\mathrm{P}(F_1, \ldots, F_J)$ から左辺を明示的に得たいが、一般に右辺の評価は簡単とは限らない。利便性の観点から、計算が容易なモデルが望まれる。

もし標本と母集団が同じ分布に従い、サイズに関する母数のみ異なるのであれば便利である。つまりそれぞれ、$\mathrm{P}(f_1, \ldots, f_J; \theta, n)$ と $\mathrm{P}(F_1, \ldots, F_J; \theta, N)$ で表される（θ は母数ベクトルに対応）ならば良い。この時、母集団モデルは標本抽出に関して「共役」と呼ぶ事にする。共役な分布族の例として、「分割構造 (Partition structure: Kingman[103])」を持つ分布族が有る。

分割構造について説明しよう。N 個のボールが確率分布 p_N に従って、各壷に分配されているとする。ここで 1 個のボールを $1/N$ の確率で（一様に）選択して取り出すとする。分割構造が成立

する場合、残りの $N-1$ 個のボールが分布 p_{N-1} に従う。分割構造は再帰的に、$p_1, p_2, p_3, \ldots$ のような確率分布列として定義される。分割構造を寸法指標について言い換えれば、$\sum_{i=1}^{N} it_i = N$ を満たす全ての非負整数の組 $(t_1, t_2, \ldots, t_N)$ について

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_N = t_N) = \sum_{i=1}^{N} \frac{it_i}{N} \mathrm{P}(\cdots, S_{i-1} = t_{i-1} + 1, S_i = t_i - 1, \cdots)$$

が成立する事である。Kingman[104] が指摘するように、分割構造を持つようなモデルならば、p_N に従う母集団から非復元単純無作為抽出で n 個の標本を取った場合、標本分布は p_n となる。言いかえれば無限母集団から直接 n 個標本を取る際の分布と、無限母集団から取られた N 個の母集団から非復元単純無作為抽出で n 個標本を取る際の分布が同じという事である。このモデルは非復元単純無作為抽出に関して共役であり、(2.5) 式の右辺の評価は問題にならない。

Takemura[199] は、共役性の簡単な十分条件を与えている。

命題 2.1 (Takemura[199], Lemma 1) 個体のラベル $(1, 2, 3, \ldots, N)$ の全て ($N!$ 通り) の並び替え $(i_1, \ldots, i_N)$ について母集団分布が不変、すなわち $\mathrm{P}(x_1, \ldots, x_N) = \mathrm{P}(x_{i_1}, \ldots, x_{i_N})$ だとする。このような超母集団モデルは非復元単純無作為抽出について共役である。

命題 2.1 より、個体のラベルが情報を持たない (交換可能な) モデルでは非復元単純無作為抽出の場合、所与の $\mathrm{P}(F_1, \ldots, F_J)$ の N に依存する部分を n におきかえれば、標本分布 $\mathrm{P}(f_1, \ldots, f_J)$ を得られる。そして応用の際は、ラベルが情報を持たないのが普通である。本稿ではこのようなモデルのみ扱う。

例 2.1 頻度 $(F_1, F_2, \ldots, F_J)$ が多項分布

$$\mathrm{P}(F_1, F_2, \ldots, F_J) = \binom{N}{F_1 \cdots F_J} \prod_{j=1}^{J} {\lambda_j}^{F_j}, \tag{2.6}$$

但し $\sum_{i=1}^{J} \lambda_i = 1$、に従うとしよう。この時 N 個の中から n 個を非復元単純無作為抽出する。ここで N 個の個体を並び替えたとしても、母集団分布 (2.6) は不変である。従って個体のラベルは情報を持たない。故に命題 2.1 より、標本分布は

$$\mathrm{P}(f_1, f_2, \ldots, f_J) = \binom{n}{f_1 \cdots f_J} \prod_{i=1}^{J} {\lambda_i}^{f_i}$$

のように書ける。—

次にベルヌーイ抽出の下での標本分布を考える。この時周辺の標本数 f_j は、所与の F_j から確率 n_0/N_0 で抽出する二項分布に従う。(2.5) 式のように考えて、容易に標本分布 $\mathrm{P}(f_1, \ldots, f_J)$ が得られるような $\mathrm{P}(F_1, \ldots, F_J)$ の条件を考えよう。ここで注意すべきなのは、N 所与で標本サイズ n が期待値 $N \cdot n_0/N_0$ の確率変数になる事である。従ってサイズ以外の形状を保存するような関係を求めるなら、母集団サイズが確率変数のモデルが必要になる。なお本稿では母集団サイズは既知の情報である。それを N の期待値と考えて、$\mathrm{E}(N) = N_0$ を所与とする。この時標本サイズの期待値は n_0 である。

では、そのような都合の良い例を挙げよう。確率変数 X が、平均 $N_0\lambda$ のポアソン分布に従うと仮定する。このような X から抽出率 n_0/N_0 でベルヌーイ抽出をして、結果として得られる確率変数を Y とおく。

$$\begin{aligned} \mathrm{P}(Y=y) &= \sum_{x=y}^{\infty} \binom{x}{y} (\frac{n_0}{N_0})^y (1-\frac{n_0}{N_0})^{x-y} \frac{(N_0\lambda)^x \exp(-N_0\lambda)}{x!} \\ &= \sum_{x=y}^{\infty} \frac{(N_0\lambda - n_0\lambda)^{x-y}(n_0\lambda)^y \exp(-N_0\lambda)}{(x-y)!y!} \\ &= \frac{(n_0\lambda)^y \exp(-n_0\lambda)}{y!}. \end{aligned}$$

つまり、Y は平均 $n_0\lambda$ のポアソン分布に従う。なお母関数を用いた証明については例 3.1 を見よ。本稿で検討される頻度 F の分布は混合ポアソン分布であり、更に λ を確率変数と考える（詳しくは 3.1 節を見よ）。このような場合、標本分布は N_0 を n_0 でおきかえて得られる。言いかえれば、混合ポアソン分布はベルヌーイ抽出について共役である。まとめると、以下の命題が成立する。

命題 2.2 確率変数 X が平均 λ のポアソン分布に従い、また λ がある確率分布に従うとする。この時、X の分布を $F(X;N_0,\theta)$ と記述する。ただし θ は、母数（ベクトル）である。このような X から抽出率 n_0/N_0 でベルヌーイ抽出をして得られる Y は、分布 $F(Y;n_0,\theta)$ に従う。

要するにベルヌーイ抽出が自然な状況では、混合ポアソン分布のモデルが便利である。ただ先に述べたようにベルヌーイ抽出と単純無作為抽出は、互いに近似的な関係にある。応用の際に厳密な区別は必要無いかもしれない。

なお母集団サイズが確率変数のモデルで非復元単純無作為抽出を前提とするにはどうしたら良いか。このような場合、条件付母集団分布 $\mathrm{P}(F_1,\ldots,F_J|N)$ から標本分布 $\mathrm{P}(f_1,\ldots,f_J|n)$ を得ると考える。必ずしも $\mathrm{P}(F_1,\ldots,F_J|N)$ が容易に評価出来るとは限らないが、サイズに関する条件付分布さえ得られれば、命題 2.1 が利用できる。後に一部のモデルについて議論されるだろう。

2.3 頻度と寸法指標

ここまで寸法指標ではなく、頻度のモデル $\mathrm{P}(F_1,F_2,\ldots,F_J)$ の標本分布を考察してきた。もし F_j の分布を決めると寸法指標の分布が一意に定まるならば、寸法指標の標本分布を特別に考察する必要はない。以下ではそのような場合を示す。

単純に、各 j について F_j が互いに独立に同一な分布に従うとしよう。この場合全ての j について $\mathrm{P}(F_j=y)=\mathrm{P}(F=y)$ と表せる。同時分布は

$$\mathrm{P}(F_1=y_1,\ldots,F_J=y_J)=\prod_{j=1}^{J}\mathrm{P}(F=y_j)$$

である。この場合組み合わせを考えて、寸法指標の分布

$$\mathrm{P}(S_0,S_1,\ldots)=\binom{J}{S_0S_1\cdots}\prod_{i=0}^{\infty}\mathrm{P}(F=i)^{S_i} \tag{2.7}$$

が得られる。これは無限項の多項分布である。つまり F_j の独立同一分布モデルでは、F_j の分布から寸法指標の分布が一意に定まる。そして寸法指標の標本分布も、f_j の分布から一意に定まる。

次に寸法指標に関する極限定理を紹介しておく。多項分布の任意の周辺分布は多項分布なので、(2.7) において周辺 S_i は二項分布に従う。故に小数法則に対応する極限をとれば、S_i の分布はポアソン分布になるはずである。厳密に考えて、寸法指標 $(S_1, S_2, \ldots)$ の確率母関数は

$$G(z_1, \ldots) = \{\sum_{i=1}^{\infty}(z_i - 1)\mathrm{P}(F = i) + 1\}^J$$

となる。もし適当な非負の実数 c_i について

$$\lim_{J\to\infty} J\mathrm{P}(F = i) = c_i, i = 1, 2, \ldots, \tag{2.8}$$

ならば

$$\lim_{J\to\infty} G(z_1, \ldots) = \exp(\sum_{i=1}^{\infty}(z_i - 1)c_i)$$

である。平均 c_i のポアソン分布の確率母関数が $\exp((z-1)c_i)$ で表される事に注意すると、以下の命題が成立する。

命題 2.3 (Engen[48]) 確率変数 F は適当な非負整数上の分布に従う。$F_j, j = 1, \ldots, J$, が独立かつ同一に F の分布に従い (2.8) が成立するとしよう。この時 $J \to \infty$ とすれば、寸法指標 $S_i, i = 1, 2, \ldots$ の極限分布は各 i 独立に平均 c_i のポアソン分布である。

命題 2.3 の成立する条件 (2.8) の意味は、$\mathrm{E}(S_i)$ が定数に収束するという事である。実は 3.1 節で明らかにされるように、F が非負整数上の無限分解可能分布に従うなら、条件 (2.8) を満たすような操作が可能である。この場合 $J \to \infty$ という極限において各セルで $\mathrm{P}(F_j > 0) \to 0$ だが、命題 2.3 より意味の有るモデルが得られる。

命題 2.3 は項数が無限大の多項分布から導出されたが、有限項数の場合は以下のような結果が知られている。

命題 2.4 (Johnson et al.[89], p.124) 多項分布 (2.6) について

$$\lim_{N\to\infty} N\lambda_i = c_i, \quad i = 1, 2, \ldots, J-1,$$

とする $(\lambda_J \to 1)$。この時 $N \to \infty$ とすれば、$F_i, i = 1, \ldots, J-1$, の極限分布は各 i 独立に平均 c_i のポアソン分布である。

命題 2.4 で得られるような有限個の独立ポアソン分布からなるモデルは、「多重ポアソン (multiple Poisson) 分布」と呼ばれている。多項分布 (2.7) を (2.6) と比較すれば、命題 2.4 が命題 2.3 の有限項数版となっている事が分かる。

第3章　頻度データのモデリング

頻度（計数）データをモデルで記述する場合、非負整数上の離散分布を用いる事になる。3章では、このような分布によるモデリング一般について議論をする。なお個別具体的な分布の下での寸法指標推測については、次章で扱う。3.1節では、重要な離散分布のクラスである混合ポアソン分布と複合ポアソン分布の関係をまとめておく。3.2節では、度数が0の集団をモデリングする方法について述べる。

3.1　混合ポアソン分布と複合ポアソン分布

実は本稿で取り上げる超母集団モデルは全て、混合ポアソン (mixed Poisson) 分布またはその極限と関係付ける事が出来る。従って個別のモデルについて議論をする前に、混合ポアソン分布の一般的性質を確認しておく。なお「混合ポアソン分布」は、後述の「複合 (compound) ポアソン」とは区別される。ただ Greenwood and Yule[65] 等は、"compound Poisson" を混合ポアソンの意味で用いている。また Satterthwaite[154] にならえば、複合ポアソンは "generalized Poisson" と言う。Godambe and Patil[60] では "Poisson stopped sum" である。これらの文脈については、Feller[53] 及び Johnson et al.[90](p.188, p.324) を見よ。

2.1節では超母集団モデルの定義に従い、ボール i が第 x_i 壷に所属するとして x_i は確率変数 X_i の実現値と考えた。離散的確率を一般に考えれば、$\mathrm{P}(X_i = j)$ を i, j に依存するように定めるべきかもしれない。しかし応用の際ラベル i は情報を持たないので、確率 $\mathrm{P}(X_i = j)$ は j のみに依存すると考える。すなわち全ての i について $\mathrm{P}(X_i = j) = \lambda_j$ $(\sum_{j=1}^{J} \lambda_j = 1)$ とする。この場合、母集団サイズ所与で N 個のボールを確率 $\lambda_j, j = 1, \ldots, J,$ で j 番目の壷に分配するような、多項分布になる。つまり母集団分布は、以前例にあげた (2.6) のように書ける。

なお Khmaladze[100] の議論では、多項分布の母数 $(\lambda_1, \lambda_2, \ldots, \lambda_J)$ がボールの総数 N に依存して変化する。例えば n 個の個体が J 個のセルに多項分布

$$\mathrm{P}(f_1, f_2, \ldots, f_J) = \binom{n}{f_1 f_2 \cdots f_J} \lambda_{1n}^{f_1} \lambda_{2n}^{f_2} \cdots \lambda_{Jn}^{f_J}$$

に従って分布すると考える。この場合 λ_{jn} と λ_{jN} の関係を仮定しなければ、母集団に関する推測は出来ない。Khmaladze の問題意識は、$(\lambda_{1n}, \lambda_{2n}, \ldots, \lambda_{Jn})$ の推測である。特に興味深いのは次の結果だろう。

命題 3.1 (Khmaladze[100]) 以下で述べる3条件は同値となる。

- 条件 (c1)

$$\liminf_{n \to \infty} \frac{\mathrm{E}(s_1)}{n} > 0.$$

- 条件 (c2) ある $z < \infty$ について

$$\liminf_{n\to\infty} \sum_{j=1}^{J} \lambda_{jn} I(\lambda_{jn} \leq \frac{z}{n}) > 0.$$

- 条件 (c3)

$$\liminf_{n\to\infty} \mathrm{E}(\sum_{j=1}^{J} |\frac{f_j}{n} - \lambda_{jn}|) > 0.$$

命題3.1の条件 (c3) を解釈すると、λ_{jn} の自然な推定量 f_j/n が必ずしも一致性を持たないという事である。このような場合が "Large Number of Rare Events" と呼ばれる。もし母数 $(\lambda_{1n}, \ldots, \lambda_{Jn})$ が n に依存せず固定されているなら、$n \to \infty$ とすれば確率1で各 $f_j \to \infty$ である。従って $s_1 \to 0$ なので条件 (c1) が満たされる事はない。以下では各 λ_j が標本数に依存しないとして話を進める。

多項分布 (2.6) の下で j 番目の壷に入るボールの数の周辺分布は、母数 N, λ_j の二項分布である。すなわち $j = 1, \ldots, J$ について

$$\mathrm{P}(F_j = y) = \binom{N}{y} \lambda_j^y (1-\lambda_j)^{N-y}, \quad y = 0, 1, \ldots, N,\ 0 < \lambda_j < 1, \tag{3.1}$$

である。故に寸法指標の期待値は (2.3) より、

$$\mathrm{E}(S_i) = \sum_{j=1}^{J} \binom{N}{i} \lambda_j^i (1-\lambda_j)^{N-i}$$

である。特に全ての j について $\lambda_j = 1/J$ の場合、

$$\mathrm{E}(S_i) = J\binom{N}{i} (\frac{1}{J})^i (1-\frac{1}{J})^{N-i}, \quad i = 0, \ldots, N,$$

となる。なお全ての壷に入る確率が等しい場合、寸法指標の同時確率は (4.61) で表され、一般に寸法指標の階乗モメントは (4.62) のように評価出来る。

等確率多項分布の場合、

$$\frac{\mathrm{E}(S_i)}{\mathrm{E}(S_{i-1})} = \frac{N-(i-1)}{(J-1)i} \tag{3.2}$$

より「L字」の程度を考察しよう。$i = 1, 2, \ldots$ と増加するにつれて (3.2) の比は

$$\frac{N}{J-1} > \frac{N-1}{2(J-1)} > \frac{N-2}{3(J-1)} > \cdots$$

と単調に減少してゆく。つまり各壷にボールが入る確率が等しい多項分布の下で、平均的に寸法指標は対数凸でない。対数凸でない寸法指標は、「L字」の記述においてはなだらか過ぎる事が多い。少なくとも、平均的に寸法指標が対数凸となるモデルが必要である。この場合一般に考えて、母数 λ_j の変動を取り扱う必要がある。ここで一つの考え方は、$(\lambda_1, \ldots, \lambda_J)$ が数個の母数で記述できるような分布に従うとする事である。利便性の観点から $(\lambda_1, \ldots, \lambda_J)$ の分布として良く使用されるのが、ディリクレ分布である。この場合に混合分布として得られる $(F_1, F_2, \ldots, F_J)$ の分布はディリクレ＝多項分布 (4.1節) と呼ばれる。

多項分布モデルは異なる F_j 同士で相関を持つ為、取り扱いが必ずしも容易ではない。その為、各 F_j が独立になるようなモデルも考えたい。しかし周辺 F_j の分布を独立とした場合、総和 $\sum_j F_j$ すなわち母集団サイズ N は、確率変数になる。この場合ベルヌーイ抽出の下で共役となる分布が望ましい。すなわち、命題 2.2 より F_j は λ_j 所与でポアソン分布に従うとするのが都合が良い。これは多項分布モデルで F_j が従う二項分布を、ポアソン分布で近似していると解釈出来る。通常 $J >> N$ なので、$F_j > 0$ となるのは稀な事象であり、ポアソン近似は小数法則より正当化される。つまり $j = 1, 2, \ldots, J$ について周辺分布 (3.1) を、独立なポアソン分布

$$\mathrm{P}(F_j = y) = \frac{(N_0\lambda_j)^y \exp(-N_0\lambda_j)}{y!}, \quad y = 0, 1, \ldots, \lambda_j > 0, \tag{3.3}$$

で近似するのが自然なモデルという事になる。なおここでも λ_j を確率変数と考えて、混合ポアソン分布を考える事が多い。Good[61] や Sichel[174] は、同様な議論を用いて混合ポアソン分布の妥当性を主張している。また混合ポアソン分布は、分割表の解析に使われるポアソン対数線形モデルと接続（Paul and Plackett[132] を見よ）出来る点でも重要である。特に個票開示リスク評価分野では、この切り口は検討を要する。

加えて同質的集団における観測度数が、経験的にポアソン分布に近い事も重要である。Neyman[124] は、同質的集団の度数をポアソン分布か二項分布（これらの分布では分散が平均を上回らない）で記述する事は、検証される事を前提に「非合理的ではない」と述べている。そしてポアソン分布のあてはまりが不十分な集団を、非同質的であるとみなした。例えば Willmot[223] が指摘するように、混合ポアソン分布の分散は平均より大きくなる (over dispersion)。従って分散が平均より大きい集団を非同質的として扱う事と、整合的である。

ここでのアプローチは、λ_j の非同質性を別の分布で記述するという事である。つまり混合ポアソン分布は、集団の集計的挙動に関する「十分柔軟な内挿式」の候補として考えられると、Neyman は主張する。なお Neyman は「汚染 (contagion)」という用語を、非同質的構造に用いている。ただ Feller[53] によれば、「汚染」という用語は (a) 事象間の相関の形容と (b) 非同質性の形容で用いられる事が有る。そして混合ポアソン分布は、どちらのモデルとしても使われた実績がある。すなわち構造と表面的分布の対応が一意ではない。故に分布の表面的あてはまりから「汚染」構造の有無は判断できないと、Feller[53] は述べている。なお一章で考察したように、我々のアプローチは根拠を突き詰めると経験的あてはまりに依存する。従って以上の議論は、構造を問わない本稿の方法論と矛盾しない。

なお頻度（計数）データを説明する有力な手法として、ポアソン回帰がある（例えば興味深い応用例を含む Cameron and Trivedi[27] を参照の事）。ポアソン回帰では、ある度数 (F_j) が従うポアソン分布の平均母数 λ_j を適当な変量で説明する。つまり j に依存した構造の違いをモデルに導入している事になる。これに対し本稿の方法論では、度数のレベル（λ_j）の散らばりを、同一分布からの実現値の散らばりとみなしている。各セルの属性の違いを同一分布からの実現値の散らばりと見なすなら、λ_j は j に依存しない。だとすれば、セルを同質的 (exchangeable) に扱う本稿の方法論で良い。逆にセルの属性所与で条件付きの推測をするなら、ポアソン回帰の考え方が適切である。本稿ではこれ以上考察しないが、セルの属性を補助情報として用いれば、（複雑になるとはいえ）寸法指標の推測を改善できる可能性が有ろう。

混合ポアソン分布の一般的性質をいくつか述べておく。ここで触れられない性質については、Johnson et al.[90] の 3 章または Haight[70]、Willmot[223] 等を見よ。確率変数 F が、平均 λ のポ

アソン分布に従うとしよう。ここで混合する分布、すなわち λ の密度関数を $f(\lambda)$ とする。つまり我々は分布

$$\mathrm{P}(F=y)=\frac{1}{y!}\int_0^\infty \exp(-\lambda)\lambda^y f(\lambda)d\lambda, \quad y=0.1,\ldots, \tag{3.4}$$

を考察する。λ の分布をうまく選べば、「L 字型」の記述が可能となる。この時 F の確率母関数は

$$G(z)=\int \exp(\lambda(z-1))f(\lambda)d\lambda \tag{3.5}$$

のようになる（Gurland[68]）。なお混合した分布のラプラース変換は

$$\mathcal{L}(s)=\int \exp(-\lambda s)f(\lambda)d\lambda = G(1-s)$$

である事が分かる。

Paul and Plackett[132] は一般の混合ポアソン分布について、近似を考えている。すなわち $\mathrm{E}(\lambda)=\mu$ として、$\mathrm{V}(\lambda)=\sigma^2$ が小さく高次のモメントが無視できる場合、

$$G(z)=\exp(\mu(z-1))\{1+\frac{1}{2}\sigma^2(z-1)^2\}$$

なので、$1/y!\cdot d^yG(z)/dz^y|_{z=0}$ を評価して

$$\mathrm{P}(F=y)\approx \mu^y\exp(-\mu)/y!+\frac{1}{2}\sigma^2\nabla^2(\mu^y\exp(-\mu)/y!), \quad y=0,1,\ldots, \tag{3.6}$$

ただし ∇ は y に関する後方差分演算子（例えば $\nabla y=y-(y-1)$）である。

混合ポアソン分布の利点として、ベルヌーイ抽出の下で標本分布が簡単に求められる事（命題 2.2）の他、モメントの導出が容易な事があげられる。Ottestad[130] によれば、混合ポアソン分布の r 次階乗モメントは、混合された分布の原点周りの r 次モメントになる。更に踏み込んだ結果を以下に示す。

命題 3.2 (Maceda[113]) 分布 (3.4) についてモメントが存在するとして、

$$\mathrm{E}(F^{(r)})=\mathrm{E}(\lambda^r), \tag{3.7}$$

ただし $F^{(r)}=F(F-1)\cdots(F-r+1)$ である。また逆に全ての次数 r について (3.7) が成立するならば、(3.4) でなければならない。

命題 3.2 を言いかえれば、階乗積率母関数が混合される分布の積率母関数と等しい事が混合ポアソン分布の必要十分条件という事である。(3.7) については $\mathrm{E}(F^{(r)}|\lambda)=\lambda^r$ から両辺の期待値をとれば、簡単に確認できる。

また混合ポアソン分布は、混合される分布が単峰ならば単峰になる。絶対連続な確率変数の密度関数を $f(x)$ としよう。もし $f(x)$ が x_0 の右では単調減少、左では単調増加するような x_0 が存在するなら、この確率変数は単峰な密度を持つと言う。また離散の場合を考えよう。まず $\boldsymbol{Z}$ を整数の集合とする。集合 $\{\theta_0+k\theta|k\in\boldsymbol{Z}\}$ の上で、k に依存して確率 p_k が決まる確率変数を考察する。もし k_0 よりも k が小さければ p_k が単調に減少し、大きければ単調に増加するような k_0 が存在するとしよう。このような確率変数を単峰格子変数という。

命題 3.3 (Holgate[78]) 分布 (3.4) について、密度 $f(\lambda)$ が絶対連続かつ単峰、$\lambda \geq 0$ だとする。この時確率変数 F は、単峰格子変数になる。

命題 3.3 より、L 字型の連続分布にポアソン分布を混合した場合、L 字という形状は結果として得られる離散分布でも保たれるように見える。

次に複合ポアソン分布について、Feller[54] の 268 ページ以下の議論に従って話を進める。$\{X_l\}_{l=1}^{N}$ が互いに独立に同一な正の整数上の分布に従う確率変数列とする。

$$S_N = X_1 + \ldots + X_N \tag{3.8}$$

と書く。もし N が各 X_l とは独立にポアソン分布に従うなら、S_N の分布を「(X_l をクラスター分布とする) 複合ポアソン分布」と呼ぶ。仮に N が平均 μ のポアソン分布に従うとしよう。この時 $X_l, l = 1, 2, \ldots,$ の確率母関数を

$$g(z) = \sum_{i=1}^{\infty} p_i z^i,$$

ただし $p_i = \mathrm{P}(X_l = i)$ と書く。条件付期待値および確率母関数の畳み込みに関する性質を利用して、複合ポアソン分布の確率母関数は、

$$G(z) = \sum_{n=0}^{\infty} \mathrm{P}(N = n) z^n g(z)^n = \exp(\mu(g(z) - 1)) \tag{3.9}$$

で表される。同じ事だが、このような確率母関数を持つ非負整数上の分布を複合ポアソン分布と呼ぶ。

Steutel[193] が、複合ポアソン分布の必要条件と十分条件をまとめている。その特徴を一言で言えば、裾が長い分布という事になる。特に Warde and Katti[212] によれば、非負整数上の分布に従う確率変数 F が $\mathrm{P}(F = 0) \neq 0, \mathrm{P}(F = 1) \neq 0$ を満たすとして、もし $\{\mathrm{P}(F = i)/\mathrm{P}(F = i - 1)\}_{i=1}^{\infty}$ が単調非減少系列ならば、F は複合ポアソン分布に従う。故に以下の命題が成立する。

命題 3.4 $F_j, j = 1, 2, \ldots, J,$ が互いに独立に同一の非負整数上の分布に従うとする。また $\mathrm{P}(F_j = 0) \neq 0. \mathrm{P}(F_j = 1) \neq 0$ とする。この時、もし平均的に寸法指標が対数凸 (1.12)、すなわち $\{\mathrm{E}(S_i)/\mathrm{E}(S_{i-1})\}_{i=1}^{\infty}$ が単調非減少ならば、F_j の分布は複合ポアソン分布である。

命題 3.4 を示すには

$$\frac{\mathrm{E}(S_i)}{\mathrm{E}(S_{i-1})} = \frac{J\mathrm{P}(F_j = i)}{J\mathrm{P}(F_j = i - 1)}$$

に注意して、Warde and Katti の十分条件を用いれば良い。本命題によれば、L 字を記述出来るモデルの多くが複合ポアソン分布モデルという事になる。

なお複合ポアソン分布は一般化 (generalized) ポアソン分布とも言われる。ポアソン分布の確率母関数を $H(z) = \exp(\lambda(z - 1))$ で表そう。この時 $g(z)$ で定義される分布をクラスター分布とする複合ポアソン分布の確率母関数は、$H(g(z))$ である。この場合 $g(z)$ は、一般化する (generalizing) 分布と呼ばれる。なお例 3.1.3.2 に見られるように、適当な制約を入れなければ複合ポアソン分布は一意に定まらない。以下では $g(z)$ を正の整数上の分布としよう。

例 3.1 ベルヌーイ抽出 (Feller[54], p.268)

確率変数 N が平均 λ のポアソン分布に従うとする。また N と独立に確率変数 $X_l, l = 1, 2, \ldots, N,$ が、互いに独立に同一な（非負整数上の）二項分布

$$X_l = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{otherwise,} \end{cases}$$

に従うとする。X_l の分布の確率母関数は

$$g(z) = (1 - f) + fz$$

であり、この場合 S_N (3.8) の確率母関数は

$$G(z) = \exp(f\lambda(z - 1))$$

である。すなわち、平均 λ のポアソン分布に従う N に確率 f のベルヌーイ抽出を適用すれば、平均 $f\lambda$ のポアソン分布になる事が示された（命題 2.2 を参照せよ）。

正の整数上の分布にする為、X_l が従う二項分布から 0 を切り落とそう。この場合、確率 1 で $X_l = 1$ となる。もし $X_l, l = 1, 2, \ldots, N,$ が互いに独立に確率 1 で 1 となるとしたら、N が独立な平均 λ' のポアソン分布に従うとして、総和 S_N の確率母関数は

$$H(z) = \exp(\lambda'(z - 1))$$

となる。もし $f\lambda = \lambda'$ なら、$G(z) = H(z)$ である。—

例 3.2 ポアソン・パスカル分布 (Willmot[225])

非負整数上の分布である負の二項分布（パスカル分布）の確率母関数は、

$$h(z) = (1 + \beta(1 - z))^{-\alpha}$$

で表される。これを一般化する分布とする一般化ポアソン分布の確率母関数は

$$H(z) = \exp(\lambda((1 + \beta(1 - z))^{-\alpha} - 1))$$

である。また負の二項分布のゼロ切り落とし分布の確率母関数は

$$g(z) = \frac{(1 + \beta(1 - z))^{-\alpha} - (1 + \beta)^{-\alpha}}{1 - (1 + \beta)^{\alpha}}$$

である。これを一般化する分布とする一般化ポアソン分布の確率母関数を

$$G(z) = \exp(\lambda'(\frac{(1 + \beta(1 - z))^{-\alpha} - (1 + \beta)^{-\alpha}}{1 - (1 + \beta)^{\alpha}} - 1))$$

で表す。ここでもし

$$\lambda' = \lambda(1 - (1 + \beta)^{-\alpha})$$

ならば、$G(z) = H(z)$ である。—

複合ポアソン分布の確率母関数 (3.9) を z のべき乗で展開すると、次のように書ける。

$$\begin{aligned}
G(z) &= \sum_{k=0}^{\infty} \frac{(\mu(\sum_{i=1}^{\infty} p_i z^i - 1))^k}{k!} \\
&= \sum_{k=0}^{\infty} \frac{\mu^k}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \sum_{l=1}^{\infty} z^l \sum_{i_1+i_2+\cdots+i_j=l} \prod_{h=1}^{j} p_{i_h} \\
&= \exp(-\mu) + \sum_{l=1}^{\infty} z^l \sum_{k=1}^{\infty} \sum_{j=1}^{k} (-1)^{k-j} \frac{\mu^k}{k!} \frac{k!}{j!(k-j)!} \sum_{i_1+i_2+\cdots+i_j=l} \prod_{h=1}^{j} p_{i_h} \\
&= \exp(-\mu) + \sum_{l=1}^{\infty} z^l \sum_{k=1}^{\infty} \sum_{j=1}^{k} (-1)^{k-j} \mu^k \frac{1}{(k-j)!} \sum_{(t_1.t_2.\ldots.t_l)\in\mathcal{T}_{l.j}} \prod_{i=1}^{l} \frac{(p_i)^{t_i}}{t_i!} \\
&= \exp(-\mu) + \sum_{l=1}^{\infty} z^l \exp(-\mu) \sum_{j=1}^{l} \sum_{(t_1.t_2.\ldots.t_l)\in\mathcal{T}_{l.j}} \mu^j \prod_{i=1}^{l} \frac{(p_i)^{t_i}}{t_i!},
\end{aligned}$$

ただし $i_h.h = 1,2,\ldots,$ は正の整数であり、$t_i.i = 1,2,\ldots,$ は非負整数である。また

$$\mathcal{T}_{l.j} = \{(t_1,t_2,\ldots,t_l) | \sum_{i=1}^{l} it_i = l. \sum_{i=1}^{l} t_i = j\}.$$

従って確率変数 F が複合ポアソン分布 (3.9) に従う場合、確率関数は

$$\mathrm{P}(F=y) = \exp(-\mu) \sum_{j=0}^{y} \sum_{(t_1.t_2.\ldots.t_y)\in\mathcal{T}_{y.j}} \mu^j \prod_{i=1}^{y} \frac{(p_i)^{t_i}}{t_i!}$$

となる。この結果は Johnson et al.[90] の (9.44) 式と同等である。また確率母関数を繰り返し微分する事で、確率関数は漸化式で表される。同じく Johnson et al.[90] (p.352) を見よ。

ところで複合ポアソン分布の確率母関数は全ての自然数 n について、

$$G(z) = \exp(\frac{\mu}{n}(g(z)-1))^n = G_n(z)^n.$$

ただし $G_n(z)$ もまた複合ポアソン分布の確率母関数、となる事が分かる。一般に全ての自然数 n について、ある分布の特性関数が別の分布の特性関数の n 乗で表されるならば、「無限分解可能」と言う（例えば Cramér[39] を見よ）。これはつまり n 回の畳み込みで表されるという事である。複合ポアソン分布は無限分解可能であり、更に Lévy の定理によれば、非負整数上の無限分解可能分布は、複合ポアソン分布に限られる（Feller[54] の 12.3 節、及び Johnson et al.[90](p.323) を見よ)。これを言い換えたのが以下の命題である。

命題 3.5 (Lévy の定理) 非負整数上の分布が無限分解可能である事の必要十分条件は、その確率母関数 $G(z)$ が (3.9) のように書ける事である。

例えば $N_l.l = 1.2.3.$ が従うポアソン分布の平均を t_l としよう。この時 $t_3 = t_1 + t_2$ ならば、

$$S_{N_3} = S_{N_1} + S_{N_2} \tag{3.10}$$

が成立する。またこのような性質が成立するのは、複合ポアソン分布だけである。Herdan[76] の経験では、任意の文法上の区分の中（壷の部分集合の上）で単語の頻度分布は、それらを集計した場合（全壷の上）の分布と性質が大きく異ならない。Herdan はこの事から、語数のモデルは複合ポアソン分布でなければならないと主張している。

興味深いのは、同時に混合ポアソン分布かつ複合ポアソン分布の場合が有るという事である。例えば Maceda[113] は、混合ポアソンかつ複合ポアソンとなる例が無限に作れる事を指摘した。複合ポアソン (3.8) について、非負確率変数 $X_i, i = 1, 2, \ldots,$ の階乗積率母関数を $\phi(z)$ とする。N が平均 μ のポアソン分布に従う時、S_N の階乗積率母関数は

$$FM(z) = \exp(\mu(\phi(z) - 1))$$

と書ける（先ほどの確率母関数の導出と同様）。ここで適当な確率変数について、積率母関数を $\omega(z)$ としよう。

$$M(z) = \exp(\mu(\omega(z) - 1))$$

も、この時積率母関数になる。命題 3.2 より、S_N の階乗積率母関数 $FM(z)$ が積率母関数 $M(z)$ に等しければ、S_N は $M(z)$ によって定義される分布で混合されたポアソン分布である。そして今度は $\phi(z)$ を確率母関数とみなせば、$FM(z)$ は確率母関数となる。つまり $FM(z)$ により、複合ポアソン分布が定義される。このように考えれば S_N の分布が、混合ポアソンかつ複合ポアソンとなっている事が分かる（以下の例 3.3 を参照の事）。その他の例については Willmot[223] を見よ。

例 3.3 負の二項分布（Willmot[223]）

形状母数 α、尺度母数 β を持つガンマ分布の積率母関数は

$$M_{gamma}(z) = (1 - \beta z)^{-\alpha}$$

で表される。確率母関数 G と階乗積率母関数 FM の間には $FM(z) = G(1 + z)$ という関係が有る (Johnson et al.[90], p.49) ので、$M_{gamma}(z)$ を階乗積率母関数とする分布の確率母関数は

$$G_{nb}(z) = (1 + \beta(1 - z))^{-\alpha}$$

で与えられる事になる。これは負の二項分布（ガンマ＝ポアソン混合分布）の確率母関数である。

この確率母関数を書き換えて、複合ポアソン表現を得る。すなわち

$$G_{nb}(z) = \exp(\mu(g(z) - 1)),$$

ただし

$$g(z) = \frac{\log(1 - \frac{\beta}{1+\beta} z)}{\log(1 - \frac{\beta}{1+\beta})}, \quad \mu = \alpha \log(1 + \beta), \tag{3.11}$$

である。この場合 $g(z)$ は対数級数分布

$$\mathrm{P}(F = y) = \frac{c(\beta/(1 + \beta))^y}{y}, \quad y = 1, 2, \ldots.$$

ただし $c = -(\log(1/(1+\beta)))^{-1}$、を定義する確率母関数になっている。すなわち $G_{nb}(z)$ という確率母関数は、混合ポアソン分布 (3.4) において $f(\lambda)$ がガンマ分布の場合、および対数級数分布をクラスター分布とする複合ポアソン分布を規定している。

逆に $g(z)$ を階乗積率母関数とみなせば、その分布の確率母関数は

$$h(z) = \frac{\log(1 - \frac{\beta}{1+\beta}(1+z))}{\log(1 - \frac{\beta}{1+\beta})}$$

で表される。ここで

$$G(z) = \exp(\mu(h(z) - 1))$$

が、ガンマ分布の積率母関数 $M_{gamma}(z)$ と等しい事を確認出来る。—

実は Maceda[113] によれば、混合ポアソン分布 (3.4) が複合ポアソン、すなわち非負整数上で無限分解可能な必要十分条件は、混合される分布が無限分解可能な事である。例えばガンマ分布は無限分解可能であり、ガンマ＝ポアソン分布も無限分解可能である。

任意の複合ポアソン分布は、各 S_i が独立なポアソン分布に従うとして、適当な $\sum_{i=1}^{\infty} iS_i$ の分布で表す事が出来る。X の確率母関数が (3.9) で表されるとしよう。この時

$$\begin{aligned} G(z) &= \exp(\mu(\sum_{i=1}^{\infty} p_i z^i - 1)) \\ &= \exp(\mu p_1(z-1)) \times \exp(\mu p_2(z^2-1)) \times \exp(\mu p_3(z^3-1)) \cdots \end{aligned}$$

と書ける (Johnson et al.[90], p.323)。平均 μp_i のポアソン分布に従う確率変数を S_i で表す。ここで $\exp(\mu p_i(z^i - 1))$ は、$i \times S_i$ の確率母関数である。すなわち、寸法指標 S_i が平均 μp_i のポアソン分布に従っている時に発生する個体数 $i \times S_i$ の分布と解釈出来る。そして $G(z)$ は、各 i 独立として $i \times S_i$ の総和の分布を表す。このような $S_i, i = 1, 2, \ldots,$ がそれぞれ独立なポアソン分布に従うモデルは、「畳み込みポアソン分布 (composed Poisson distributions)」と呼ばれる。

複合ポアソン分布が畳み込みポアソン分布で表されるという性質は、超母集団モデル構築で大きな意味を持つ。$F_j, j = 1, 2, \ldots, J.$ が任意の非負整数上の同一無限分解可能分布に各 j 独立に従うとしよう。F_j の分布の積率母関数を (3.9) で表す。この場合母集団サイズ N の積率母関数は

$$\exp(J\mu(g(z) - 1))$$

となる。すなわち独立な F_j の総和である母集団サイズ N は、やはり無限分解可能分布に従う。従って無限分解可能な F_j で構成された母集団モデルにおいて N の分布は、独立にポアソン分布に従う S_i と i の積の総和という解釈が成立する。

例 3.4 負の二項分布（例 3.3 より続く）

例 3.3 において、$g(z)$ が規定する分布は対数級数分布であった。従って負の二項分布の確率母関数は以下のように書ける。

$$G(z) = \exp(\alpha \log(1+\beta) \frac{\alpha(\beta/(1+\beta))}{1}(z-1))$$

$$\times \exp(\alpha \log(1+\beta)\frac{\alpha(\beta/(1+\beta))^2}{2}(z^2-1))$$
$$\times \exp(\alpha \log(1+\beta)\frac{\alpha(\beta/(1+\beta))^3}{3}(z^3-1))$$
$$\times \cdots$$
$$= \exp(\alpha \log(1+\beta)(g(z)-1)).$$

ここで $F_j, j = 1, \ldots, J,$ が負の二項分布に従うような母集団モデルを考えよう。母集団サイズ N の挙動を求めるために $G(z)$ を J 回畳み込むと、平均母数 $\mu = \alpha \log(1+\beta)$ を J 倍する事になる (負の二項分布が畳み込みに関して閉じているという事である)。この場合 N の分布はやはり無限分解可能となり、各 S_i が独立なポアソン分布に従う場合の $\sum iS_i$ の分布という解釈が出来る。—

では F_j の分布が無限分解可能でない場合に、N の分布は無限分解可能だろうか。もし N の分布が非負整数上の無限分解可能分布であるならば、前述の Lévy の定理よりその確率母関数は (3.9) のように書ける。この時 $F_j, j = 1.2, \ldots, J,$ が互いに独立同一分布に従っているならば、その分布の確率母関数は

$$\exp(\frac{\mu}{J}(g(z)-1)) \tag{3.12}$$

となり、無限分解可能でなければならない。すなわち「N の分布が無限分解可能ならば、F_j の分布は無限分解可能である。」この命題の対偶をとれば、「F_j の分布が無限分解可能でないならば、N の分布は無限分解可能でない。」という事になる。

しかし F_j の分布に無限分解可能性を仮定していない命題 2.3 の議論でも、仮定 (2.8) を満たせばたたみこみポアソン分布が得られるという事であった。もし $S_i, i = 1, 2, \ldots,$ が互いに独立に平均 c_i のポアソン分布に従うなら、N の分布は確率母関数

$$G_c(z) = \prod_{i=1}^{\infty} \exp(c_i(z^i - 1)) \tag{3.13}$$

で表される。そしてもし $\sum_{i=1}^{\infty} c_i = C < \infty$ ならば、

$$G_c(z) = \exp(C(g_c(z) - 1)),$$

ただし

$$g_c(z) = \sum_{i=1}^{\infty} \frac{c_i}{C} z^i$$

と書き直すことが出来る。ここで $\sum_{i=1}^{\infty} c_i/C = 1, c_i \geq 0$ なので、$g_c(z)$ は確率母関数と考えられる。この場合、N の分布は無限分解可能である。これはあくまでも極限についての話なので、前段落で述べた事と矛盾はしない。ただし命題 2.3 の仮定 (2.8) と無限分解可能性の関係を考察しておくべきだろう。

任意の J について

$$\mathrm{P}(F_j = i) = \frac{c_i}{J}, \quad j = 1.2, \ldots, J. \; i = 1.2, \ldots. \tag{3.14}$$

が成立するとしよう。ただし各 $c_i, i = 1.2, \ldots.$ は非負の定数である。この場合、

$$\mathrm{E}(N) = J\mathrm{E}(F_j) = J\sum_{i=0}^{\infty} i\frac{c_i}{J} = \sum_{i=0}^{\infty} ic_i$$

である。すなわち (3.14) が成立する場合、N の期待値は J に依存しない。故に命題 2.3 の仮定 (2.8) が成立すれば、$J \to \infty$ という極限において $\mathrm{E}(N)$ は J に依存しない。なお極限において、N の分布の確率母関数は $G_c(z)$ である。これを微分して $\mathrm{E}(N)$ を求めても、同じ結果が得られる (Satterthwaite[154] を見よ)。

逆に $\mathrm{E}(N)$ を定数として $J \to \infty$ とした場合、何が起こるのだろうか。このような極限操作を、以下では「小数法則」と呼ぼう (これは二項分布からポアソン分布を導出するのと同等な極限操作である)。$J \to \infty$ という操作は、J が非常に大きい場合の近似と解釈出来る。個票開示の分野ではナイーブに決まる J が非常に大きい (例えば佐井・竹村 [149] では $J = 5.644 \times 10^{12}$) ので、「小数法則」の適用は妥当である。また後述されるが、モデルから J への依存を除去する事には重要な意義が有る。

条件 (2.8) を満たすモデルについて $J \to \infty$ ならば $\mathrm{E}(N)$ は定数になるので、そのようなモデルについて小数法則を適用した場合にはたたみこみポアソン分布が得られる。そもそも $\mathrm{E}(N)$ が定数ならば、

$$\mathrm{E}(N) = N_0 = J\mathrm{E}(F_J) = J\sum_{i=0}^{\infty} i\mathrm{P}(F_j = i)$$

であり、右辺は J に依存してはならない。すなわち

$$\sum_{i=0}^{\infty} i\mathrm{P}(F_j = i) = \mathrm{E}(F_j) = \frac{N_0}{J}$$

でなければならない。だがこのような条件を満たす場合は、明らかに無数にある。従って意味のある結果を得るには、より限定的に考える必要があろう。

$\mathrm{E}(N)$ が J の変化について定数となる場合として、N の分布が J に依存せず変化しない場合が考えられる。これはそもそも N の分布が無限分解可能という事に他ならない。例えば N の確率母関数が (3.9) ならば、F_j の分布の確率母関数は (3.12) であり、無限分解可能となる。実は Lévy の定理より、J に依存しないような N の分布は複合ポアソン分布に限られる。すなわち、以下の命題 3.6 が成立する。

命題 3.6 (Hoshino[82], Remark 1) $F_j, j = 1, 2, \ldots, J$, が互いに独立に同一の非負整数上の分布に従うとする。もし $N = \sum_{j=1}^{J} F_j$ の分布が全ての J について変化しないとしたら、F_j および N の分布は複合ポアソンに限る。

従って、F_j が非負整数上の無限分解可能分布に従うようなモデルの挙動を明らかにする事は重要である。今度はそのようなモデルについて、小数法則を適用しよう。この時たたみこみポアソン分布が得られる (正確には、以下の命題 3.7 が成り立つ)。ここでは $J\mu$ を固定することが母集団分布を固定すること、すなわち $\mathrm{E}(N)$ を定数とする事と等価である。

命題 3.7 (Hoshino[82], Theorem 1) $F_j, j = 1, 2, \ldots, J$, が互いに独立に同一の非負整数上の分布に従い、その確率母関数が (3.9) だとする。ここで $J\mu = A$ として固定する。$J \to \infty$ $(\mu \to 0)$ の極限で、$S_i, i = 1, 2, \ldots$ はそれぞれ独立に平均 Ap_i のポアソン分布に従う。

混乱しがちなので、注意しておこう。式 (2.7) で示されるように、J が有限の時に寸法指標の同時確率は、独立なポアソン分布の積ではない。確率変数 X と Y の分布が等しいことを $X \stackrel{d}{=} Y$ と

書く。そして確率変数 T_i が、平均 Ap_i のポアソン分布に従うとする。命題 3.7 の状況で、全ての J について

$$N = \sum_{j=1}^{J} F_j = \sum_{i=1}^{\infty} iS_i \stackrel{d}{=} \sum_{i=1}^{\infty} iT_i$$

が成立する。しかし $J < \infty$ について

$$S_i \stackrel{d}{\neq} T_i, \quad i = 1, 2, \ldots,$$

という事である。ただし $\stackrel{d}{\neq}$ は分布が異なる事を表す。

例 3.5 負の二項分布（例 3.3 より続く）

$F_j, j = 1, 2, \ldots, J,$ が確率母関数 $G_{nb}(z)$ で表される負の二項分布に従うとする。$g(z)$ は対数級数分布の確率母関数で、$\theta = \beta/(1+\beta)$ とおけば

$$p_i = -\frac{1}{\log(1-\theta)} \frac{\theta^i}{i}, \quad i = 1, 2, \ldots,$$

となる。ここで $J\mu = A$ を固定して、$J \to \infty$ とする。極限で $S_i, i = 1, 2, \ldots,$ は互いに独立に平均

$$Ap_i = \frac{A}{-\log(1-\theta)} \frac{\theta^i}{i}$$

のポアソン分布に従う（後述の対数級数モデル）。—

命題 3.7 の重要性は、（適当な正則条件を満たす）任意の L 字型曲線を確率モデルとして正当化できる事にある。つまり本論文の方法論においてモデル構築は、パラメトリックな曲線の構想に問題を縮小できる。$i = 1, 2, \ldots$ について、S_i が c_i に比例するような形状を「L 字」として考えているとしよう。曲線を連続関数 $f(\cdot)$ として考えているならば、$c_i = f(i)$ とすれば良い。ただし

$$\sum_{i=1}^{\infty} ic_i < \infty, \quad c_i \geq 0, \quad i = 1, 2, \ldots,$$

でないと母集団モデルとして意味がない。この時

$$\sum_{i=1}^{\infty} c_i = C < \sum_{i=1}^{\infty} ic_i < \infty$$

であり、(3.13) 以下の複合ポアソン分布に関する議論を用いることが出来る。$F_1, F_2, \ldots, F_J$ が互いに独立に同一分布 (3.13) に従う場合、$JC = A$ を固定して $J \to \infty$ としよう。命題 3.7 より、極限で各 S_i は互いに独立に平均 Ac_i/C のポアソン分布に従う。つまり $\mathrm{E}(S_i) \propto c_i$ となって、望ましい形状が得られたことになる。しかも各 S_i が独立なポアソン分布に従うという簡潔な構造は、応用の際の数値評価を容易にする。

例 3.6 Zipf モデル (渋谷 [169])

Zipf 分布の確率関数は

$$\mathrm{P}(X=x)=\frac{1}{x(x+1)},\quad x=1,2.\ldots,$$

のように書ける。この分布の確率母関数は

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

となる。$X_1,X_2,\ldots$ が互いに独立に Zipf 分布に従うとする。この時 (3.8) の分布、すなわち Zipf 分布をクラスター分布とする複合ポアソン分布は正の μ について

$$G(z;\mu)=\exp(\mu(g(z)-1))=(1-z)^{-\mu(1-1/z)}$$

という確率母関数で表される。$F_j, j=1.2,\ldots,J,$ が互いに独立に同一なこの分布に従うとしよう。小数法則の結果として、

$$\mathrm{E}(S_i)\propto\frac{1}{i(i+1)},\quad i=1,2,\ldots,$$

を満たす極限分布を得る。—

ではどのような「L 字型」の $\{c_i\}_{i=1}^{\infty}$ を用いるべきだろうか。そのような系列は、適当に基準化すれば正の整数上の分布になる。従って、既存の離散分布に関する研究成果を援用すれば良い。正の整数上の分布のクラスとして、例えばベキ級数 (Power series) 分布が考えられる。確率関数が正のベキ母数 θ について

$$\mathrm{P}(X=x)=\frac{a_x\theta^x}{\eta(\theta)},\quad x=0,1,\ldots,\tag{3.15}$$

ただし $a_x\geq 0$ で

$$\eta(\theta)=\sum_{x=0}^{\infty}a_x\theta^x$$

のように書ける場合にベキ級数分布という。詳しくは Johnson et al.[90] の 2.2 節を参照せよ。任意のゼロ切り落としベキ級数分布は、$\mathrm{P}(X=0)=0$ を満たすようなベキ級数分布である。従って、正の整数上の分布としてもベキ級数分布を用いることが出来る。なおベキ級数分布 (3.15) の確率母関数は、

$$G(z)=\frac{\eta(\theta z)}{\eta(\theta)}$$

のように書ける。例えば対数級数分布の確率母関数が (3.11) で与えられている。ここで

$$\eta(\theta)=\log(1-\theta),\quad \theta=\frac{\beta}{1+\beta},$$

と見れば、対数級数分布がベキ級数分布のクラスに属することを確認できる。

ベキ級数分布のクラスは、多くの基本的離散分布を含む。従ってベキ級数分布のクラスの中で、望ましい分布の選択を議論する事は重要である。もちろんこれは一章で考察したように、データに依存する。故にデータが特定の分布が要求する構造に近いか否かを、判断する事になる。ここでもっとも簡潔かつ有効な方法として、データの図示について考察しよう。

	傾き a	切片 b	母数との関係
二項分布： $\eta(\theta)=(1+\theta)^n$ $\theta>0, n=2,3,\ldots$	$-\theta$ (-)	$(n+1)\theta$ (+)	$\theta=-a$ n=-(b/a+1)
ポアソン： $\eta(\theta)=\exp(\theta)$ $\theta>0$	0	θ (+)	$\theta=b$
負の二項分布： $\eta(\theta)=(1-\theta)^{-k}$ $0<\theta<1, k>0$	θ (+)	$(k-1)\theta$ (+/-)	$\theta=a$ $k=b/a+1$
対数級数分布： $\eta(\theta)=-\log(1-\theta)$ $0<\theta<1$	θ (+)	$-\theta$ (-)	$\theta=a,$ $\theta=-b$
拡張負の二項分布： $\eta(\theta)=1-(1-\theta)^r$ $0<\theta<1.0<r<1$	θ (+)	$-(r+1)\theta$ (-)	$\theta=a$ $r=-(b/a+1)$

表 3.1: 標準ベキ級数分布の性質 (Wani and Lo[211], Table 1)

Dubey[44] および Ord[128] は、あるベキ級数分布の確率関数 $\mathrm{P}(x)$ について、

$$r_x=\frac{x\mathrm{P}(x)}{\mathrm{P}(x-1)}$$

が x の線形関数となると指摘した。この議論を受けて Wani and Lo[211] は、適当な定数 a, b について

$$r_x=\frac{x\mathrm{P}(x)}{\mathrm{P}(x-1)}=ax+b,\quad x=2,3,4,\ldots,$$

となるベキ級数分布を「標準ベキ級数 (Standard power series) 分布」と呼んだ。このクラスには少なくとも、二項分布、ポアソン分布、負の二項分布、対数級数分布、拡張負の二項分布が所属する。これらの分布について、直線 $r_x=ax+b$ の切片と傾きが満たす条件を表 3.1 にまとめておく。

標準ベキ級数分布の確率関数について、$\mathrm{P}(x)=p_x$ と書く。$S_i\propto p_i$ を視覚的に確認するには、

$$\frac{iS_i}{S_{i-1}}$$

が i の線形関数か否かを判断すればよい。実際は標本寸法指標について

$$\rho_i=\frac{is_i}{s_{i-1}}$$

を縦軸、i を横軸にとって図示する。$s_{i-1}=0$ の時は、例えば $s_{i-1}=1$ とすれば良いだろう。線形性の視覚的確認は厳密ではないが、費用対効果の優れた手法と考える。また傾きと切片に関する制約条件から、(最小二乗法を用いるまでもなく) その正負を見ればあてはめるべき分布の見当が付

く。例えば表 3.1 に示された分布の中で、切片と傾きが共に正で有りうるのは負の二項分布のみである。故にデータセット $\{(i, \rho_i)|i = 1, 2, \ldots\}$ にあてはめた直線の傾き、切片が正ならば、負の二項分布をクラスター分布とする複合ポアソン分布から派生するモデルが有力な候補となる。

実は「L 字」の記述という観点から、標準ベキ級数分布の中でも使える分布を更に絞る事が出来る。これについては

$$\frac{s_i}{s_{i-1}} = \frac{\rho_i}{i}$$

について望ましい変化を考える方が分かりやすい。もし標本寸法指標を対数凸と考えるなら、定義 (1.12) より

$$\frac{s_2}{s_1} \leq \frac{s_3}{s_2} \leq \frac{s_4}{s_3} \leq \cdots$$

でなければならない。故に $\rho_i = ai + b$ ならば、標本寸法指標が対数凸という事は

$$\frac{b}{i} \leq \frac{b}{i+1}, \quad i = 1, 2, \ldots$$

と同値となる。これが成立するには b が非正でなければならない。表 3.1 より二項分布とポアソン分布の b は正である。従ってこれらをクラスター分布とする複合ポアソン分布から構成されるモデルの極限では、(平均的に) 寸法指標は対数凸ではない。「L 字」の中でも歪みの強い対数凸のクラスは経験的に有力な候補なので、これを満たさない分布は適用可能性で劣る。また実データでは、傾き a はしばしば正である。図 3.1 から 3.7 は、例 1.8 の労働力調査データセットについて、横軸を $i = 2, 3, \ldots$、縦軸を ρ_i としてプロットしたものである。それぞれ直線を当てはめた場合、切片は 0 の周辺で正か負か定かでない。しかし全ての場合に、右上がりの傾向が見られる。二項分布やポアソン分布の場合、傾きは非正である。従ってこれらの分布をクラスター分布とする複合ポアソン分布を、本稿ではモデルとして考察しない。

なお本節の冒頭で考察した多項分布の場合、(3.2) を変形して $i = 1, 2, \ldots, N$ について

$$\rho'_i = \frac{i\mathrm{E}(S_i)}{\mathrm{E}(S_{i-1})} = \frac{N-(i-1)}{J-1} = -\frac{1}{J-1}i + \frac{N}{J-1}$$

となる。これを i に関する線形関数と見た場合、切片 $N/(J-1)$ は正である。従って既に確認したように、この場合の寸法指標は平均的に対数凸ではない。

ここまで母集団サイズ N が確率変数となるモデリングを、主に考察してきた。次は N が定数となるモデリングを考察する。標本抽出の名簿が固定されている場合には、N が定数という仮定は期待値を制約するよりもリアルである。しかしこのようなモデルは、必然的に組み合わせ論的評価を伴い、操作が容易とは限らない。そもそもセル間の従属性がモデル構築を複雑にする為、各セルで F_j が独立という母集団サイズが確率変数となる場合を考察してきたのだった。以下では N が固定されたモデル構築の一般的議論をする。しかし、結果として得られるモデルが解析的に扱いやすいか否かは個別的問題である。

母集団サイズが固定されたモデルを構築する一つの方法は、N が確率変数のモデルについて、N 所与の条件付分布を求める事である。つまり J が有限のモデルならば

$$\mathrm{P}(F_1, F_2, \ldots, F_J|N) \quad \text{または} \quad \mathrm{P}_J(S_0, S_1, \ldots, S_N|N)$$

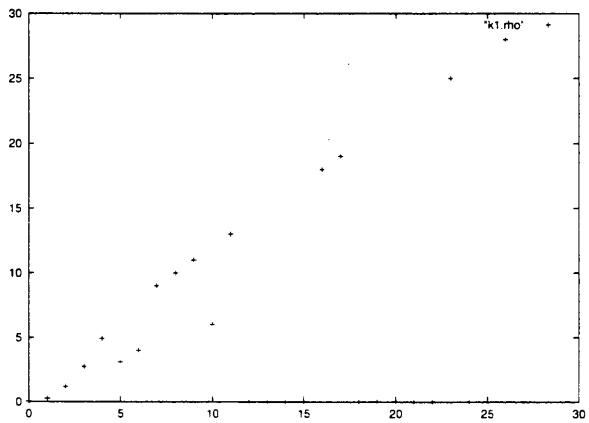

図 3.1: 労働力調査より (竹村 [198], Case 1)

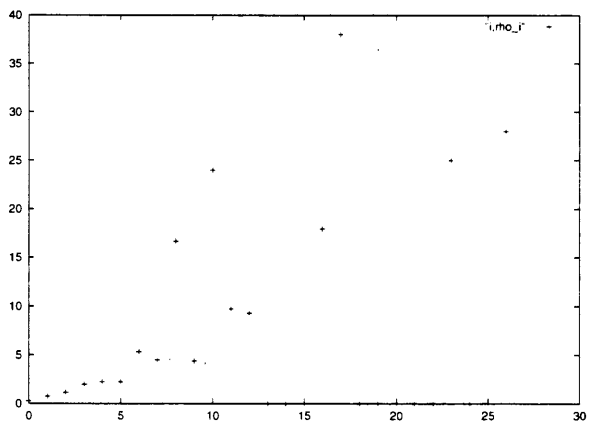

図 3.2: 労働力調査より (竹村 [198], Case 2)

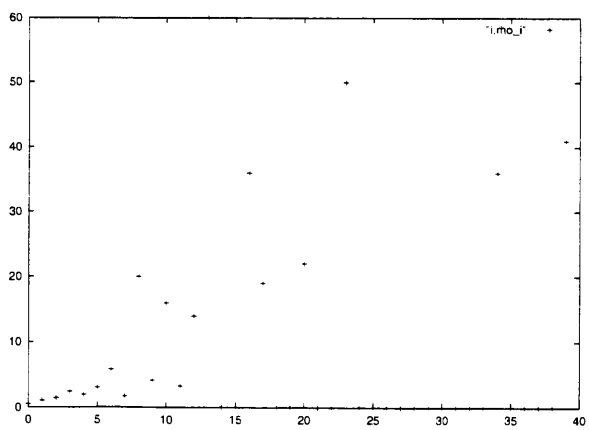

図 3.3: 労働力調査より (竹村 [198], Case 3)

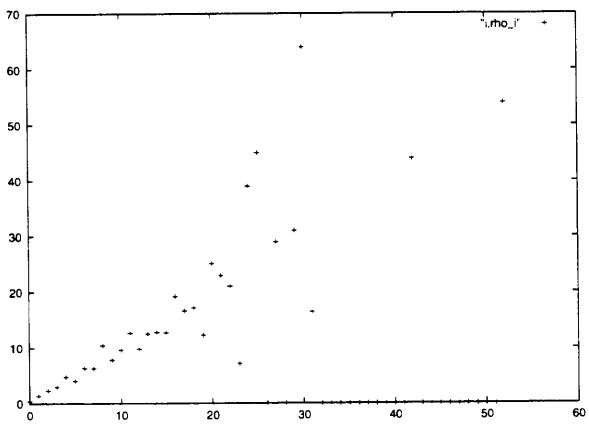

図 3.4: 労働力調査より (竹村 [198], Case 4)

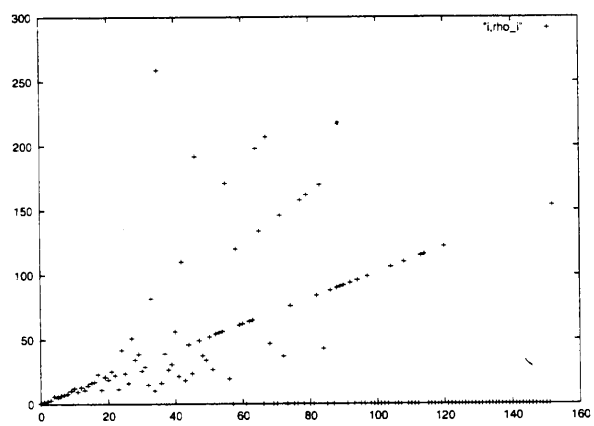

図 3.5: 労働力調査より (竹村 [198], Case 5)

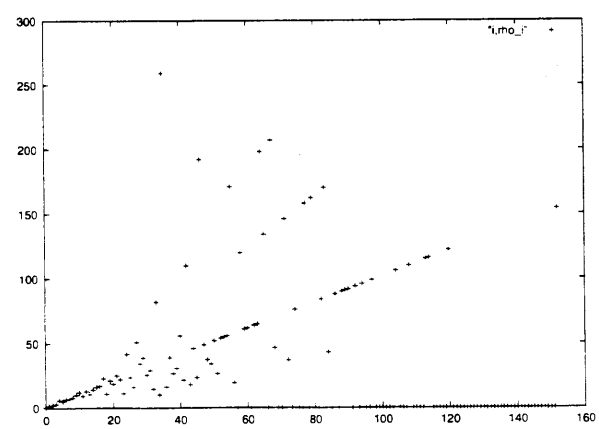

図 3.6: 労働力調査より (竹村 [198], Case 6)

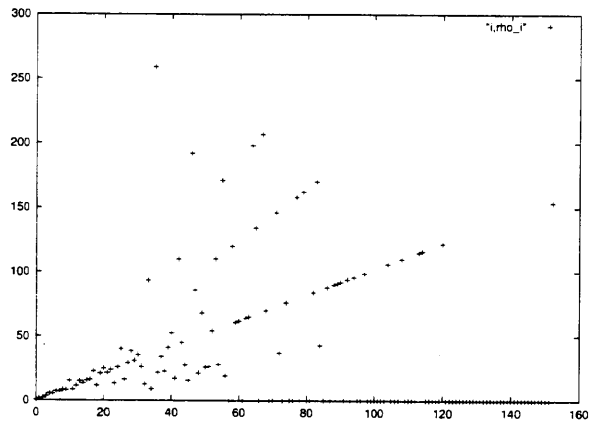

図 3.7: 労働力調査より (竹村 [198], Case 7)

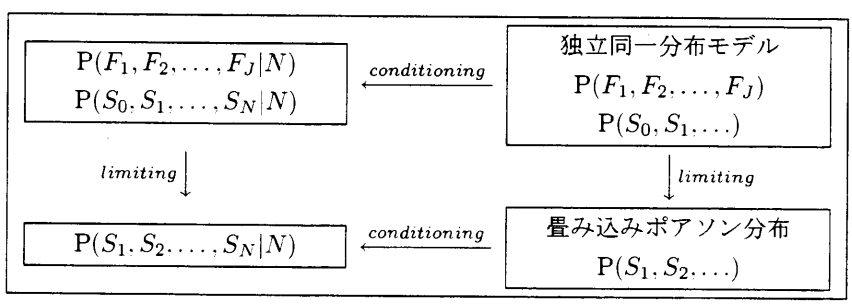

図 3.8: 複合ポアソン分布モデルの 4 形態

を評価する。それから畳み込みポアソン分布では $J \to \infty$ とした結果、S_0 が定義されていない。この場合

$$\mathrm{P}(S_1, S_2, \ldots, S_N | N)$$

を求めれば良い。このように考えれば、今までの議論を活かすことが出来る。すなわち複合ポアソン分布によるモデリングは、4通り有りえるという事であり、その関係を図3.8で表す。なお S_0 の取り扱いについては、次節で詳しく検討する。

問題は N の分布の評価である。一般に N の確率関数の導出は、必ずしも容易ではない。しかし特に F_j が独立同一分布に従う場合、確率母関数を用いる事で見通しが良くなる。すなわち F_j の確率母関数が $G(z)$ ならば、N の確率母関数は $G(z)^J$ であった。ここで N の確率関数評価が簡単になる $G(z)$ のクラスを考察しよう。例えば適当な定数 a, b を用いて $G(z)^J = G(az+b)$ と書けるとする。この時 N の分布は、F_j の分布の位置とスケールを調整する事で得られる。この場合 F_j の確率関数が簡潔ならば、N の確率関数も簡潔となろう。このような性質を持つ非負整数上の分布は有りえるだろうか。

残念ながら、答えは否定的である。確率変数 $X_0, X_1, \ldots, X_J$ が、互いに独立に同一の分布 R に従うとする。$Y_J = \sum_{j=1}^{J} X_j$ と書く。原点に集中していない分布 R について、各 J で

$$Y_J \stackrel{d}{=} a_J X_0 + b_J$$

となるような定数 $a_J > 0, b_J$ が存在すれば、R は「(広義に) 安定」と言う。そして特に $b_J = 0$ を満たす安定な分布は、「狭義に安定」と呼ばれる。つまり安定分布の和の分布の評価は、その位置とスケールを調整するだけで良い。なお狭義の安定分布の定義は

$$X_0 \stackrel{d}{=} \alpha X_1 + (1-\alpha^\gamma)^{1/\gamma} X_2, \quad 0 < \alpha < 1, \tag{3.16}$$

と同値である。より詳しくは Feller[55](p.167)、Uchaikin and Zolotarev[207] (p.43) 等を見よ。容易に確認できるように、安定分布は無限分解可能分布のサブクラスである。しかし Feller[55](p.167) によれば、全ての安定分布は連続である。故に非負整数上の (離散) 無限分解可能分布は、安定な分布を含まないのである。

では非負整数上の無限分解可能分布に対して、別の適当なサブクラスを考えられないだろうか。Steutel and van Harn[195] は、離散分布の安定性を以下のように考察した。まず非負整数上の確率変数 F_0 が、確率母関数 $G(z)$ で定義されるとする。また確率変数 $N_j, j = 1, 2, \ldots,$ は、$\mathrm{P}(N_j = 1) = 1 - \mathrm{P}(N_j = 0) = \alpha$ を満たす。ここで全て独立な確率変数について

$$\alpha \circ F_0 = \sum_{i=1}^{F_0} N_j$$

のように定義する。このように考えれば、$\alpha \circ F_0$ は非負整数上の分布となる。また $1 \circ F_0 \stackrel{d}{=} F_0$、$\mathrm{E}(\alpha \circ F_0) = \alpha \mathrm{E}(F_0)$ のように、スカラーの乗算と同様の性質を持つ。この演算で定義される確率変数の確率母関数は

$$\mathrm{E}(z^{\alpha \circ F_0}) = \sum_{i=0}^{\infty} \mathrm{P}(F_0 = i) \sum_{j=0}^{i} \binom{i}{j} (\alpha z)^j (1-\alpha)^{i-j} = G(1-\alpha+\alpha z)$$

となる。非負整数上の分布 R に従う独立な確率変数 $F_0 \stackrel{d}{=} F_1 \stackrel{d}{=} F_2$ について

$$F_0 \stackrel{d}{=} \alpha \circ F_1 + (1-\alpha^\gamma)^{1/\gamma} \circ F_2, \quad 0 < \alpha < 1. \tag{3.17}$$

が成立する場合、R は（狭義）離散安定と言う。(3.17) 式は、連続の場合の定義式 (3.16) と並行している事に注意すべきである。Steutel and van Harn[195] の Theorem 3.2 によれば、指数 γ $(0 < \gamma \leq 1)$ の離散安定分布は確率母関数

$$\begin{aligned} G(z) &= \exp(-\lambda(1-z)^\gamma), \quad |z| \leq 1, \lambda > 0, \\ &= \exp(\lambda(g(z)-1)), \end{aligned}$$

ただし

$$g(z) = 1-(1-z)^\gamma \tag{3.18}$$

で表される。離散安定分布を複合ポアソンとみなした場合、(3.18) 式で定義されるクラスター分布は「渋谷分布」(Sibuya[163]) と呼ばれる。その確率関数は $0 < \gamma \leq 1$ について

$$\mathrm{P}(X=x) = (-1)^{x+1}\binom{\gamma}{x}, \quad x = 1, 2, \ldots.$$

となる。

離散安定分布を用いたモデリングを考察しよう。各 $F_j, j = 1, 2, \ldots, J,$ が独立に同一離散安定分布に従うとする。この時母集団サイズ N の確率母関数は

$$G(z) = \exp(-J\lambda(1-z)^\gamma), \quad |z| \leq 1, \lambda > 0.$$

で表される。しかしここで

$$G'(1) = \lim_{z \to 1} \frac{J\lambda\gamma}{(1-z)^{1-\gamma}}$$

なので、N の期待値は $\gamma = 1$（ポアソン分布）の場合を除いて発散してしまう。従って離散安定分布モデルにおいて、一般には $\mathrm{E}(N) = N_0$ という制約をおけないという事になる。これでは母集団寸法指標を推測する目的で使いにくい。もっとも N 所与の条件付モデルなら目的に合致するのだが、確率関数が簡潔に書ける場合は限られている。すなわち $\gamma = 1, 1/2$ 及び $\gamma \to 0$ の場合のみ、操作が容易なモデルとなる。ただ各 F_j がポアソン分布に従うようなモデルは本節冒頭で議論したように、経験的に観測される over dispersion のデータを記述しきれない。従って本稿では、$\gamma = 1$ の場合は特に興味の対象とならない。以下では離散安定分布を特殊ケースとして含むような一般化された分布を考察する。この議論の中で、$\gamma = 1/2$ と $\gamma \to 0$ の場合が取り上げられるだろう。

先に複合ポアソンのクラスター分布がベキ級数分布の場合を考察した。この中で拡張負の二項分布（表 1 参照）の確率母関数は

$$g(z) = \frac{1-(1-z\theta)^r}{1-(1-\theta)^r}.$$

ただし $0 < \theta < 1, 0 < r < 1$ であった。ここで $\theta \to 1$ の場合が渋谷分布 (3.18) に他ならない（以下では拡張負の二項分布の母数空間に $\theta = 1$ を含める場合も有る）。そして $r = 1$ の場合がポアソン分布である。従って拡張負の二項分布をクラスター分布とする複合ポアソンの母集団モデルについ

て挙動を明らかにすれば、離散安定分布モデルの性質も特殊ケースとして理解できるだろう。4.12節で詳述されるが、特に $r = 1/2$ の拡張負の二項分布をクラスター分布とする複合ポアソン分布は、逆ガウシアン分布による混合ポアソン分布である。また $r \to 0$ の場合、クラスター分布として

$$g(z) = \frac{\log(1 - z\theta)}{\log(1 - \theta)}$$

が得られる。これは対数級数分布（表1参照）であり、この場合複合ポアソンとして負の二項分布が得られる（例3.3から例3.5を見よ）。逆ガウシアン＝ポアソン混合分布や負の二項分布で構成されるモデルでは、N 所与で条件付分布の確率関数も明示的に求められる。また、小数法則の適用結果も簡潔に表現される。もともと母集団サイズの分布が容易に求められる場合を考察するという事だったが、目的は部分的には達成された。

より一般的に考えて、ベキ級数分布をクラスター分布とする複合ポアソン分布モデルを仮定する。このようなモデルについて、小数法則を適用した結果及び N 所与の条件付分布を確認しておこう。一般に $F_j, j = 1, 2, \ldots, J,$ が互いに独立に同一のベキ級数分布に従う場合、そのベキ母数 θ の十分統計量は N である。従って命題3.8において、N 所与の条件付モデルは θ に依存しない。

命題 3.8 (Hoshino[82], Theorem 2) 確率母関数

$$g(z) = \sum_{i=1}^{\infty} \frac{a_i(z\theta)^i}{\eta(\theta)},$$

ただし

$$\eta(\theta) = \sum_{i=1}^{\infty} a_i\theta^i, \quad a_i \geq 0, \theta > 0,$$

で定義されるベキ級数分布をクラスター分布とする複合ポアソン分布の確率母関数を

$$G(z) = \exp(\alpha\eta(\theta)(g(z) - 1))), \quad \alpha > 0,$$

と書く。$G(z)$ もまたベキ級数分布となる。

$F_j, j = 1, 2, \ldots, J,$ が互いに独立に同一の $G(z)$ で定義される分布に従うとする。この時

$$\mathrm{P}(F_j = i) = \frac{b_i h(\theta)^i}{\exp(\alpha\eta(\theta))},$$

ただし $b_0 = 1$ かつ $b_{i+1} = \alpha(i+1)^{-1} \sum_{j=0}^{i} (i + 1 - j)a_{i+1-j}b_j$ である。

母集団サイズ $N = \sum_{j=1}^{J} F_j$ 所与で $(F_1, F_2, \ldots, F_J)$ の条件付分布は次式で表される。

$$\mathrm{P}(F_1, F_2, \ldots, F_J|N) = \prod_{j=1}^{J} b_{F_j}/d_N, \tag{3.19}$$

ここで d_N は $d_0 = 1$ かつ $d_{i+1} = J\alpha(i+1)^{-1} \sum_{j=0}^{i}(i + 1 - j)a_{i+1-j}d_j$ という漸化式で定まる。

もし $J\alpha$ が A で固定されているならば、$J \to \infty$ とした時に (3.19) の極限分布は

$$\mathrm{P}(S_1, S_2, \ldots, S_N|N) = \frac{A^U}{d_N} \prod_{i=1}^{N} \frac{a_i{}^{S_i}}{S_i!}$$

である。

命題3.8で漸化式を用いて定義したb_i, d_iが簡潔に表現出来るような場合が、便利なモデルという事になる。これを確認するには、個別的な作業が避けられない。次章において、便利なモデルをいくつか紹介しよう。

さて、一般論から離れて具体的に混合ポアソン分布モデルを構築してゆこう。以下では利便性の観点から壷に関する交換可能性、すなわち各jについてλ_jが独立に同一の分布に従う事を仮定する。故に$j = 1, 2, \ldots, J$について

$$\mathrm{P}(F_j = y|\lambda) = \mathrm{P}(F = y|\lambda) = \frac{(N_0\lambda)^y \exp(-N_0\lambda)}{y!}, \quad y = 0, 1, \ldots, \lambda > 0, \tag{3.20}$$

とする。問題はλの密度の仮定だが、「L字」という形状を記述する為には、右に歪んだ分布が必要である。まずλがガンマ分布に従う場合、Fの分布はガンマ＝ポアソン分布（例3.3）であった。λが対数正規分布に従う場合は、Fの分布を対数正規＝ポアソン分布と言う。ただしこの分布は、解析的に扱いやすくはない。更に同様に考えて、（一般化）逆ガウシアン＝ポアソン分布も使える。これはガンマ＝ポアソン分布を特殊ケースとして含むような、一般化された分布である。実はここで挙げた混合ポアソン分布は、全て複合ポアソン分布でもある。故にこれらの独立同一分布モデルに小数法則と条件付けを適用すれば、派生モデルを導出可能である（図3.8を見よ）。

まずN所与の条件付分布を見てゆこう。Sibuya et al.[170]によれば、ガンマ＝ポアソンモデルの母集団サイズに関する条件付分布は、ディリクレ＝多項分布(4.1節)になる。なおTakemura[199]が言うように、ディリクレ＝多項分布の母集団サイズを負の二項分布に従うように混合すると、当然ながらガンマ＝ポアソンモデルを得られる。同様に逆ガウシアン＝ポアソンモデル(4.5節)で母集団サイズNを所与とする場合も評価出来、これを条件付逆ガウシアン＝ポアソン分布(4.9節)と呼ぶ。対数正規＝ポアソンモデル(4.3節)および一般化逆ガウシアン＝ポアソンモデル(4.4節)の条件付分布は、特殊な場合を除いて簡潔な表現は知られていない。

次に、小数法則（$J \to \infty$）を適用して導出されるモデルを示す。例3.5で確認したように、ガンマ＝ポアソンモデルから対数級数モデル(4.6節)を得られる。なお後述されるように、対数級数分布の解釈には幅が有る。本稿の解釈はAnscombe[6]によるものであり、必ずしも標準的なものではない。また我々は同様な極限操作を用いて、ディリクレ＝多項分布よりEwens分布(4.7節)を導く事が出来る。更に本稿は、Pitman分布(4.8節)を考察する。Pitman分布は、Ewens分布とディリクレ＝多項分布を特殊ケースとして含むような一般化された分布である。また逆ガウシアン＝ポアソンモデルからも、小数法則による極限分布を導出可能である。対数級数分布と対数級数モデルの関係から類推して、これを拡張負の二項モデル(4.10節)と呼ぶ。条件付逆ガウシアン＝ポアソン分布についても、小数法則が適用可能である。その極限分布を、極限条件付逆ガウシアン＝ポアソン分布(4.11節)と呼ぶ。

図3.8等では条件付分布を求める操作しか書き込まれていないが、Nについて適当な分布を混合する逆操作も可能である。ここまでの説明から明らかなように、Ewens分布の母集団サイズを負の二項分布に従うように混合すると、対数級数モデルが得られる。また同様に、極限条件付逆ガウシアン＝ポアソン分布の母集団サイズを逆ガウシアン＝ポアソン分布に従うように混合すると、拡張負の二項モデルが得られる。モデル同士の関係については、図3.9及び3.10と、各節の説明を参照の事。

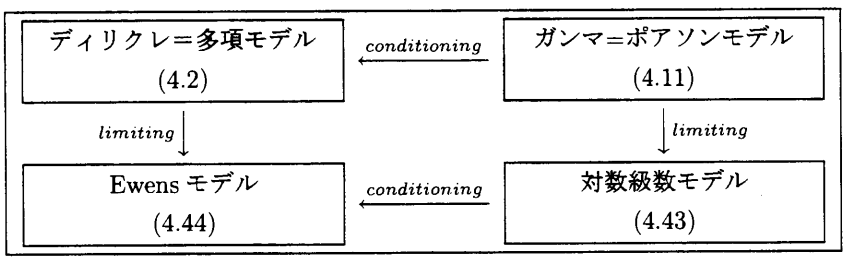

図 3.9: 負の二項分布によるモデリング (Hoshino and Takemura[85])

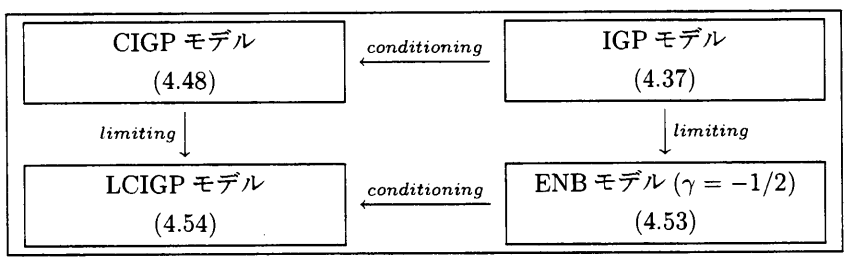

図 3.10: 逆ガウシアン=ポアソン分布よるモデリング (Hoshino[81])

3.2 度数が 0 の集団について

前節の超母集団モデリングでは、各セルで母集団の度数 F_j が 0 かもしれないと考えた。しかし例えば F_j が j 番目の種の個体数だとして、度数が 0 の種はこの世に存在しない事になる。そして母集団に存在しない集団をモデルに取り込む必要はないという批判は、十分考えられる。だとしたら F_j が正の整数上の分布に従うとみなし、$J = U$ としても良いはずである。もし前節の一般的議論が $\mathrm{P}(F_j = 0) = 0$ の場合も含むなら、新たに検討を加えなくても良い。しかし Johnson et al.[90] (p.352) などで指摘されるように、F_j が有限の期待値を持つ複合ポアソン分布に従うならば、$\mathrm{P}(F_j = 0) > 0$ でなければならない（逆に $\mathrm{P}(F_j = 0) = 0$ かつ $\mathrm{E}(F_j) < \infty$ ならば、F_J の分布は複合ポアソンではない)。従って正の整数上の分布によるモデリングは、別に考察する必要が有る。また $J > U$ だとしても、J が正確な情報として得られる場合は限られている。(例えば寄生虫の数を宿主毎に調べるような場合、宿主の数が J である。このような例外を除く。) もし $S_0 = J - U$ が曖昧ならば、その取り扱いに工夫が入る余地があろう。本節では度数 0 の集団の扱いが前節とは異なるモデリングについて、考察を加える。

非負整数上の分布を s_0 が大きいデータに当てはめるとして、s_0 のみ小さくする事であてはまりが大きく改善する場合がある。例えば個票データを用いた Hoshino[80] の実験では、s_0 が多い時に対数正規=ポアソン分布のあてはまりは良くない。この場合 $(s_1, s_2, \ldots)$ へのあてはまりを優先して対数正規=ポアソン分布の母数を定めると、$\mathrm{P}(F = 0)$ が s_0/J よりもかなり小さくなった。またプランクトンの調査ではしばしば s_0 が得られ、なおかつ s_0 が大きくなる。Pennington[133] の議論では、このような場合頻度 0 の一点分布を（対数正規分布に）混合すれば、当てはまりが良くなる。つまり割合 δ で $X = 0$、残り $1 - \delta$ では対数正規分布（$X > 0$）に従うような確率変数を考

えている。Aitchison and Brown[3] は、このような X の分布を Δ-分布と呼んだ。

Δ-分布は、対数正規分布をどうしても使いたい為の工夫に見える（生態学では、個体数の分布が対数正規曲線で記述出来るという主張が有力であった)。しかし明かに対数正規分布以外にも、このような調整を施す事が出来る。Δ-分布の基本的なアイデアは、多すぎるセル数を $J^* = (1-\delta)J$ とみなして減らす事にある。つまり非負整数上の分布を当てはめる時に正の整数上のあてはまりを優先、つじつまは度数 0 の割合で合わせるという事である。一般に考えて非負確率変数 X について、$\mathrm{E}(X|X \neq 0) = \mu, \mathrm{V}(X|X \neq 0) = \sigma^2$ とする。$\mathrm{P}(X \neq 0) = \delta$、$\mathrm{E}(X) = \delta\mu, \mathrm{V}(X) = \delta(1-\delta)\mu^2 + \delta\sigma^2$ である。$x_1, \ldots, x_m, \ldots, x_n$ が標本データ、かつ最初の m 個が非零とする。$a_{(m)}, e_{(m)}, f_{(m)}$ がそれぞれ μ, μ^2, σ^2 の不偏推定量を表す。Aitchison[2] によれば、$\mathrm{E}(X), \mathrm{V}(X)$ の不偏推定量はそれぞれ ($a_{(0)}$ が定義されていないものとして)

$$c = \begin{cases} \frac{ma_{(m)}}{n} & m > 0, \\ 0 & m = 0, \end{cases}$$

$$d = \begin{cases} \frac{mf_{(m)}}{n} + \frac{m(n-m)e_{(m)}}{n(n-1)} & m > 0, \\ 0 & m = 0. \end{cases}$$

のようになる。また $a_{(m)}$ が μ の完備十分統計量ならば、$(m/n, a_{(m)})$ は (δ, μ) の完備十分統計量である事も示されている。この場合 c は MVUE であり、d についても同様の結果が成立する。Pennington[133] の貢献は、

$$\mathrm{V}(c) = \frac{1}{n^2}\mathrm{E}(m^2\mathrm{V}(a_{(m)})) + \frac{\delta(1-\delta)\mu^2}{n},$$

を指摘した事、および $g_{(m)}$ が $\mathrm{V}(a_{(m)})$ の不偏推定量として、

$$\hat{\mathrm{V}(c)} = \begin{cases} \frac{m(m-1)}{n(n-1)}g_{(m)} + \frac{m(n-m)}{n^2(n-1)}a^2_{(m)}, & m > 0 \\ 0 & m = 0, \end{cases}$$

を $\mathrm{V}(c)$ の不偏推定量として提案した事である。特に Δ-分布に依存する結果は、Aitchison and Brown[3] を見よ。また Smith[190] は、Δ-分布について c の厳密な分散を与えている。

Δ-分布を用いる場合、モデルのあてはまりはセル数を $(1-\delta)J$ として前節のようなモデリングをした場合と似るはずである。同様の結果を得るために複雑な手順を踏む必要はないので、Δ-分布を用いたモデリングはこれ以上考察しない。代わりに、モデルの当てはまりが良くなるように J を減らせば良い。例えば前述の Hoshino[80] では、J を母数と考えて対数正規＝ポアソン分布モデルをあてはめている。すなわち、尤度が最大になるようにセル総数を選択した。しかしこのように所与である J を適当に調整してしまうのは、批判を受けるかもしれない。

もし J に関する事前知識が正確でなければ、Δ-分布のような J の調整は許容しやすい。特に個票開示リスク評価問題では、ナイーブに決まる J よりも真の J は小さいという議論（5.3 節で詳述）がある。もし J が大きく定められているならば、J を減らす事で当てはまりが良くなるモデルは、不当に低く評価されているという事になる。また語彙数の計測等、多くの場合に s_0 はデータとして得られない。つまりこの場合、J は定まらない。これらの例は、J を固定する事に疑問を投げかける。Williams[222] が生態学における度数 0 の集団の取り扱いを議論しているが、「標本数が多くほぼ母集団をあまねくカバーしているような場合を除いて、度数 0 の集団に関する正確な知識

を要求するモデルは使用が難しい」と述べている。もしこのような不確実性を考慮するならば、J に依存する推測は避けるべきである。以下では J をモデルから消去する方法として、$J = U$ すなわち $S_0 = 0$ の場合と、$J = \infty$ すなわち $S_0 = \infty$ の場合を考察しよう。

本節冒頭で述べたとおり、各 j 独立に F_j が正の整数上の分布に従うならば $J = U$ となる。この場合形式的にはモデルから J が消え、今度は U の情報確度が問題になる。これで何か変わるのだろうか。ここでは正の整数上の分布の構成方法が問題になる。例えば Chen and Keller-McNulty[36] は、$F_j - 1$ が負の二項分布に従うとしている。この場合 F_j は、正の整数上の分布である。つまり非負整数上の確率変数に 1 を足せば、正の整数上の確率変数が得られる。Chen and Keller-McNulty は気が付いていないようだが、この方法については前節の議論をわずかに修正するだけで良い。つまり $F'_j = F_j - 1, j = 1.2,\ldots,U,$ が非負整数上の分布に従うので、その総和 $N' = N - U = \sum_{j=1}^{U} F'_j$ も非負整数上の分布である。要するに、$J' = U$ 個のセル上で前節のモデリングを利用すれば良い。そして J' 個のセル上で度数が $i, i = 0, 1, 2, \ldots,$ のセル数が、S_{i+1} である。故にこの方法によるモデリングは、特別に議論するまでもない。考察すべきなのは、より伝統的なゼロ切り落とし分布によるモデリングである。

生態学、言語学の分野では s_0 が得られなかった場合、当てはめたい分布の度数 0 を切り落とす例が多い。非負整数上の分布に従う確率変数 F について、ゼロ切り落とし分布は確率関数

$$\mathrm{P}(F = x | F \geq 1) = \frac{\mathrm{P}(F = x)}{1 - \mathrm{P}(F = 0)}$$

で定義される。なお F の確率母関数が $G(z)$ の時 $\mathrm{P}(F = 0) = G(0)$ である事に注意すれば、F のゼロ切り落とし分布の確率母関数は

$$\frac{G(z) - G(0)}{1 - G(0)} \tag{3.21}$$

で表される。ゼロ切り落とし分布が用いられた具体例として、負の二項分布については Sampford[150]、対数正規＝ポアソン分布については Bulmer[23]、逆ガウシアン＝ポアソン分布については Ord and Whitmore[129]、Sichel[177]、一般化逆ガウシアン＝ポアソン分布については Sichel[180] 等を挙げておく。

しかし使われてきたとはいうものの、ゼロ切り落とし分布によるモデリングは利便性に欠ける。確率変数 F が期待値有限の複合ポアソン分布に従い、X がそのゼロ切り落とし分布に従うとしよう。先に述べたとおり、X の分布は複合ポアソンではない。そして $X - 1$ は非負整数上の分布となるが、複合ポアソンになるとは限らない。例えばポアソン分布は複合ポアソンである。しかし Kemp[97] によれば、F がポアソン分布に従う場合 $X - 1$ の分布は複合ポアソンではない。故に正の整数上の分布によるモデリングでは、前節の一般論が必ずしも適用できない。また母集団分布が正の整数上の分布だとしても、非復元単純無作為抽出とベルヌーイ抽出の下では、一般に標本分布は非負整数上の分布である。従って、適当な標本抽出法についてモデルが閉じるという命題 2.1、2.2 のような結果は、正の整数上の分布によるモデリングでは期待できない。つまり超母集団モデルによる寸法指標の推定を簡潔だからこそ採用する本稿の立場では、$J = U$ となるモデリングは意味を持たない。

更に問題なのは s_0 が得られない場合、非負整数上の分布モデルの母数を、ゼロ切り落とし分布の当てはめにより「推定」するケースである。最尤法の例で説明しよう。適当なベルヌーイ抽出の結果、$f_j. j = 1.2.\ldots.J.$ が各 j 独立に $\mathrm{P}(f_j = i) = p_i. i = 0.1.2.\ldots$ という分布に従うとする。対

数尤度は

$$\ell = \sum_{i=0}^{\infty} s_i \log p_i + \text{const.}$$

となる。ここで s_0 が未知な為、生態学などで切り落とし分布 $\mathrm{P}(i) = p_i/(1-p_0), i = 1, 2, \ldots$ を当てはめる例が見られる。すなわち ℓ を最大化する代わりに

$$l = \sum_{i=1}^{\infty} s_i \log p_i - u \log(1-p_0) + \text{const.}$$

を最大化するという事である（$\chi^2 = \sum_{i=1}^{\infty}((s_i - \mathrm{E}(s_i))/\mathrm{E}(s_i))^2$ を最小化している場合が多い）。ℓ と l の最大化が一致するのは

$$p_0^{s_0} = \frac{1}{(1-p_0)^u}$$

の時だけであり、これは明らかに一般的でない。この条件は $p_0 = 0$ なら満たされるが、この場合 $u = U$ となって全く観測されなかった集団が無い事になってしまう。このようなモデリングは事実上、役に立たない。また l を最大化して得られた母数の「推定値」の下で計算した p_0 を $\hat{p_0}$ とする。Bulmer[23] の対数正規＝ポアソン分布モデルに関する議論にならえば、未知の J を

$$\hat{J} = \frac{u}{1-\hat{p_0}}$$

で推定出来るかもしれない。しかし $\hat{p_0}$ が依存する母数の「推定値」が望ましい性質を持たない以上、Bulmer 等の議論には疑問が残る。Engen[49](p.101) は、対数正規=ポアソン分布の母数の推定値が（同一母集団からのデータであるにも関わらず）標本数を変える事で変化する例を示している。これは ℓ と l の違いを意識していない事が原因かもしれない。

要するに J が有限の場合、母集団寸法指標を推定する為の情報として s_0 を無視してはならない。応用局面では、しばしば何らかの理由で特定の大きさの寸法指標が重視される。例えば個票開示リスク評価では、個体数が小さなセル数が興味の対象である。故に Chen and Keller-McNulty[36] や Skinner and Holmes[185] は、負の二項分布のあてはまりを重視するあまり右裾のデータ（i が大きいところの s_i）を切り落とす事を推奨している。また生態学の例で、$(s_1, s_2, \ldots)$ という情報を適当な分布で要約・記述する事が目的なら、s_0 を無視してゼロ切り落とし分布をあてはめても良いだろう。しかし母集団寸法指標を推測するには、特定の大きさの標本寸法指標の尤度への貢献を捨てるべきではない。結局不確実な J に依存しない推測の方法として、J が有限となるモデリングは適切ではない。

前節において小数法則を適用する際、J を無限大とした。この極限操作によって、モデルの J への依存が除去されている事に注目すべきである。また小数法則の極限において、空でない各セルでの度数は、以下で説明するようなゼロ切り落とし分布に従うと解釈出来る。だとすれば、$J = \infty$ ではなく $J = U$ とする理由は無い。

確率変数 F が複合ポアソン分布に従い、その確率母関数が

$$G(z) = \exp(\mu(g(z) - 1)). \quad g(z) = \sum_{i=1}^{\infty} z^i p_i, \quad \sum_{i=1}^{\infty} p_i = 1\,(p_i \geq 0),$$

で表されるとしよう。Kemp[97] によれば、

$$g(z) = \lim_{\mu \to 0} \frac{G(z) - G(0)}{1 - G(0)}, \tag{3.22}$$

つまり F のゼロ切り落とし分布 (3.21) の極限が、クラスター分布 $g(z)$ である。例えば対数級数分布は、負の二項分布のゼロ切り落とし分布の極限として得られる。また拡張負の二項分布の特殊な場合を、逆ガウシアン＝ポアソン分布の極限として得る事が出来る。渋谷 [169] の Lemma 1 によれば、平均 λ のポアソン分布でゼロを打ち切り $\lambda \to 0$ とすれば、確率 1 で 1 の値をとる分布に退化する。(3.8) 式で表される複合ポアソン分布の N が確率 1 で 1 をとる確率変数ならば、S_N の分布はクラスター分布と同等である。従ってゼロ打ち切り複合ポアソン分布において、平均母数を 0 とする極限 (3.22) はクラスター分布でなければならない。

(3.22) 式で表される関係は、命題 3.7 と整合的である。直感的に説明してみよう。$F_j, j = 1, 2, \ldots, J$, が独立同一に F の分布に従う場合、$\mathrm{E}(N)$ を固定して $J \to \infty$ とすれば、F のゼロ切り落とし分布は退化しない。そして $\mu \to 0$ となるので、各 F_j が 1 以上という条件付分布は $g(z)$ で定まる。この場合極限で $F_j = i, i = 1, 2, \ldots,$ という事象の生起数 S_i の平均は、p_i に比例する。またセル数が無限大で $F_j = i, i = 1.2, \ldots,$ という事象の確率が 0 に近ければ、二項分布のポアソン近似が成立する。これをまとめたものが、命題 3.7 である。つまり複合ポアソン分布に小数法則を適用して得られるモデルでは、各セルで度数のゼロ切り落とし分布がクラスター分布に従うとみなしている事になる。

ゼロ切り落とし分布だけでなく、より一般的に考えて任意の正の整数上の分布を、複合ポアソン分布のクラスター分布として用いる事が出来よう。この場合小数法則（$J \to \infty$）の極限において、空でない各セルの度数がその正の整数上の分布に従うと解釈可能である。つまり正の整数上の分布でモデリングする場合、セル数を無限大とし、度数が 1 以上の条件付分布として用いるべきである。このようなモデルは J に依存しないので、大変都合が良い。ただし応用時は、J が無限大という暗黙の前提について検討が必須である。

本節の議論をまとめておく。前節で展開した複合ポアソン分布モデルの一般論は、非負整数上の分布によるモデリングと正の整数上の分布によるモデリングを共に包含する。J が既知（有限）ならば、非負整数上の分布を用いてモデリングすれば良い。また J が未知ならば、正の整数上の分布 $\mathrm{P}(i) = p_i, i = 1, 2, \ldots,$ によるモデリングが利用できる。ただし伝統的な $J = U < \infty$ とする方法ではなく、$J = \infty$ として各 S_i が独立なポアソン分布、$\mathrm{E}(S_i) \propto p_i$ となるように構成すべきである。このようなモデルは母集団と標本の関係が単純なだけでなく、複合ポアソン分布モデルの極限として正当化される。また母集団サイズ N の分布を、容易に評価出来る。

第4章　超母集団モデル各論

本章では寸法指標推測に用いられる超母集団モデルについて、個別の議論をまとめる。各超母集団モデルの、分布に依存した寸法指標の挙動が整理される。具体的には、分布の導出、他分布との関係、モメント、母数推定に関する結果が示される。なおモデルに用いられる分布の引数が満たす制約は、全て共通である。母集団に関して $F_j, j = 1, 2, \ldots, J,$ は非負整数であり、(1.1), (1.2), (1.3) を満たす。また標本に関して $f_j, j = 1, 2, \ldots, J,$ は非負整数であり、(1.4), (1.5), (1.6) が成立する。

4.1　ディリクレ＝多項モデル

4.1.1　定義等

母集団が確率的に生成されるとして最も直感的な発想では、各壷が $(\lambda_1, \ldots, \lambda_J)$ という確率を持つような多項分布 (2.6) を考えるという事であった。ディリクレ＝多項分布では、各壷の確率についてディリクレ分布

$$\mathrm{P}(\lambda_1, \ldots, \lambda_J) = \frac{\Gamma(\gamma_1 + \cdots + \gamma_J)}{\Gamma(\gamma_1) \cdots \Gamma(\gamma_J)} {\lambda_1}^{\gamma_1 - 1} \cdots {\lambda_J}^{\gamma_J - 1}, \quad \lambda_j > 0, \gamma_j > 0, j = 1, 2, \ldots, J,$$

を混合する。なお $\lambda_j, j = 1, 2, \ldots, J,$ が互いに独立に密度

$$f(\lambda_j) = \frac{1}{\Gamma(\gamma_j)} {\lambda_j}^{\gamma - 1} \exp(-\lambda_j), \quad \lambda_j > 0, \gamma_j > 0,$$

のガンマ分布に従う場合、$\sum_{j=1}^{J} \lambda_j$ を一定にしたシンプレックス上の分布がディリクレ分布である。なお本節の議論全般については、Johnson et al.[89] の 35.13.1 節、及び Takemura[199] に詳しい。

定義　ディリクレ＝多項モデルでは母集団サイズ N は定数であり、$\gamma_j > 0, j = 1, 2, \ldots, J,$ について

$$\mathrm{P}(F_1 = y_1, \ldots, F_J = y_J) = \frac{N! \Gamma(\gamma_1 + \cdots + \gamma_J)}{\Gamma(\gamma_1 + \cdots + \gamma_J + N)} \frac{\Gamma(\gamma_1 + y_1)}{\Gamma(\gamma_1) y_1!} \cdots \frac{\Gamma(\gamma_J + y_J)}{\Gamma(\gamma_J) y_J!}$$

と定義される。なお壷に関する交換可能性 $\gamma_1 = \cdots = \gamma_J = \gamma$ を仮定すれば、$\gamma > 0$ について

$$\mathrm{P}(F_1 = y_1, \ldots, F_J = y_J) = \frac{N! \Gamma(J\gamma)}{\Gamma(J\gamma + N)} \frac{\Gamma(\gamma + y_1)}{\Gamma(\gamma) y_1!} \cdots \frac{\Gamma(\gamma + y_J)}{\Gamma(\gamma) y_J!} \tag{4.1}$$

となる。この時寸法指標に関して

$$\mathrm{P}(S_0, \ldots, S_N) = \frac{N! J! \Gamma(J\gamma)}{\Gamma(J\gamma + N)} \prod_{i=0}^{N} \left(\frac{\Gamma(\gamma + i)}{\Gamma(\gamma) i!}\right)^{S_i} \frac{1}{S_i!} \tag{4.2}$$

を得る。

以下では壷に関して交換可能なモデル (4.1) のみ考察する。このモデルでは母集団サイズが固定されているが、N が負の二項分布 (4.14) に従うとしよう。この場合 (4.1) に (4.14) を掛けて、N が確率変数となるモデル (4.11)、すなわちガンマ=ポアソンモデルを得られる。逆にガンマ=ポアソンモデルの N に関する条件付分布がディリクレ=多項分布である (Sibuya et al.[170])。次に $J\gamma = 1/\beta = \theta$ とおく。ここで θ を固定して、$J \to \infty, \gamma \to 0$ という小数法則に対応した極限操作を考える。Watterson[216], Takemura[199] 等の議論によれば、結果として Ewens 分布 (4.44) が得られる。また N を固定して $\gamma \to \infty$ の時、モデルは等確率多項分布 (4.61) に収束する。その他の極限については Paul and Plackett[132] を見よ。なおディリクレ=多項分布は後に議論する Pitman 分布の特殊ケースである。

ディリクレ=多項分布はベータ=二項分布の自然な一般化である。Mosimann[119] によって導出され、古くから様々な名前で研究されている。例えば Ishii and Hayakawa[86] では、多変数ベータ=二項分布と呼ばれている。実際多項分布の周辺二変数の分布は二項分布であり、ディリクレ分布の周辺二変数の分布はベータ分布である。この事から分かるように、(4.1) で F_j の周辺分布は、二項分布 $\mathrm{Bin}(N, \lambda_j)$ について λ_j がベータ分布 $\mathrm{Beta}(\gamma, (J-1)\gamma)$ に従う。すなわち、

$$\begin{aligned} \mathrm{P}(F_j = y) &= \binom{N}{y} \frac{\Gamma(J\gamma)}{\Gamma(\gamma)\Gamma((J-1)\gamma)} \int_0^1 \lambda_j{}^{y+\gamma-1}(1-\lambda_j)^{N-y+(J-1)\gamma-1} d\lambda_j \\ &= \frac{\Gamma(N+1)\Gamma(J\gamma)\Gamma(y+\gamma)\Gamma(N-y+(J-1)\gamma)}{\Gamma(y+1)\Gamma(N-y+1)\Gamma(\gamma)\Gamma((J-1)\gamma)\Gamma(N+J\gamma)} \\ &= \binom{y+\gamma-1}{y}\binom{N-y+(J-1)\gamma-1}{N-y} \Big/ \binom{N+J\gamma-1}{N} \end{aligned} \tag{4.3}$$

である。なお (4.3) 式は γ が整数の時、負の超幾何分布である。また同様に考えれば、$j = 1, 2, \ldots, J, l = 1, 2, \ldots, J\ (l \neq j)$ について

$$\mathrm{P}(F_j = y_j, F_l = y_l) = \frac{\Gamma(N+1)\Gamma(J\gamma)\Gamma(\gamma+y_j)\Gamma(\gamma+y_l)\Gamma((J-2)\gamma+N-y_j-y_l)}{\Gamma(J\gamma+N)\Gamma(\gamma)^2\Gamma(y_j)\Gamma(y_l)\Gamma((J-2)\gamma)\Gamma(N-y_j-y_l)} \tag{4.4}$$

のように書ける。

4.1.2 モメント

本節でも (4.1) 式を前提とする。確率の総和が 1 になる事を利用して、階乗モメントを導出出来る。すなわち非負整数の $r_j, j = 1, 2, \ldots, N,$ について

$$\mathrm{E}\Big(\prod_{j=1}^{J} F_j^{(r_j)}\Big) = \frac{N^{(R)} \prod_{j=1}^{J} \gamma^{[r_j]}}{(J\gamma)^{[R]}}. \tag{4.5}$$

但し $\sum_{j=1}^{J} r_j = R\,(\leq N)$, $x^{(r)} = x(x-1)\cdots(x-r+1)$, $x^{[r]} = x(x+1)\cdots(x+r-1)$ である。なお壷が交換可能でない場合についても同様な議論が出来る (Sibuya et al.[170])。特に $j = 1, \ldots, J$ について

$$\mathrm{E}(F_j) = \frac{N}{J}.$$

$$\mathrm{V}(F_j) = \frac{N(J-1)(J\gamma+N)}{J^2(J\gamma+1)}$$

が成立する。寸法指標についても同様に評価可能で、非負整数の $r_i, i = 0, 1, \ldots, N$, について

$$\mathrm{E}(\prod_{i=0}^{N} S_i^{(r_i)}) = \frac{N!J!\Gamma(J\gamma)\Gamma((J-r)\gamma+N-R)}{(N-R)!(J-r)!\Gamma((J-r)\gamma)\Gamma(J\gamma+N)} \prod_{i=0}^{N} (\frac{\Gamma(\gamma+i)}{\Gamma(\gamma)i!})^{r_i},$$

ただし $R = \sum_{i=0}^{N} ir_i, r = \sum_{i=0}^{N} r_i$ となる。または (2.3) より (4.3) を利用して、特に

$$\mathrm{E}(S_i) = \sum_{j=1}^{J} \mathrm{P}(F_j = i) = J\binom{i+\gamma-1}{i}\binom{N-i+(J-1)\gamma-1}{N-i} / \binom{N+J\gamma-1}{N}$$

と評価出来る。また (2.4) と (4.4) より、分散は

$$\mathrm{V}(S_i) = \mathrm{E}(S_i) - \mathrm{E}(S_i)^2 + J(J-1)\frac{\Gamma(N+1)\Gamma(J\gamma)\Gamma(\gamma+i)^2\Gamma((J-2)\gamma+N-2i)}{\Gamma(J\gamma+N)\Gamma(\gamma)^2\Gamma(i)^2\Gamma((J-2)\gamma)\Gamma(N-2i)}$$

となる。

4.1.3 母数推定

母数 γ の推定については、Mosimann[119], Takemura[199] 等で議論されている。非復元単純無作為抽出で、大きさ n の標本が取られたとする。まず標本分布は命題 2.1 より、以下のように書ける。

$$\mathrm{P}(s_0, \ldots, s_n) = \frac{n!J!\Gamma(J\gamma)}{\Gamma(J\gamma+n)} \prod_{i=0}^{n} (\frac{\Gamma(\gamma+i)}{\Gamma(\gamma)i!})^{s_i} \frac{1}{s_i!}. \tag{4.6}$$

まず最尤法から考察しよう。Levin and Reeds[109] は、尤度関数 (4.6) を詳細に分析した。最も重要な結果は、尤度が γ に関して単峰という Good[62] の予想を証明した事である。壷に関して交換可能な場合について、彼らの結果を引用しておく。

命題 4.1 (Levin and Reeds[109], Theorem 1) (4.6) 式の右辺を $l(\gamma)$ と書く。尤度関数 $l(\gamma)$ は高々一つ極大をとる。

$$\chi^2 = \frac{J}{n} \sum_{i=0}^{n} s_i (i - \frac{n}{J})^2$$

と書く。もし有限の γ について

$$\chi^2 > J - 1 \tag{4.7}$$

ならば、$\log l(\gamma)$ に関する全ての次数の微分係数は $(0, \infty)$ でちょうど一箇所 0 となる。もし条件 (4.7) が満たされないなら、いかなる $\log l(\gamma)$ の微分係数も $(0, \infty)$ で 0 とならない。

命題 4.1 は、最尤推定値が極大を求めるアルゴリズムで得られる事を保証する。Hoshino[80] は、Vetterling et al.[210] の一変数関数極大化アルゴリズムを用いて最尤解を計算した。または (4.6) 式の右辺の対数を取れば

$$L(\gamma) = \log l(\gamma) = \sum_{i=1}^{n} s_i \{\sum_{j=0}^{i-1} \log(\gamma+j)\} - \sum_{j=0}^{n-1} \log(J\gamma+j) + \mathrm{const.}$$

のように対数尤度が評価出来る。微分係数

$$\frac{dL(\gamma)}{d\gamma} = \sum_{i=1}^{n} s_i \{\sum_{j=0}^{i-1} \frac{1}{\gamma + j}\} - \sum_{j=0}^{n-1} \frac{J}{J\gamma + j}$$

$$\frac{d^2L(\gamma)}{d\gamma^2} = \sum_{i=1}^{n} s_i \{\sum_{j=0}^{i-1} \frac{-1}{(\gamma + j)^2}\} + \sum_{j=0}^{n-1} \frac{J^2}{(J\gamma + j)^2}$$

が明かなので、ニュートン＝ラフソン法など高速に収束する数値解法が利用できる。

次にモメント法を試す。(4.5) 式と命題 2.1 より、

$$\mathrm{E}[f_j(f_j - 1)] = \frac{n(n-1)(\gamma + 1)}{J\gamma + 1}$$

である。故に

$$T = \frac{1}{n(n-1)} \sum_{i=2} i(i-1)s_i$$

の時、

$$\mathrm{E}(T) = \frac{\gamma + 1}{J\gamma + 1}$$

である。Takemura[199] では、$T = (\hat{\gamma} + 1)/(J\hat{\gamma} + 1)$ を解いて

$$\hat{\gamma}_{Takemura} = \frac{1 - T}{JT - 1},$$

を推定量としている。また標本分散を

$$v = \frac{\sum_{i=0} s_i (i - n/J)^2}{J - 1} \tag{4.8}$$

と書く。この時

$$\frac{n-1}{J\hat{\gamma}_{Takemura} + 1} = \frac{J}{n} v - 1$$

が成立する。ここで左辺を $n/(J\hat{\gamma}_{Takemura})$ で近似すれば、Bethlehem et al.[15] の推定量

$$\hat{\gamma}_{Bethlehem} = \frac{n}{J(Jv/n - 1)} \tag{4.9}$$

を得る。ただし元々これは、ガンマ＝ポアソン分布の推定量として提案された。他に Mosimann[119] は、多項分布とディリクレ＝多項分布の分散共分散行列の関係から、母数の推定を考察している。

4.2 ガンマ＝ポアソンモデル

4.2.1 定義等

ガンマ＝ポアソンモデルでは、(3.20) 式の λ がガンマ分布

$$f(\lambda) = \frac{\lambda^{\gamma - 1} \exp(-\lambda/\beta)}{\Gamma(\gamma)\beta^{\gamma}}, \quad \lambda > 0,$$

ただし $\gamma > 0, \beta > 0,$ に従うとみなす。なおガンマ分布については、Johnson et al.[88] の Chap. 17 を見よ。混合された分布は Greenwood and Yule[65] によれば

$$\begin{aligned} \mathrm{P}(F = y) &= \int_0^\infty \frac{(\lambda N_0)^y \exp(-\lambda N_0)}{y!} \frac{\lambda^{\gamma-1} \exp(-\lambda/\beta)}{\Gamma(\gamma)\beta^\gamma} d\lambda \\ &= \frac{\Gamma(y+\gamma)}{\Gamma(\gamma)y!} \frac{1}{(N_0\beta)^\gamma} \left(\frac{N_0\beta}{N_0\beta+1}\right)^{y+\gamma} \\ &= \frac{\Gamma(y+\gamma)}{\Gamma(\gamma)y!} p^\gamma \cdot q^y, \quad y = 0, 1, 2, \ldots, \end{aligned} \tag{4.10}$$

但し $p = 1/(N_0\beta + 1), q = 1 - p$ であり、これは負の二項分布である。負の二項分布の性質については Johnson et al.[90] の Chap. 5 を見よ。特に $\gamma = 1$ の時、幾何分布と言う。

定義 ガンマ＝ポアソンモデルでは N が確率変数となり、同時確率関数

$$\mathrm{P}(F_1 = y_1, \ldots, F_J = y_J) = \prod_{j=1}^{J} \frac{\Gamma(y_j+\gamma)}{\Gamma(\gamma)y_j!} p^\gamma \cdot q^{y_j} \tag{4.11}$$

で定義される $(p = 1/(N_0\beta + 1), q = 1 - p)$。ここで γ, β は正である。なお寸法指標で表せば

$$\mathrm{P}(S_0, \ldots) = J! \prod_{i=0}^{\infty} \left(\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)Ki!} p^\gamma \cdot q^i\right)^{S_i} \frac{1}{S_i!} \tag{4.12}$$

となる。なお期待値制約 $\mathrm{E}(N) = N_0$ を満たすため、本稿では

$$\gamma\beta = 1/J \tag{4.13}$$

となる。

本モデルは個票開示リスク評価の分野では、Bethlehem et al.[15] 以来「ポアソン・ガンマモデル」と呼ばれる。ガンマ＝ポアソンはモデルとして古典的であり、寸法指標推測に関しても Engen[49] にて詳しく解説されている。

負の二項分布が畳み込みに関して閉じている事から、母集団サイズ N の分布は負の二項分布

$$\mathrm{P}(N) = \frac{\Gamma(J\gamma+N)}{N!\Gamma(J\gamma)} p^{J\gamma} q^N, \quad N = 0, 1, \ldots, \tag{4.14}$$

但し $p = 1/(N_0\beta), q = 1 - p$ になる。この事から Sibuya et al.[170] は、母集団サイズ N の条件付ガンマ＝ポアソンモデル $\mathrm{P}(S_0, \ldots, S_N|N)$ がディリクレ＝多項モデル (4.2) になると指摘する。なおこの結果は、壷が交換可能でないモデルについても成立する。また Anscombe[6] によると、$J\gamma = 1/\beta$ を固定して $\gamma \to 0\,(J \to \infty)$ とすれば、対数級数モデル (4.43) が得られる。

ガンマ＝ポアソンモデルは母数 γ を大きくする事で、S_i のモードが $i \geq 2$ にあるようなデータも記述が出来る。このようなデータは「L 字型」の仮定から外れるが、統計的生態学・計量言語学分野では時々あらわれる。従って「L 字型」の記述に特化したモデルにあてはまりで負ける場合は多いとしても、母集団に関する事前知識が曖昧な場合は本モデルも適用してみるべきである。

4.2.2 モメント

周辺の負の二項分布 (4.10) については階乗モメント

$$\mathrm{E}(F^{(r)}) = (\frac{q}{p})^r \gamma^{[r]}$$

が、確率の総和が 1 になる事を利用して導出できる。ただし r は非負整数である。もちろん命題 3.2 を用いて、ガンマ分布のモメントから求めても良い。$\mathrm{E}(N) = J\mathrm{E}(F) = Jq\gamma/p = N_0$ なので、確かに母集団サイズの制約 (4.13) を満たしている事が分る。また特に

$$\mathrm{E}(F) = \frac{q\gamma}{p} = \frac{N_0}{J},$$

$$\mathrm{V}(F) = (\frac{p^2}{q^2} + \frac{p}{q})\gamma = \frac{N_0}{J}(N_0\beta + 1)$$

である。寸法指標の期待値は、(2.3) を利用して (4.10) から

$$\mathrm{E}(S_i) = \sum_{j=1}^{J} \mathrm{P}(F_j = i) = J\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)i!}(\frac{1}{N_0\beta+1})^\gamma \cdot (\frac{N_0\beta}{N_0\beta+1})^i \tag{4.15}$$

となる。また分散は $\mathrm{P}(F_k = i, F_l = i) = \mathrm{P}(F = i)^2$ なので、(2.4) より

$$\mathrm{V}(S_i) = \mathrm{E}(S_i) - \mathrm{E}(S_i)^2 + J(J-1)(\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)i!}(\frac{1}{N_0\beta+1})^\gamma \cdot (\frac{N_0\beta}{N_0\beta+1})^i)^2$$

である。

4.2.3 母数推定

負の二項分布に関する母数の推定は、古典的な議論が充実している。例えば Johnson et al.[90] の 5.8 節、Engen[49] の 3.4 節を見よ。標本分布については母集団サイズがランダムなので、抽出率 n_0/N_0 のベルヌーイ抽出を利用する事とする。この時命題 2.2 より

$$\mathrm{P}(s_0.\ldots) = J!\prod_{i=0}^{\infty}(\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)i!}(\frac{1}{n_0\beta+1})^\gamma \cdot (\frac{n_0\beta}{n_0\beta+1})^i)^{s_i}\frac{1}{s_i!}, \tag{4.16}$$

となる事が分る。(4.16) 式から対数尤度は

$$L = \sum_{i=0}^{\infty} s_i\{\log(\gamma\cdot(\gamma+1)\cdots(\gamma+i-1)) - \gamma\log(n_0\beta+1) + i\log(\beta) - i\log(n_0\beta+1)\} + \mathrm{const.}$$

ここで最尤法を考えよう。期待値制約 (4.13) が無ければ、

$$\frac{\partial L}{\partial\gamma} = \sum_{i=0}^{\infty} s_i\{\frac{1}{\gamma} + \cdots + \frac{1}{\gamma+i-1} - \log(n_0\beta+1)\} = 0.$$

$$\frac{\partial L}{\partial\beta} = \sum_{i=0}^{\infty} s_i\{\frac{i}{\beta} - \frac{(\gamma+i)n_0}{n_0\beta+1}\} = 0$$

を数値的に解けば良い。二次の微分係数

$$\frac{\partial^2 L}{\partial \gamma^2} = \sum_{i=0}^{\infty} s_i \{ \frac{-1}{\gamma^2} + \cdots + \frac{-1}{(\gamma+i-1)^2} \},$$

$$\frac{\partial^2 L}{\partial \beta^2} = \sum_{i=0}^{\infty} s_i \{ \frac{-i}{\beta^2} - \frac{(\gamma+i){n_0}^2}{(n_0\beta+1)^2} \},$$

$$\frac{\partial^2 L}{\partial \gamma \partial \beta} = \sum_{i=0}^{\infty} s_i (\frac{-n_0}{n_0\beta+1})$$

が利用できる。制約式 (4.13) の下では、ラグランジュ乗数法を用いるか、$\beta = 1/(J\gamma)$ を代入するかして解く事ができるだろう。代入した場合は

$$\frac{\partial L}{\partial \gamma} = \sum_{i=0}^{\infty} s_i \{ \frac{1}{\gamma} + \cdots + \frac{1}{\gamma+i-1} - \log(\frac{n_0}{J\gamma}+1) + \frac{(\gamma+i)n_0}{\gamma(n_0+J\gamma)} - 1 - \log\gamma \} = 0$$

を解けば良い。なお

$$\frac{\partial^2 L}{\partial \gamma^2} = \sum_{i=0}^{\infty} s_i \{ \frac{-1}{\gamma^2} + \cdots + \frac{-1}{(\gamma+i-1)^2} + \frac{n_0}{\gamma(n_0+J\gamma)} + \frac{n_0\gamma(n_0+J\gamma) - (n_0+2J\gamma)(n_0\gamma+in_0)}{\gamma^2(n_0+J\gamma)^2} - \frac{1}{\gamma} \}$$

である。

次に非復元単純無作為抽出を考えよう。この時 4.2.1 節で言及したように、$\mathrm{P}(F_1,\ldots,F_J|N)$ はディリクレ＝多項モデルになる。この場合命題 2.1 から $\mathrm{P}(s_0,\ldots|n)$ は、ディリクレ＝多項モデル (4.6) となる。つまりこの場合母数の推定は、4.1 節の議論に従えば良い。モメント推定量 (4.9) は本節の議論での反復数値解法において、初期値として用いる事も出来るだろう。

4.3 対数正規＝ポアソンモデル

4.3.1 定義等

対数正規＝ポアソンモデルでは、(3.20) 式の λ が対数正規分布

$$f(\lambda) = \frac{1}{\lambda\sqrt{2\pi V}} \exp(-(\log\lambda - M)^2/2V), \quad \lambda > 0,$$

ただし $-\infty < M < \infty, 0 < V$, に従うとみなす。すなわち $\log\lambda$ は平均 M、分散 V の正規分布に従う。対数正規分布は代表的な右裾の長い分布であり、様々な観測データの記述に用いられる。なお対数正規分布が良く用いられる事を、中心極限定理により正当化する場合が有る (4.7.1 節の Residual Allocation Model に関する議論を見よ)。ただし本論文の立場では、「L 字型」を生成する構造には興味が無い。対数正規分布については、Johnson et al.[88] の Chap. 14、Crow and Shimizu[40]、Aitchison and Brown[3] 等を参照の事。

結局混合された分布は

$$\mathrm{P}(F=y) = \frac{1}{y!\sqrt{2\pi V}} \int_0^{\infty} \lambda^{y-1} \exp(-\lambda - (\log\lambda - M)^2/2V) d\lambda, \quad y = 0.1.2.\ldots. \qquad (4.17)$$

となり、解析的には都合がよくない。応用の際は数値積分、または数表が必要である。例えばAitchison and Ho[5] では、エルミート積分を用いて数値評価をしている。数表については Grundy[66], Brown and Holgate[22] を参照の事。また Bulmer[23] は、$y \geq 10$ について近似

$$\mathrm{P}(F=y) \approx \frac{1}{y\sqrt{2\pi V}} \exp(-(\log y - M)^2/2V)[1 + \frac{1}{2yV}\{\frac{(\log y - M)^2}{V} + \log y - M - 1\}] \quad (4.18)$$

が成立すると主張している。ただし Reid[142] は自分のデータに関し、この近似に否定的である。なお本節の議論は、Shaban[161] に大部分含まれている。

定義 対数正規＝ポアソンモデルの確率関数は、

$$\mathrm{P}(F_1 = y_1, \ldots, F_J = y_J) = \prod_{j=1}^{J} \frac{1}{y_j!\sqrt{2\pi V}} \int_0^\infty \lambda^{y_j - 1} \exp(-\lambda - (\log\lambda - M)^2/2V) d\lambda \quad (4.19)$$

で与えられる。ただし $V > 0, \infty > M > -\infty$ とする。寸法指標について表せば

$$\mathrm{P}(S_0, \ldots) = J! \prod_{i=0}^{\infty} \{\frac{1}{i!\sqrt{2\pi V}} \int_0^\infty \lambda^{i-1} \exp(-\lambda - (\log\lambda - M)^2/2V)\}^{S_i} \frac{1}{S_i!}$$

となる。本モデルでは、母集団サイズが確率変数となる。従って期待値制約 $\mathrm{E}(N) = N_0$ を満たすため、本稿では

$$M = \log N_0 - \log J - V/2 \quad (4.20)$$

とする。

対数正規＝ポアソンモデルは、S_i のモードが $i \geq 2$ にあるようなデータも記述出来る。従って「L字型」を外れる場合にも適用可能である。以下では各 F_j が独立なモデルのみ考察するが、Aitchison and Ho[5] は多変数対数正規分布と複数のポアソン分布の混合を考える事で、モデルに相関を導入している。

対数正規＝ポアソン分布は、特に生態学分野で良く使われる。例えば Preston[138] は、種と個体数の関係を記述するのに、(切り落とし) 対数正規分布がふさわしいと主張する。対数正規＝ポアソン分布の事を Anscombe[6] 等は、「離散対数正規分布」と呼ぶ。ただし Cassie[30] は、離散的な生物データに対数正規曲線をあてはめて、これを「離散対数正規分布」と呼んでいる。また Anscombe は対数正規＝ポアソン分布が単峰格子になると予想したが、Holgate[78] によって証明された (命題 3.3)。研究史の解説は、Aitchison and Ho[5] を見よ。

なお Thorin[203] によれば、対数正規分布は無限分解可能である。従って 3.1 節で紹介した Maceda の議論から、対数正規＝ポアソン分布も無限分解可能と結論される。しかし確率母関数の初等的な表記は知られていない。

4.3.2 モメント

階乗モメントは Bulmer[23] が示している。すなわち非負整数の r について

$$\mathrm{E}(F^{(r)}) = \exp(rM + \frac{1}{2}r^2V) \quad (4.21)$$

である。特に

$$\mathrm{E}(F) = \exp(M + V/2)$$

となる。また

$$\mathrm{V}(F) = \exp(M + V/2) + \exp(2M + 2V) - \exp(2M + V)$$

である。母集団サイズ N の期待値は $J\mathrm{E}(F)$ になるので、制約 (4.20) を確認できる。寸法指標の期待値は (2.3), (4.17) を利用して

$$\mathrm{E}(S_i) = J\frac{1}{i!\sqrt{2\pi V}}\int_0^\infty \lambda^{i-1}\exp(-\lambda - (\log\lambda - M)^2/2V)d\lambda$$

と書ける。分散は (2.4),(4.17) から

$$\mathrm{V}(S_i) = \mathrm{E}(S_i) - \mathrm{E}(S_i)^2 + J(J-1)\{\frac{1}{i!\sqrt{2\pi V}}\int_0^\infty \lambda^{i-1}\exp(-\lambda - (\log\lambda - M)^2/2V)d\lambda\}^2$$

となる。異なる F_j 同士が独立な事に注意。

4.3.3 母数推定

次に母数の推定を考察する。母集団サイズが確率変数のモデルについては、抽出率 n_0/N_0 のベルヌーイ抽出を適用するという事であった。この場合命題 2.2 より標本分布は

$$\mathrm{P}(s_0,\ldots) = J!\prod_{i=0}^\infty\{\frac{1}{i!\sqrt{2\pi V}}\int_0^\infty \lambda^{i-1}\exp(-\lambda - (\log\lambda - m)^2/2V)d\lambda\}^{s_i}\frac{1}{s_i!},$$

ただし

$$m = \log n_0 - \log J - V/2 \tag{4.22}$$

となる。

最尤法から考察しよう。対数尤度は

$$L = \sum_{i=0}^\infty s_i\{-\frac{1}{2}\log V + \log(\int_0^\infty \lambda^{i-1}\exp(-\lambda - (\log\lambda - m)^2/2V)d\lambda)\} + \mathrm{const.}$$

のように書ける。

$$\frac{\partial L}{\partial m} = \sum_{i=0}^\infty s_i\{\frac{\mathrm{P}(F=i)'}{\mathrm{P}(F=i)}\} = 0$$

$$\frac{\partial L}{\partial V} = \sum_{i=0}^\infty s_i\{-\frac{1}{2V} + \frac{\mathrm{P}(F=i)' + \frac{1}{2V}\mathrm{P}(F=i)}{\mathrm{P}(F=i)}\} = \sum_{i=0}^\infty s_i\{\frac{\mathrm{P}(F=i)'}{\mathrm{P}(F=i)}\} = 0$$

を数値的に解けば最尤推定値が得られる。ここで $\mathrm{P}(F=i)'$ の評価が必要になるが、Bulmer[23] によれば期待値制約が無いものとして、

$$\frac{\partial \mathrm{P}(F=i)}{\partial m} = i\mathrm{P}(F=i) - (i+1)\mathrm{P}(F=i+1).$$

$$\frac{\partial \mathrm{P}(F=i)}{\partial V} = \frac{1}{2}\{i^2\mathrm{P}(F=i) - (i+1)(2i+1)\mathrm{P}(F=i+1) + (i+1)(i+2)\mathrm{P}(F=i+2)\}.$$

何故なら

$$\begin{aligned}\frac{\partial \mathrm{P}(F=i)}{\partial m} &= \frac{1}{i!\sqrt{2\pi V}}\int_0^\infty -\lambda^i \frac{\partial \exp(-(\log\lambda - m)^2/2V)}{\partial\lambda}\exp(-\lambda)d\lambda \\ &= \frac{1}{i!\sqrt{2\pi V}}\{[-\exp(-(\log\lambda-m)^2/2V)\lambda^i\exp(-\lambda)]_0^\infty \\ &\quad + \int_0^\infty \exp(-\lambda-(\log\lambda-m)^2/2V)(i\lambda^{i-1}-\lambda^i)d\lambda\} \\ &= i\mathrm{P}(F=i)-(i+1)\mathrm{P}(F=i+1),\end{aligned}$$

$$\frac{\partial \mathrm{P}(F=i)}{\partial V} = \frac{1}{2}\frac{\partial^2 \mathrm{P}(F=i)}{\partial m^2}.$$

いずれにしても数値積分は避けられない。なお期待値制約の下ではラグランジュ乗数法を用いれば、数値解が得られるだろう。すなわち連立方程式

$$\frac{\partial L}{\partial m}+\lambda(-1)=0,$$

$$\frac{\partial L}{\partial V}+\lambda(-\frac{1}{2})=0$$

および (4.22) を解く。

非復元単純無作為抽出の場合はどうだろうか。問題は標本サイズ n 所与の尤度だが、$\mathrm{P}(n)$ は解析的に捉えがたい。故に例えば Hoshino[80] では、n について正規近似を用いている。すなわち n の期待値 n_0、分散 $T=J(\exp(m+V/2)+\exp(2m+2V)-\exp(2m+V))$ から $\mathrm{P}(n=n_0)=1/\sqrt{2\pi T}$ と評価する。この場合

$$\mathrm{P}(s_0,\ldots|n=n_0)\approx J!\prod_{i=0}^{n_0}\{\frac{1}{i!\sqrt{2\pi V}}\int_0^\infty \lambda^{i-1}\exp(-\lambda-(\log\lambda-m)^2/2V)d\lambda\}^{s_i}\frac{1}{s_i!}\sqrt{2\pi T}$$

と考える。ここから期待値制約の下で Hoshino は、Vetterling et al.[210] の一変数関数極大化アルゴリズムを用いて最尤解を求めた。

次にモメント法を考察する。命題 2.2 と (4.21) より全ての j について

$$\mathrm{E}(f_j^2-f_j)=\exp(2\log n-2\log J+V),$$

ただし f_j は標本での壷内ボール数である。f に関する標本分散 (4.8) を利用して、

$$\log(v-n/J)=2\log n-2\log J+V$$

を解けば推定量

$$\hat{V}_{Moment}=\log(v-n/J)-2\log n+2\log J$$

を得る。

4.4 一般化逆ガウシアン＝ポアソンモデル

4.4.1 定義等

一般化逆ガウシアン＝ポアソン（Generalized Inverse Gaussian=Poisson, GIGP）分布では、(3.20) 式の λ が γ 次一般化逆ガウシアン (GIG) 分布

$$f(\lambda;\gamma,\alpha,\theta)=\frac{(2\sqrt{1-\theta}/(\alpha\theta))^{\gamma}}{2K_{\gamma}(\alpha\sqrt{1-\theta})}\lambda^{-1+\gamma}\exp(-(\frac{1}{\theta}-1)\lambda-\frac{\alpha^2\theta}{4\lambda}),\quad \lambda>0, \tag{4.23}$$

に従う。但し $0<\theta<1,\alpha>0$ で、$K_{\gamma}(\cdot)$ は γ 次第三種変形ベッセル関数

$$K_{\gamma}(\kappa)=\frac{\pi}{2}\frac{I_{-\gamma}(\kappa)-I_{\gamma}(\kappa)}{\sin n\pi}, \tag{4.24}$$

ただし

$$I_{\gamma}(\kappa)=\sum_{r=0}^{\infty}\frac{1}{r!\Gamma(n+r+1)}(\frac{\kappa}{2})^{2r+n}$$

のように定義される。$K_{\gamma}(\cdot)$ は、文献によっては「第二種変形ベッセル関数」と呼ばれる。(変形) ベッセル関数に関する結果については、Watson[213], Jørgensen[91] の Appendix, Abramowitz and Stegun[1] の Chap. 9 等を見よ。特に漸化式

$$K_{\gamma+1}(\kappa)=\frac{2\gamma}{\kappa}K_{\gamma}(z)+K_{\gamma-1}(\kappa) \tag{4.25}$$

は数値計算の際、便利である。GIG 分布はピアーソン・タイプ 3 分布とピアーソン・タイプ 5 分布の中間として、Good[61] が導入した。ただし Good は、解析的に扱いがたいと指摘したのみである。Jørgensen[91] が GIG 分布に関する詳しいレビューを与えている。なお Jørgensen[91] では GIG 分布を対数正規分布で近似している。この事から示唆されるように、GIG 分布は「L 字」の記述に適する。

Johnson et al.[88](p.284) によれば、もし $\gamma>0$ ならば $\alpha\to 0$ の時 (4.23) はガンマ分布（ピアーソン・タイプ 3）

$$f(\lambda)=\frac{((1-\theta)/\theta)^{\gamma}}{\Gamma(\gamma)}\lambda^{\gamma-1}\exp(-\frac{1-\theta}{\theta}\lambda)$$

になる。$\gamma<0$ ならば、$\theta\to 1$ の時レシプロカル・ガンマ分布（ピアーソン・タイプ 5、またはインバーティド・ガンマ）

$$f(\lambda)=\frac{\lambda^{-1+\gamma}}{\Gamma(-\gamma)}(\frac{4}{\alpha^2})^{\gamma}\exp(-\frac{\alpha^2}{4\lambda})$$

になる。$\gamma=0$ の場合、双曲線分布と言う。特に (4.23) 式で $\gamma=-1/2$ とした時を逆ガウシアン (Inverse Gaussian, IG) 分布と言う。詳しくは 4.5 節を参照のこと。

混合の結果である GIGP 分布は

$$\mathrm{P}_{GIGP}(F=y;\gamma,\alpha,\theta)=\frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^{y}}{K_{\gamma}(\alpha\sqrt{1-\theta})y!}K_{y+\gamma}(\alpha),\quad y=0,1,2,\ldots, \tag{4.26}$$

ただし $0<\alpha,0<\theta<1,-\infty<\gamma<\infty$、となる (Sichel[171])。特に $\gamma=-1/2$ の時を逆ガウシアン＝ポアソン (IGP) 分布と呼び、次節で扱う。一般の γ、すなわち (4.26) について漸化式

$$\mathrm{P}(F=y)=\theta(\frac{y+\gamma-1}{y})\mathrm{P}(F=y-1)+\frac{(\alpha\theta)^2}{4\gamma(\gamma-1)}\mathrm{P}(F=y-2)$$

が成立する（Sichel）。

定義 一般化逆ガウシアン＝ポアソンモデルは

$$\mathrm{P}_{GIGP}(F_1 = y_1, \ldots, F_J = y_J) = \prod_{j=1}^{J} \frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^{y_j}}{K_\gamma(\alpha\sqrt{1-\theta})y_j!} K_{y_j+\gamma}(\alpha) \tag{4.27}$$

で定義される。ここで $0 < \alpha, 0 < \theta < 1, -\infty < \gamma < \infty$ である。なお母集団サイズは確率変数となる。寸法指標について表せば

$$\mathrm{P}_{GIGP}(S_0, \ldots) = J! \prod_{i=0}^{\infty} \{\frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^i}{K_\gamma(\alpha\sqrt{1-\theta})i!} K_{i+\gamma}(\alpha)\}^{S_i} \frac{1}{S_i!} \tag{4.28}$$

である。期待値制約 $\mathrm{E}(N) = N_0$ を満たすため、本稿では

$$N_0 = J\frac{\alpha\theta}{2\sqrt{1-\theta}} \frac{K_{\gamma+1}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} \tag{4.29}$$

とする。

GIGP 分布については Johnson et al.[90] の 11.15 節を見よ。IGP 分布が比較的簡単に数値評価出来るのに対し、一般の GIGP 分布の計算は厄介である。Atkinson and Yeh[8] は GIGP 分布で $\gamma = -3/2, -1/2, 1/2$ について尤度を評価、三点を通る唯一の放物線で尤度を近似する事を提案した。なお GIGP 分布は「Sichel 分布」とも呼ばれ、Stein et al.[192] は多変数 Sichel 分布を考察している。

例えば Willmot[223] が、GIGP 分布の確率母関数を示している。本稿のパラメトライゼーションに合わせると

$$G(z) = (\frac{1-\theta}{1-\theta z})^{\frac{\gamma}{2}} \frac{K_\gamma(\alpha\sqrt{1-z\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} \tag{4.30}$$

となる。なお

$$\eta(\theta) = (\frac{1}{\sqrt{1-\theta}})^\gamma K_\gamma(\alpha\sqrt{1-\theta})$$

とおけば、$G(z) = \eta(\theta z)/\eta(\theta)$ である。これは GIGP 分布が（α, γ を固定すれば）ベキ級数分布である事を意味する。さらに

$$G(z) = \exp(\log\eta(\theta)(\frac{\log\eta(\theta z)}{\log\eta(\theta)} - 1))$$

と変形できる。ここで

$$g(z) = \frac{\log\eta(\theta z)}{\log\eta(\theta)}$$

もベキ級数分布の確率母関数になっている。故に GIGP 分布も非負整数上の無限分解可能分布である事が確認された（3 章における Lévy の定理に関する議論を参照の事）。ここで N の分布の確率母関数は

$$G(z)^J = (\frac{1-\theta}{1-\theta z})^{\frac{J\gamma}{2}} \left(\frac{K_\gamma(\alpha\sqrt{1-z\theta})}{K_\gamma(\alpha\sqrt{1-\theta})}\right)^J$$

となってベキ級数分布であるが、GIGP 分布ではない。すなわち、GIGP 分布は畳み込みに関して閉じていない。$J > 1$ の場合、N の確率関数の簡潔な表現は明らかでない。

なお Hoshino [82] によれば、$-1 < \gamma < 0$ の時 $J\alpha^{-2\gamma}$ を固定して $J \to \infty (\alpha \to 0)$ という小数法則を GIGP モデル (4.27) に適用すると、拡張負の二項モデル (4.53) が得られる。また γ が正の場合は、$\alpha \to 0$ とすれば GIG 分布がガンマ分布になるので、GIGP 分布モデルはガンマ＝ポアソン（負の二項分布）モデルになる。故に $\gamma > 0, \alpha \to 0$ の場合、更に小数法則を適用した極限 $(J \to \infty, \gamma \to 0)$ は対数級数モデルである。

GIGP 分布は負の二項分布を特殊ケースとして含む事からも分かるように、S_i のモードが $i \neq 1$ となるデータも記述出来る。

4.4.2 モメント

命題 3.2 から GIGP 分布の階乗モメントは、GIG 分布の r 次モメントを見れば良い。Willmot による確率母関数 (4.30) の導出もこの関係を利用している。Jørgensen[91] によれば、非負整数の r について

$$\mathrm{E}(F^{(r)}) = (\frac{\alpha\theta}{2\sqrt{1-\theta}})^r \frac{K_{\gamma+r}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})}.$$

特に $r = 1$ の場合を Sichel[173] が考察しており、

$$\mathrm{E}(F_{GIGP}) = \frac{\alpha\theta}{2\sqrt{1-\theta}} \frac{K_{\gamma+1}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})},$$

$$\mathrm{V}(F_{GIGP}) = \frac{(\alpha\theta)^2}{4(1-\theta)} \frac{K_{\gamma+2}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} + \mathrm{E}(F_{GIGP})(1 - \mathrm{E}(F_{GIGP})).$$

寸法指標の期待値は (2.3) より

$$\mathrm{E}_{GIGP}(S_i) = J\frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^i}{K_\gamma(\alpha\sqrt{1-\theta})i!} K_{i+\gamma}(\alpha),$$

分散も各壷独立な事に注意して、形式的には (2.4) を用いて評価出来る。すなわち

$$\mathrm{V}_{GIGP}(S_i) = \mathrm{E}(S_i) - \mathrm{E}(S_i)^2 + J(J-1)\{\frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^i}{K_\gamma(\alpha\sqrt{1-\theta})i!} K_{i+\gamma}(\alpha)\}^2.$$

4.4.3 母数推定

次に標本分布を考察する。母集団サイズが確率変数なので、抽出率 n_0/N_0 のベルヌーイ抽出を考えよう。この時 (4.28) と命題 2.2 より、GIGP モデルの標本分布は以下のように書ける。

$$\mathrm{P}_{GIGP}(s_0, \ldots) = J! \prod_{i=0}^{\infty} \{\frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^i}{K_\gamma(\alpha\sqrt{1-\theta})i!} K_{i+\gamma}(\alpha)\}^{s_i} \frac{1}{s_i!}.$$

期待値制約 (4.29) については

$$n_0 = J\frac{\alpha\theta}{2\sqrt{1-\theta}} \frac{K_{\gamma+1}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})}$$

となる。

GIGP モデルの対数尤度は

$$L = \sum_{i=0}^{\infty} s_i\{i\log\xi + (\gamma+i)\log\omega - (\gamma+i)\log\alpha + \log K_{i+\gamma}(\alpha) - \log K_\gamma(\omega)\} + \text{const.}$$

なので、

$$\frac{\partial L}{\partial \alpha} = \sum_{i=0}^{\infty} s_i[\frac{\alpha}{\sqrt{\xi^2+\alpha^2}}\{R_\gamma(\omega) + \frac{i}{\omega}\} - R_{i+\gamma}(\alpha)],$$

ただし $R_\gamma(\omega) = K_{\gamma+1}(\omega)/K_\gamma(\omega)$ とする。なおここで

$$\frac{K'_\gamma(\kappa)}{K_\gamma(\kappa)} = -R_\gamma(\kappa) + \frac{\gamma}{\kappa}$$

を用いた。同様に

$$\frac{\partial L}{\partial \xi} = \{\frac{\omega}{\omega+\xi}\}\sum_{i=0}^{\infty} s_i\{\frac{i}{\xi} - R_\gamma(\omega)\},$$

従って尤度方程式

$$\frac{J\alpha}{\sqrt{\xi^2+\alpha^2}}R_\gamma(\omega) + \frac{n}{\omega} - \sum_{i=0}^{\infty} s_i R_{i+\gamma}(\alpha) = 0,$$

$$\frac{n}{\xi} - JR_\gamma(\omega) = 0$$

を解けば、γ 所与の最尤解が得られる。ここで γ 所与での最大対数尤度を、$L^*(\gamma) = \max_{\alpha,\xi} L(\alpha,\xi|\gamma)$ で表す。Stein et al.[192] は、以下のアルゴリズムで三母数の最尤解 $(\gamma^*,\alpha^*,\xi^*)$ が得られる、ただし α^*,ξ^* において $L^*(\gamma^*) = L(\alpha^*,\xi^*|\gamma^*)$、と主張している。

1. 試行錯誤によって、$L^*(\gamma_1) < L^*(\gamma_2) > L^*(\gamma_3)$ となるような $\gamma_1 < \gamma_2 < \gamma_3$ を見つける ($\gamma = -1/2$ は通常良い初期値)。$g_i = L^*(\gamma_i), i = 1,2,3,$ とおく。
2. 三点 $(\gamma_1,g_1),(\gamma_2,g_2),(\gamma_3,g_3)$ を通る二次式の最大値を与えるような γ^* を計算。すなわち

$$\gamma^* = \frac{1}{2}[\frac{(g_1-g_3)(\gamma_2^2-\gamma_3^2) - (g_2-g_3)(\gamma_1^2-\gamma_3^2)}{(g_1-g_3)(\gamma_2-\gamma_3) - (g_2-g_3)(\gamma_1-\gamma_3)}].$$

3. $g^* = L^*(\gamma^*)$ を計算。
4. • もし $\gamma_1 < \gamma^* < \gamma_2$ ならば :
   If $g^* < g_2$, then $\gamma_1 \leftarrow \gamma^*, g_1 \leftarrow g^*$. Else $\gamma_3 \leftarrow \gamma_2, g_3 \leftarrow g_2, \gamma_2 \leftarrow \gamma^*, g_2 \leftarrow g^*$.
   • もし $\gamma_2 < \gamma^* < \gamma_3$ ならば :
   If $g^* < g_2$, then $\gamma_3 \leftarrow \gamma^*, g_3 \leftarrow g^*$. Else $\gamma_1 \leftarrow \gamma_2, g_1 \leftarrow g_2, \gamma_2 \leftarrow \gamma^*, g_2 \leftarrow g^*$.
   • もし $\gamma^* = \gamma_2$ ならば :
   $\gamma^* \leftarrow .99\gamma_2 + .01\gamma_3$; ステップ 3 へ。
5. 収束を確認。もし $\gamma_3 - \gamma_1$ が十分小さければ停止し $\hat{\gamma} = \gamma^*$. Else ステップ 2 へ。

なお計算精度によっては、$(g_2 - \min(g_1,g_3))/\min(g_1,g_3)$ が十分小さいか確認する事が推奨されている。このアルゴリズムは極大を与えるが、それは必ずしも最大と限らないように思える。

期待値制約 (4.29) 付きの最尤推定を考察しよう。パラメトライゼーションを合わせると、制約は

$$\log J + \log \xi + \log K_{\gamma+1}(\omega) - \log K_{\gamma}(\omega) - \log n_0 = 0 \tag{4.31}$$

のように書ける。(4.31) の左辺を T と書けば

$$\frac{\partial T}{\partial \alpha} = \frac{\alpha}{\omega+\xi}\{(R_{\gamma+1}(\omega) - R_{\gamma}(\omega)) + \frac{1}{\omega}\},$$

$$\frac{\partial T}{\partial \xi} = \frac{1}{\xi} + \frac{1}{\omega+\xi}\{\omega(R_{\gamma+1}(\omega) - R_{\gamma}(\omega)) - 1\}$$

となり、ラグランジュ乗数法が使えるだろう。なお

$$\frac{\partial R_{\gamma-1/2}(\kappa)}{\partial \kappa} = R^2_{\gamma-1/2}(\kappa) - \frac{2\gamma}{\kappa} R_{\gamma-1/2}(\kappa) - 1$$

は計算の際、便利である。

次に母集団サイズ N 所与の場合を考える。GIGP モデルについては一般に N の確率が評価出来ていないので、正規近似を用いるのが一つの考えだろう。N は独立な J 個の確率変数の和なので、3.4.2 節の結果を利用して、分散も容易に評価出来る。すなわち

$$\mathrm{V}(N) = J\{\frac{(\alpha\theta)^2}{4(1-\theta)} \frac{K_{\gamma+2}(\alpha\sqrt{1-\theta})}{K_{\gamma}(\alpha\sqrt{1-\theta})} + \frac{\alpha\theta}{2\sqrt{1-\theta}} \frac{K_{\gamma+1}(\alpha\sqrt{1-\theta})}{K_{\gamma}(\alpha\sqrt{1-\theta})} - (\frac{\alpha\theta}{2\sqrt{1-\theta}} \frac{K_{\gamma+1}(\alpha\sqrt{1-\theta})}{K_{\gamma}(\alpha\sqrt{1-\theta})})^2\}.$$

ここで N が期待値制約 (4.29) を満たすとして、$\mathrm{P}(N = N_0) \approx 1/\sqrt{2\pi \mathrm{V}(N)}$ のように考えれば良い。つまり

$$\mathrm{P}(S_0, S_1 \ldots | N = N_0) \approx J! \prod_{i=0}^{\infty} \{\frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^i}{K_{\gamma}(\alpha\sqrt{1-\theta})i!} K_{i+\gamma}(\alpha)\}^{S_i} \frac{1}{S_i!} \sqrt{2\pi \mathrm{V}(N)} \tag{4.32}$$

が近似的条件付 GIGP モデルという事になる。標本分布は命題 2.1 より、(4.32) の母集団に依存する量を標本のそれに置きかえれば良い。ここから最尤解を数値的に評価出来る可能性が有る。

モメント推定については、どうしても変形ベッセル関数の数値的評価を伴うと思われる。だとすれば簡便性が無いため、本稿では考察しない。

4.5 逆ガウシアン＝ポアソンモデル

4.5.1 定義等

逆ガウシアン (IG) 分布は GIG 分布で $\gamma = -1/2$ とおいた特殊ケースである。密度関数は以下のように書ける。

$$f(\lambda; \alpha. \theta) = \frac{(2\sqrt{1-\theta}/(\alpha\theta))^{\frac{-1}{2}}}{2K_{-1/2}(\alpha\sqrt{1-\theta})} \lambda^{-\frac{3}{2}} \exp(-(\frac{1}{\theta} - 1)\lambda - \frac{\alpha^2\theta}{4\lambda}), \quad \lambda > 0. \tag{4.33}$$

但し $0 < \alpha. 0 < \theta \leq 1$ で $K_{-1/2}(\cdot)$ は $-1/2$ 次第三種変形ベッセル関数である。

$$K_{-1/2}(\kappa) = \sqrt{\frac{\pi}{2}} \kappa^{-1/2} \exp(-\kappa) = K_{1/2}(\kappa) \tag{4.34}$$

が成立するので、(4.33) は

$$f(\lambda)=\frac{a\sqrt{\theta}}{2\sqrt{\pi}}\lambda^{\frac{-3}{2}}\exp(\frac{-1}{2}(\frac{a^2\theta}{2}\lambda^{-1}+(\frac{2}{\theta}-2)\lambda)+a\sqrt{1-\theta}) \tag{4.35}$$

とも書ける。

GIG 分布の場合、母数 $\theta=1$ の時に分布が退化する。しかし IG 分布については $\theta=1$ と出来る。つまり GIG 分布は $\gamma=-1/2$ の時のみ、$\theta=1$ を許す。なお GIG 分布について γ が負なら、$\theta\to 1$ の時はレシプロカル・ガンマ分布になると述べた。(4.33) は $\theta=1$ の時、

$$f(\lambda)=\frac{\alpha}{2\sqrt{\pi}}\lambda^{-3/2}\exp(-\frac{\alpha^2}{4\lambda})$$

と書ける。これはレシプロカル・ガンマ分布の密度だが、指数 1/2 の安定分布になっている。なお指数 1/2 の安定分布は、Lévy 分布とも呼ばれる（例えば Uchaikin and Zolotarev[207] を見よ)。本稿では IGP 分布のベキ級数分布としての性質を用いるため、Sichel のパラメトライゼーション (4.33) を用いる。IG 分布の他のパラメトライゼーションについては Johnson et al.[88] (p.261) 等を見よ。なお IG 分布については、Seshadri[159] のモノグラフが応用も含めて網羅的である。また Seshadri[158] は、指数族という視点から IG 分布について議論している。Folks and Chhikara[57] のレビュー以来、IG 分布は右に歪んだデータを記述する分布としてかなり研究されている（生存時間解析、信頼性評価等)。右に歪んだ分布としては対数正規分布が代表的だが、Folks and Chhikara によれば、IG 分布は対数正規分布よりも挙動をとらえやすく、代替・近似的に使用できる。なお IG 分布の近似については、Mudholkar and Natarajan[121] による、渋谷 [165] の近似を含む各種手法の総合的比較を参照の事。IG 分布が対数正規分布に比べて都合が良い最大の理由は、以下で示されるように混合ポアソン分布が解析的に扱いやすい事である。

IG 分布でポアソン分布を混合する、すなわち (3.20) 式の λ が IG 分布 (4.33) に従う場合、$0<\alpha, 0<\theta\leq 1$ について逆ガウシアン＝ポアソン (IGP) 分布

$$\mathrm{P}_{IGP}(F=y;\alpha,\theta)=\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^y}{y!}K_{y-1/2}(\alpha),\quad y=0,1,2,\ldots, \tag{4.36}$$

を得る (Holla[79])。なお漸化式 (4.25) と初期値 (4.34) より

$$K_{y-1/2}(\alpha)=\sqrt{\frac{\pi}{2\alpha}}\exp(-\alpha)(\sum_{i=0}^{y-1}\frac{(y-1+i)!}{(y-1-i)!i!}(2\alpha)^{-i}),\quad y=1,2,\ldots,$$

が示される。やはり GIGP 分布では $\theta\neq 1$ だった事に注意。(4.36) は $\theta=1$ の時、指数 1/2 の離散安定分布 (Steutel and van Harn[195]) である。ところが安定分布の性質から類推されるように、この時 $\mathrm{E}(F)$ は無限となる。N の期待値が有限でないモデルは、寸法指標の推測には使いづらい。従ってモデルにおいては $\theta\neq 1$ とし、期待値制約が入れられるようにする。

定義 GIGP モデル (4.27) において $\gamma=-1/2$ の時を、逆ガウシアン＝ポアソンモデルと呼ぶ。頻度の同時確率は

$$\mathrm{P}_{IGP}(F_1=y_1,\ldots,F_J=y_J)=\prod_{j=1}^{J}\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^{y_j}}{y_j!}K_{y_j-1/2}(\alpha) \tag{4.37}$$

となる。ただし $0<\alpha, 0<\theta<1$ である。寸法指標については

$$\mathrm{P}_{IGP}(S_0,\ldots)=J!\prod_{i=0}^{\infty}\{\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^i}{i!}K_{i-1/2}(\alpha)\}^{S_i}\frac{1}{S_i!}$$

となる。期待値制約については、GIGP モデルの制約 (4.29) が単純化され

$$\alpha=\frac{2N_0\sqrt{1-\theta}}{J\theta}$$

と書ける。

IGP 分布については、Seshadri[159] の 7.1 節の議論が詳しい（または Johnson et al.[90] の 11.15 節を見よ)。Shaban[160] は IGP 分布と、その極限（レシプロカル・ガンマ＝ポアソン分布およびポアソン分布）について、形状の差異を数値的に評価、近似として差は小さくないと結論している。しかし Ong[127] によれば、レシプロカル・ガンマ＝ポアソン分布に適当な修正項を付け加える事で、近似の改善がみこめると言う。Shaban[161] では、対数正規＝ポアソン分布、Bulmer の近似 (4.18)、Paul and Plackett の近似 (3.6) と IG ＝ポアソン分布が数値的に比較されており、IGP 分布は対数正規＝ポアソン分布と比較的近い場合が多いという。また Tweedie[206] によれば、IG 分布は正の単峰絶対連続な密度関数を持つ。故に命題 3.3 より、IGP 分布は単峰格子変数である。

Sankaran[152] が IGP 分布の確率母関数を導出している。(3.5) 式において混合する分布を IG 分布 (4.35) とすれば、確率母関数

$$G(z)=\exp(\alpha(\sqrt{1-\theta}-\sqrt{1-z\theta})), \tag{4.38}$$

ただし $0<\alpha, 0<\theta\leq 1$ を得る。確率の総和が 1 となる事を前提とすれば $G(z)$ の導出は容易だが、直接的に評価するには Spiegel[191] 等を参照の事。特性関数については Sichel[173] を見よ。もし $\theta\neq 1$ ならば

$$G(z)=\exp(-\alpha\sqrt{1-\theta}((g(z)-1)),$$

ただし

$$g(z)=\frac{\sqrt{1-z\theta}}{\sqrt{1-\theta}}$$

と書き直す事が出来る。ここで $g(z)$ は、拡張負の二項分布 (4.51) の $\gamma=-1/2$ の場合の確率母関数である。以上の議論から、IGP 分布は非負整数上の無限分解可能分布である事が分かる。従って、小数法則の適用で IGP モデルから意味のある極限分布（$\gamma=-1/2$ の拡張負の二項モデル (4.53)）が得られる。4.4 節と 4.12 節の議論も参照のこと。

ここで $\alpha\theta=c$ を固定して $\alpha\to\infty, \theta\to 0$ という極限を考える。実はこの時 IGP 分布の確率母関数 (4.38) は、平均 $c/2$ のポアソン分布の確率母関数 $\exp(c(z-1)/2)$ と一致する。何故なら、

$$\begin{aligned}\alpha(\sqrt{1-\theta}-\sqrt{1-z\theta}) &= \alpha(\sqrt{1-\theta}-\sqrt{1-z\theta})\frac{(\sqrt{1-\theta}+\sqrt{1-z\theta})}{(\sqrt{1-\theta}+\sqrt{1-z\theta})}\\ &= \frac{\alpha\theta(z-1)}{\sqrt{1-\theta}+\sqrt{1-z\theta}}\\ &\to \frac{c(z-1)}{2}.\end{aligned}$$

また確率母関数 (4.38) の形状から、IGP 分布がが畳み込みに関して閉じている事が分る。故に母集団サイズ N の分布は、少なくとも IGP 分布の場合、明示的に評価出来る。すなわち

$$\mathrm{P}_{IGP}(N=y)=\sqrt{\frac{2J\alpha}{\pi}}\exp(J\alpha\sqrt{1-\theta})\frac{(J\alpha\theta/2)^y}{y!}K_{y-1/2}(J\alpha),\quad y=0,1,2,\ldots. \tag{4.39}$$

のように、(4.36) 式で母数 α を $J\alpha$ で置き換えれば良い。

4.5.2 モメント

IGP 分布の r 次階乗モメントは IG 分布の r 次モメントに等しい。GIGP 分布に関する議論の特殊ケースとして、

$$\mathrm{E}(F_{IGP})=\frac{\alpha\theta}{2\sqrt{1-\theta}},$$
$$\mathrm{V}(F_{IGP})=\frac{\alpha\theta(2-\theta)}{4(1-\theta)^{\frac{3}{2}}}.$$

を得る。もちろん (4.38) から求める事も出来る。

寸法指標の期待値は GIGP モデルについての議論より

$$\mathrm{E}_{IGP}(S_i)=J\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^i}{i!}K_{i-1/2}(\alpha),$$

$$\mathrm{V}_{IGP}(S_i)=\mathrm{E}(S_i)-\mathrm{E}(S_i)^2+J(J-1)\{\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^i}{i!}K_{i-1/2}(\alpha)\}^2$$

である。

4.5.3 母数推定

母数推定もやはり、GIGP モデルに関する議論の特殊ケースである。抽出率 n_0/N_0 のベルヌーイ抽出を仮定すると、IGP モデルの標本分布は

$$\mathrm{P}_{IGP}(s_0,\ldots)=J!\prod_{i=0}^{\infty}\{\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^i}{i!}K_{i-1/2}(\alpha)\}^{s_i}\frac{1}{s_i!}.$$

ただし期待値制約が $\alpha=2n_0\sqrt{1-\theta}/(J\theta)$ のようになる。

IGP 分布に関する最尤推定については、例えば Seshadri[159] で説明されている。Sichel[175] は、最尤推定量 $(\hat{\alpha},\hat{\theta})$ の相関を評価した。そして通常の応用においてこれらは強い負の相関を持ちうる為、別のパラメトライゼーションが妥当かもしれないと述べている。Stein et al.[192] はこれを受けて、数値解の不安定性を避けるため変数変換

$$\xi=\frac{\alpha\theta}{2\sqrt{1-\theta}}$$

を推奨している。この時 $\omega=(\xi^2+\alpha^2)^{1/2}-\xi=\alpha\sqrt{1-\theta}$ とおけば

$$\mathrm{P}(F=y)=\frac{(\omega/\alpha)^\gamma(\xi\omega/\alpha)^y K_{y+\gamma}(\alpha)}{y!K_\gamma(\omega)}$$

と書ける。ここでξは$\gamma = -1/2$の時、平均に比例する母数である。なおIGPモデルではNがθの十分統計量となる。従って3.8節で議論されるIGPモデルのN所与の条件付分布では、θが消える。しかし条件付IGP分布は、実質的にIGP分布の二母数の情報を持つ。故にIGP分布の数値的不安定性を避けるには、条件付IGP分布を用いると良い(Hoshino[83])。

Sichel[175]は$\theta = 0.97$としてシミュレーションでIGP分布の推定量の効率を評価した。なお現実のデータあてはめでは、多くの場合$\theta > .9$になると主張されている(これは右裾がかなり長い事を意味する)。シミュレーションの結果が示唆する事は以下の通りである。すなわち$\alpha \doteq 0$の時(つまり頻度0の割合が多い場合)モメント法の効率は悪い。標本数が非常に多い場合は別にして、モメント法は変動係数が35%よりも小さい時(正規分布に近い)のみ望ましい。

標本分散(4.8)を用いたモメント推定量は、以下の通りになる。

$$\hat{\theta} = 1 - \frac{1}{2vJ/n - 1}, \quad \hat{\alpha} = \frac{2n/J\sqrt{1-\hat{\theta}}}{\hat{\theta}}. \tag{4.40}$$

Anscombe[6]によれば、観測された頻度0の標本割合が寸法指標の推定の効率性に大きく影響するという。Sichel[172]は、この議論を受けたIGP分布の推定量を提案した。これを利用すれば

$$\hat{\theta} = 1 - (\frac{-\log\hat{\phi_0}}{2n/J + \log\hat{\phi_0}})^2, \quad \hat{\alpha} = -\frac{1}{2}\log\hat{\phi_0}(1 + \frac{n/J}{n/J + \log\hat{\phi_0}}), \tag{4.41}$$

ただし$\hat{\phi_0}$は観測された頻度0の標本割合s_0/Jである。

4.6 対数級数モデル

4.6.1 定義等

対数級数分布のアイディアは、Fisher et al.[56]の有名な論文に遡る。この論文は大きな影響力を持ったが、一意にとどまらない解釈を許す事でも知られている(例えばJohnson et al.[90]の7章、Watterson[215]を見よ)。通常「対数級数分布」といえば、Holgate[77]のような解釈をする。すなわち負の二項分布(4.10)から度数0を切り落とした

$$\mathrm{P}(F = y) = (\frac{\Gamma(y+\gamma)}{\Gamma(\gamma)y!}p^\gamma q^y)/(1-p^\gamma), \quad y = 1, 2, \ldots,$$

について、$\gamma \to 0$の極限を適用した結果

$$\mathrm{P}(F = y) = \frac{cq^y}{y}, \quad y = 1, 2, \ldots, \tag{4.42}$$

ただし$c = -1/\log(1-q)$、をさす。(なお切り落とし負の二項分布については、Sampford[150], Rider[143]等を見よ。)しかしこの分布は$F = 0$の確率を定義していない為、自明な方法で空のセルを許す超母集団モデルを構成できない。

我々はこれと区別して、Anscombe[6]の解釈によるモデルを「対数級数モデル」と呼ぶ。すなわち対数級数モデルはガンマ=ポアソンモデル(4.12)において$J\gamma = 1/\beta$を固定し、$J \to \infty, \gamma \to 0$という極限操作(小数法則)の結果として得られる。3.2節で考察したように、このモデルは無限

大個の各セルで、度数 F_j の 1 以上という条件付分布が対数級数分布に従っていると解釈出来る。なお対数級数モデルについて N 所与とした場合、Ewens モデル (4.44) を得る。逆に Ewens モデルの母集団サイズに負の二項分布 (4.14) を混合すれば、対数級数モデルを得る。本節の議論は、Hoshino and Takemura[85] に詳しい。

定義 $N_0 > 0,\ 0 < p < 1, q = 1 - p$ について、$p = 1/(N_0\beta + 1)$、

$$\lambda_i = N_0 \frac{p \cdot q^{i-1}}{i}, \quad i = 1, 2, \ldots,$$

とする。この時母集団サイズが確率変数である対数級数モデルは

$$\mathrm{P}(S_1, S_2, \ldots) = \prod_{i=1}^{\infty} \frac{\lambda_i^{S_i} \exp(-\lambda_i)}{S_i!} \tag{4.43}$$

のように定義される。この時期待値制約 $\mathrm{E}(N) = N_0$ を満たす。

対数級数モデルにおいて、$S_i, i = 1, 2, \ldots,$ は各 i 独立に平均 λ_i のポアソン分布に従う。母集団サイズ $N = \sum_{i=1}^{\infty} i\, S_i$ は確率変数であり、負の二項分布 (4.14) に従う（Hoshino and Takemura[85] を見よ)。期待値は以下の様にも評価出来る。

$$\mathrm{E}(N) = \sum_{i=1}^{\infty} i \cdot \mathrm{E}(S_i) = \sum_{i=1}^{\infty} i\lambda_i = N_0 p \sum_{i=1}^{\infty} i \cdot \frac{q^{i-1}}{i} = N_0.$$

4.6.2 モメント

対数級数モデルでは、S_i のモメントはポアソン分布のそれである。従って良く知られているように、階乗モメントは非負整数の r について

$$\mathrm{E}(S_i^{(r)}) = {\lambda_i}^r, \quad i = 1, 2, \ldots,$$

のように書ける。Johnson et al.[90] の 4 章を見よ。なお各 i 独立であり、特に

$$\mathrm{E}(S_i) = \lambda_i,$$

$$\mathrm{V}(S_i) = \lambda_i.$$

4.6.3 母数推測

まず抽出率 n_0/N_0 のベルヌーイ抽出を考えよう。命題 2.2 の証明で確認したように、ポアソン分布からのベルヌーイ標本はポアソン分布に従う。従って対数級数モデルの標本分布も対数級数モデルであり、母集団分布の N_0 を n_0 に置き換えればよい。すなわち、以下の標本分布を得る。

$$\mathrm{P}(s_1, s_2, \ldots) = \prod_{i=1}^{\infty} \frac{\tilde{\lambda}_i^{s_i} \exp(-\tilde{\lambda}_i)}{s_i!},$$

ただし

$$\tilde{\lambda}_i = \frac{\beta^{i-1}}{i}(\frac{n_0}{n_0\beta+1})^i$$

である。

従って対数尤度は

$$L = (n-u)\log\beta - (n+1/\beta)\log(n_0\beta+1) + \text{const.}$$

応用上は $n = n_0$ とみなす事に注意。この時 L を β で微分して、最尤方程式を得る。すなわち最尤推定量 $\hat{\beta}$ は、

$$-u\beta + \log(n_0\beta+1) = 0$$

の解を数値的に評価すれば良い。上式の左辺をもう一度微分して

$$-u + \frac{n_0}{n_0\beta+1}$$

が得られるので、ニュートン＝ラフソン法等が使える。

モメント推定量の導出は、例えば

$$\frac{\mathrm{E}(s_2)}{\mathrm{E}(s_1)} = \frac{n_0\beta}{2(n_0\beta+1)}$$

という事から、

$$\hat{\beta} = \frac{2s_2}{n(s_1-2s_2)}$$

等が考えられる。

単純無作為抽出の場合は n 所与の分布、すなわち Ewens モデルの議論に従う。

4.7 Ewens モデル

4.7.1 定義等

Ewens 分布は、"Multivariate Ewens Distribution" や "Ewens Sampling Formula" とも呼ばれる。そもそも Ewens[51] では、遺伝子の複製に関して突然変異が無作為に起きるモデルとして提唱された。壷モデルとしての解釈が可能であり、多様な応用がなされている。本分布については、Johnson et al.[89] の 41 章で詳しく解説されている。

定義 母数 $\theta > 0$ について、母集団サイズ N が固定されたモデル

$$\mathrm{P}(S_1,\ldots,S_N) = \frac{\theta^U}{\theta^{[N]}}\frac{N!}{\prod_{i=1}^N i^{S_i}S_i!}, \tag{4.44}$$

ただし $\theta^{[N]} = \theta(\theta+1)(\theta+2)\cdots(\theta+N-1)$、を Ewens モデルとする。

Ewens 分布は、後に取り上げる Pitman 分布 (4.46) で母数 α が 0 の特殊ケースである。またディリクレ＝多項モデル (4.2) で $J\gamma = \theta$ を固定し、$J \to \infty.\gamma \to 0$ の極限を適用すると、Ewens 分布となる。もしくは対数級数モデル (4.43) において、母集団サイズ所与の条件付モデルである。関連して次の命題が成立する。

命題 4.2 (Sibuya[166], Proposition 2.2) m を固定した正の有限な整数として $N \to \infty$ とすれば、*Ewens* 分布 (4.44) の最初の m 個の要素 $(S_1, S_2, \ldots, S_m)$ は独立な平均 $(\theta, \theta/2, \ldots, \theta/m)$ のポアソン分布の同時分布に収束する。

命題 4.2 での極限分布は、対数級数モデル (4.43) でベキ級数母数 q が 1 の場合に近づく。何故なら、N の期待値が大きな対数級数モデルは裾が長く、ベキ母数 q が 1 となる事に対応する。これと Ewens 分布の極限が一致するのは自然である。この命題と並行し、極限 CIGP 分布について命題 4.4 が得られる。

Ewens 分布は "Residual Allocation Model (RAM)" として構成可能である。$(W_1, W_2, \ldots)$ を $0 \le W_i \le 1, i = 1, 2, \ldots,$ となるような確率変数列とする。$\bar{W}_i = 1 - W_i$ と定義する。

$$P_i = \bar{W}_1 \cdots \bar{W}_{i-1} W_i, \quad i = 1, 2, \ldots. \tag{4.45}$$

と書く。この時 $(P_1, P_2, \ldots)$ は、第 i グループが割合 P_i を持つような確率的分類を構成する。ここで $P_i = (1 - P_1 - P_2 - \cdots - P_{i-1})W_i$、つまり余りを割り付けているので、この過程が独立な W_i によって構成される場合、RAM と呼ばれる。ここで $X_j, j = 1, \ldots, N,$ は $(P_1, P_2, \ldots)$ 所与で、独立かつ同一に $\mathrm{P}(X_j = i | P_1, P_2, \ldots) = P_i, i = 1, 2, \ldots,$ という分布に従うとする。もし W_i が母数 $(1, \theta)$ のベータ分布に従う場合、$X_1, \ldots, X_N$ の寸法指標の分布は、Ewens 分布という事になる。RAM のサーベイについては Pitman[135] を見よ。

RAM では

$$\log P_i = \log \bar{W}_1 + \cdots + \log \bar{W}_{i-1} + \log W_i$$

と書ける。ここで上式の右辺は確率変数の和である。故に適当な正則条件の下で中心極限定理が成立し、$\log P_i$ は正規分布に従う。すなわち P_i は多くの場合、対数正規分布となる。この議論は Halmos[72] によるが、多くの著者が中心極限定理を用いて対数正規分布を正当化している。例えば Brown and Sanders[21] の議論を見よ。更に $F_i = \sum_{i=1}^{N} I(X_j = i)$ の周辺分布

$$\mathrm{P}(F_i = y | P_1, P_2, \ldots) = \binom{N}{y} P_i{}^y (1 - P_i)^{N-y}, y = 0, 1, \ldots, N,$$

は二項分布である。これは平均 NP_i のポアソン分布で近似されるので、$\log(NP_i)$ が中心極限定理により正規分布に従う場合、周辺分布を対数正規＝ポアソン分布で近似するのは自然である。要するに RAM として構成される分布は、対数正規＝ポアソン分布と同様の構造を背後に持つ (Hoshino[80])。

4.7.2 モメント

寸法指標の同時階乗モメントは Sibuya[166] が与えている。すなわち非負整数の $r_i, i = 1, 2, \ldots, N,$ について

$$\mathrm{E}(\prod_{i=1}^{N} S_i{}^{(r_i)}) = \frac{\theta^r \theta^{[N-R]} N^{(R)}}{\theta^{[N]}} \prod_{i=1}^{N} (\frac{1}{i!})^{r_i},$$

ただし $r = r_1 + \cdots + r_N, R = \sum_{i=1}^{N} i r_i \le N$。特に

$$\mathrm{E}(S_i) = \frac{\theta}{i} \prod_{j=1}^{i} \frac{N - j + 1}{\theta + N - j},$$

$$\mathrm{V}(S_i) = \frac{\theta^2\theta^{[N-2i]}N^{(2i)}}{\theta^{[N]}}(\frac{1}{i!})^2 + \frac{\theta}{i}\prod_{j=1}^{i}\frac{N-j+1}{\theta+N-j} - (\frac{\theta}{i}\prod_{j=1}^{i}\frac{N-j+1}{\theta+N-j})^2.$$

4.7.3 母数推測

Ewens 分布は分割構造を持つ事が知られている。また命題 2.1 の前提条件（個体のラベルに依存しない事）も満たされる。従って大きさ N の母集団から n 個の標本を非復元単純無作為抽出する場合、標本分布は

$$\mathrm{P}(s_1,\ldots,s_n) = \frac{\theta^u}{\theta^{[n]}}\frac{n!}{\prod_{i=1}^{n} i^{s_i}s_i!}$$

のようになる。この時対数尤度は以下の様に書ける。

$$L = u\log\theta - \log(\theta+n-1) - \log(\theta+n-2) - \cdots - \log\theta + \mathrm{const.}$$

故に母数 θ の最尤推定量は

$$\frac{u}{\theta} - \sum_{j=1}^{n}\frac{1}{\theta-1+j} = 0$$

の解という事になる。二次の微分係数

$$\frac{\partial^2 L}{\partial\theta^2} = -\frac{u}{\theta^2} + \sum_{j=1}^{n}\frac{1}{(\theta-1+j)^2}$$

を利用してニュートン＝ラフソン法等を用いれば良い。例えば Sibuya[164], Hoshino and Takemura[85] を見よ。なお Sibuya[164] が指摘するように、Ewens 分布は指数分布族に所属するので尤度方程式は一意な解を持つ。また数値解を求めるための反復計算において、良好に収束する。なお u は明らかに母数 θ の十分統計量である。

簡単なモメント推定量を挙げよう。$\mathrm{E}(s_1) = \theta n/(\theta+n-1)$ より

$$\hat{\theta} = \frac{s_1(n-1)}{n-s_1}$$

を導く事が出来る。

4.8 Pitman モデル

4.8.1 定義等

Pitman 分布は、“Pitman Sampling Formula” とも呼ばれる。Pitman[134] が、Ewens 分布を二母数へ拡張したものである。Ewens モデルと同様に RAM として構成可能である。すなわち Pitman[134] の Construction 16 によれば、過程 (4.45) において、W_i が母数 $(1-\alpha, \theta+i\alpha)$ のベータ分布に従う場合、Pitman 分布が導出される。この意味で、対数正規分布と関連を持つ。

定義 母数 $0 \leq \alpha < 1, \theta > -\alpha$ またはある自然数 m について $\alpha < 0, \theta = -m\alpha$ を満たすような組み合わせについて、Pitman モデルは以下の様に定義される。すなわち固定された母集団サイズ N について、

$$\mathrm{P}(S_1, \ldots, S_N) = N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=1}^{N} \left(\frac{(1-\alpha)^{[j-1]}}{j!}\right)^{S_j} \frac{1}{S_j!}. \tag{4.46}$$

ここで $\theta^{[U:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(U-1)\alpha)$, $\theta^{[N]} = \theta(\theta+1)\cdots(\theta+N-1)$, である。

もし α が 0 ならば、(4.46) は Ewens モデルの定義 (4.44) と一致する。また $\alpha < 0$ の場合 $\theta = -J\alpha > 0, \gamma = -\alpha > 0$ とおけば、(4.46) はディリクレ=多項モデル (4.2) となる。故に Pitman モデルとしては、$0 < \alpha < 1, \theta > -\alpha$ の領域が重要である。

4.8.2 モメント

Yamato and Sibuya[227] が、寸法指標の同時階乗モメントを評価している。すなわち非負整数の $r_i, i = 1, 2, \ldots, N$, について

$$\mathrm{E}\left(\prod_{i=1}^{N} {S_i}^{(r_i)}\right) = \frac{\theta^{[r:\alpha]}(\theta+r\alpha)^{[N-R]}N^{(R)}}{\theta^{[N]}} \prod_{i=1}^{N} \left(\frac{(1-\alpha)^{[i-1]}}{i!}\right)^{r_i},$$

ただし $r = r_1 + \cdots + r_N, R = \sum_{i=1}^{N} i r_i \leq N$。特に

$$\mathrm{E}(S_i) = \frac{(1-\alpha)^{[i-1]}N^{(i)}}{i!}\theta\left(\frac{(\theta+\alpha)^{[N-i]}}{\theta^{[N]}}\right),$$

$$\begin{aligned}\mathrm{V}(S_i) &= \left(\frac{(1-\alpha)^{[i-1]}}{i!}\right)^2 \theta^{[2:\alpha]}\left(\frac{(\theta+2\alpha)^{[N-2i]}N^{(2i)}}{\theta^{[N]}}\right) \\ &\quad + \frac{(1-\alpha)^{[i-1]}N^{(i)}}{i!}\theta\left(\frac{(\theta+\alpha)^{[N-i]}}{\theta^{[N]}}\right) \\ &\quad - \left(\frac{(1-\alpha)^{[i-1]}N^{(i)}}{i!}\theta\left(\frac{(\theta+\alpha)^{[N-i]}}{\theta^{[N]}}\right)\right)^2\end{aligned}$$

となる。

Hoshino[80] によれば、$\alpha \neq 0$ の時

$$\mathrm{E}(U) = \frac{\theta}{\alpha}\left(\frac{(\theta+\alpha)^{[N]}}{\theta^{[N]}} - 1\right),$$

と書ける。そして $N \to \infty$ の時 $\Gamma(\theta+\alpha+N-i)/\Gamma(\theta+N) \approx N^{\alpha-i}$ より、$\alpha \geq 0$ ならば $N \to \infty$ の時

$$\mathrm{E}(S_i) \approx \frac{(1-\alpha)^{[i-1]}}{i!}\frac{\Gamma(\theta)}{\Gamma(\theta+\alpha)}N^{\alpha}$$

である。故に

$$\lim_{N\to\infty} \frac{E(S_i)}{E(U)} = \alpha\frac{(1-\alpha)^{[i-1]}}{i!} \tag{4.47}$$

が成立する。

4.8.3 母数推測

Pitman 分布はそもそも分割構造を満たすように構成された（Pitman[134]）。また分布が個体のラベルに依存しない事からも、命題 2.1 が適用出来る事が分かる。すなわち大きさ N の母集団から n 個の標本を非復元単純無作為抽出した場合、標本分布は

$$\mathrm{P}(s_1,\ldots,s_n) = n!\frac{\theta^{[u:\alpha]}}{\theta^{[n]}}\prod_{j=1}^{n}(\frac{(1-\alpha)^{[j-1]}}{j!})^{s_j}\frac{1}{s_j!}$$

のように書ける。

この時、対数尤度は

$$L = \sum_{i=1}^{u-1}\log(\theta+i\alpha) - \sum_{i=1}^{n-1}\log(\theta+i) + s_1 + \sum_{i=2}^{n}s_i(\sum_{j=2}^{i}\log(j-1-\alpha)) + \mathrm{const.}$$

のように表される。従って最尤推定量は、以下の同時方程式の解である。

$$\frac{\partial L}{\partial\theta} = \sum_{i=1}^{u-1}\frac{1}{\theta+i\alpha} - \sum_{i=1}^{n-1}\frac{1}{\theta+i} = 0,$$

$$\frac{\partial L}{\partial\alpha} = \sum_{i=1}^{u-1}\frac{i}{\theta+i\alpha} - \sum_{i=2}^{n}s_i\sum_{j=1}^{i-1}\frac{1}{J-\alpha} = 0.$$

最尤解は数値的に評価する必要が有る。例えば Hoshino[80] は二次の微分係数

$$\frac{\partial^2 L}{(\partial\theta)^2} = -\sum_{i=1}^{u-1}\frac{1}{(\theta+i\alpha)^2} + \sum_{i=1}^{n-1}\frac{1}{(\theta+i)^2},$$

$$\frac{\partial^2 L}{(\partial\alpha)^2} = -\sum_{i=1}^{u-1}\frac{i^2}{(\theta+i\alpha)^2} - \sum_{i=2}^{n}s_i\sum_{j=1}^{i-1}\frac{1}{(j-\alpha)^2} < 0,$$

$$\frac{\partial^2 L}{\partial\theta\partial\alpha} = -\sum_{i=1}^{u-1}\frac{i}{(i\alpha+\theta)^2} < 0$$

を利用し、ニュートン＝ラフソン法を用いた。なお最尤推定について、以下の命題が成立する。つまり数値計算において収束する初期値を探索する場合、まず Ewens モデルの最尤推定値を求めれば、その範囲を限定する事が出来る。

命題 4.3 (Hoshino[80], Proposition 1) *Pitman* モデルの最尤推定値を α^*, θ^* と書く。*Ewens* モデルの最尤推定値を θ_E とする。この時 $\alpha^* \leq 0$ でない限り、$\theta^* < \theta_E$ である。

Pitman モデルの母数空間のうち、α が非負のケースのみ考えてみよう。すなわち、

$$0 \leq \alpha < 1, \theta > -\alpha$$

に母数空間を限る。最尤解がこの内点 $(0 < \alpha < 1)$ に有るならば尤度方程式を解けば良いが、端点に解が有る時は尤度方程式を満たさない。最尤解 $\alpha^* = 0$ となるのはどのような場合だろうか。

手掛かりは、$u \geq 2$ について

$$\frac{\partial^2 L}{\partial \alpha^2} = -\sum_{i=1}^{u-1} \frac{i^2}{(\theta + i\alpha)^2} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{(j-\alpha)^2} < 0$$

および

$$\frac{\partial^2 L}{\partial \alpha \partial \theta} = -\sum_{i=1}^{u-1} \frac{i}{(i\alpha + \theta)^2} < 0$$

を利用する。つまり

- (a) $\partial L/\partial \alpha$ は θ, α について単調減少。
- (b) $\partial L/\partial \theta$ は α について単調減少。

まず $\partial L/\partial \alpha \mid_{\alpha=0} = 0$ を満たす $\theta = \theta_0$ を求める。

$$\frac{\partial L}{\partial \alpha} \mid_{\alpha=0} = \frac{u(u-1)}{2\theta} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j}$$

より、

$$\theta_0 = \frac{u(u-1)}{2T} > 0,$$

但し $T = \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j}$ なので、$u = n$ の時 $T = 0$ に注意。ここで単調性 (a) より

$$\begin{aligned} \partial L/\partial \alpha \mid_{\alpha=0} &> 0 \quad \text{for } \theta < \theta_0, \\ \partial L/\partial \alpha \mid_{\alpha=0} &= 0 \quad \text{for } \theta = \theta_0, \\ \partial L/\partial \alpha \mid_{\alpha=0} &< 0 \quad \text{for } \theta > \theta_0. \end{aligned}$$

次に $\partial L/\partial \theta \mid_{\alpha=0} = 0$ となる $\theta = \theta_1$ を考える。θ_1 は

$$\frac{u-1}{\theta} - \sum_{i=1}^{n-1} \frac{1}{\theta + i} = 0$$

の解である。先ほどの結果より

$$\frac{\partial L}{\partial \alpha} \mid_{\alpha=0, \theta=\theta_1} > 0, \quad \forall \theta_1 < \theta_0$$

なので、この時 $\alpha^* \neq 0$ である。つまり $\alpha^* = 0$ の必要条件として、$\theta_0 \leq \theta_1$ が言える。

次に

$$\frac{\partial L}{\partial \theta} \mid_{\alpha=0, \theta=\theta_1+\delta} < 0 \quad \text{for } \delta > 0$$

が言える。($\alpha = 0$ の時、Pitman 分布は Ewens 分布に退化する。Ewens 分布は指数族に属するから上の単調性を満たす。もちろん直接に $\partial L/\partial \theta \mid_{\alpha=0, \theta=\theta_1} - \partial L/\partial \theta \mid_{\alpha=0, \theta=\theta_1+\delta} > 0$ を示せる。) だとすると、単調性 (b) より

$$\frac{\partial L}{\partial \theta} < 0 \quad \text{for } \theta > \theta_1, \alpha \geq 0$$

なので、

$$\theta^* \leq \theta_1$$

が言える。また $\theta_0 < \theta_1$ の時、$\theta^* \leq \theta_0$ または $\theta^* = \theta_1$ のいずれかとなる。つまり $\theta_0 < \theta < \theta_1$ について、$\partial L/\partial\alpha < 0$ なので $\alpha = 0$ が最適である。だとすれば、$\theta = \theta_1$ の時の尤度を上回らない。

次に近似的モメント推定量を紹介する。

$$\hat{\theta} = \frac{nuc - s_1(n-1)(2u+c)}{2s_1u + s_1c - nc},$$

$$\hat{\alpha} = \frac{\hat{\theta}(s_1 - n) + (n-1)s_1}{nu};$$

ただし $c = s_1(s_1 - 1)/s_2$。導出は Hoshino[80] の Appendix を見よ。また n が大きい時に、(4.47) より

$$\hat{\alpha} = \frac{s_1}{u}$$

が使えるかもしれない。

4.9 条件付逆ガウシアン＝ポアソン (CIGP) モデル

4.9.1 定義等

Hoshino[83] は、逆ガウシアン＝ポアソンモデル (4.37) の母集団サイズ N を固定した条件付分布を CIGP 分布と呼ぶ。すなわち (4.37) を母集団サイズの確率 (4.39) でわれば、CIGP 分布を得る。

定義 母数 $\alpha > 0$ について、母集団サイズ N 所与の CIGP モデルは以下の様に定義される。

$$\mathrm{P}(F_1 = y_1, \ldots, F_J = y_J | N) = (\frac{2\alpha}{\pi})^{\frac{J-1}{2}} \frac{N!}{J^{N+1/2} K_{N-1/2}(J\alpha)} \prod_{j=1}^{J} \frac{K_{y_j - 1/2}(\alpha)}{y_j!}.$$

同値だが、寸法指標については

$$\mathrm{P}(S_0, \ldots, S_N) = (\frac{2\alpha}{\pi})^{\frac{J-1}{2}} \frac{J!N!}{J^{N+1/2} K_{N-1/2}(J\alpha)} \prod_{i=0}^{N} \{\frac{K_{i-1/2}(\alpha)}{i!}\}^{S_i} \frac{1}{S_i!} \tag{4.48}$$

と書ける。

逆に CIGP 分布 (4.48) の集団サイズ N が分布 (4.39) に従うような混合をすれば、IGP モデル (4.37) を得る。また $J\alpha = A$ を固定し、$J \to \infty, \alpha \to 0$ という極限をとれば、極限 CIGP 分布 (4.54) となる。

4.9.2 モメント

寸法指標の同時階乗モメントのみ考察する。Hoshino[83] によれば、非負整数の $r_j, j = 1, 2, \ldots, N$, について

$$\mathrm{E}(\prod_{j=1}^{N} S_j^{(r_j)}) = (\frac{2\alpha}{\pi})^{\frac{r}{2}} \frac{N!J!K_{N-R-1/2}((J-r)\alpha)(J-r)^{N-R+1/2}}{(N-R)!(J-r)!J^{N+1/2}K_{N-1/2}(J\alpha)} \prod_{j=1}^{N} (\frac{K_{j-1/2}(\alpha)}{j!})^{r_j}, \tag{4.49}$$

ただし $r=\sum_{j=1}^{N} r_j\,(\leq J)$, $R=\sum_{j=1}^{N} jr_j\,(\leq N)$, となる。特に

$$\mathrm{E}(S_i)=\sqrt{\frac{2\alpha}{\pi}}\frac{K_{i-1/2}(\alpha)}{i!}\frac{N!K_{N-i-1/2}((J-1)\alpha)(J-1)^{N-i+1/2}}{(N-i)!J^{N-1/2}K_{N-1/2}(J\alpha)},$$

$$\begin{aligned}\mathrm{V}(S_i) &= (\frac{2\alpha}{\pi})\frac{N!J!K_{N-2i-1/2}((J-2)\alpha)(J-2)^{N-2i+1/2}}{(N-2i)!(J-2)!J^{N+1/2}K_{N-1/2}(J\alpha)}(\frac{K_{i-1/2}(\alpha)}{i!})^2\\ &+\sqrt{\frac{2\alpha}{\pi}}\frac{K_{i-1/2}(\alpha)}{i!}\frac{N!K_{N-i-1/2}((J-1)\alpha)(J-1)^{N-i+1/2}}{(N-i)!J^{N-1/2}K_{N-1/2}(J\alpha)}\\ &-(\sqrt{\frac{2\alpha}{\pi}}\frac{K_{i-1/2}(\alpha)}{i!}\frac{N!K_{N-i-1/2}((J-1)\alpha)(J-1)^{N-i+1/2}}{(N-i)!J^{N-1/2}K_{N-1/2}(J\alpha)})^2.\end{aligned}$$

また Hoshino[83] によれば、$N\to\infty$ という極限において

$$\mathrm{E}(S_1)=\exp(-\alpha)\alpha\frac{J-1}{2}+O(N^{-1}).$$

なお $j=1,2,\ldots$ について、もし $S_j\geq 1$ ならば

$$\begin{aligned}S_j\,\mathrm{P}_J(S_0,\ldots,S_N|N) &= \mathrm{P}_J(S_0+1,\ldots,S_{j-1},S_j-1,S_{j+1},\ldots,S_N|N-j)\\ &\times(S_0+1)(\frac{K_{j-1/2}(\alpha)}{j!})^1\frac{N!K_{N-j-1/2}(J\alpha)}{(N-j)!J^jK_{N-1/2}(J\alpha)}\frac{1}{K_{-1/2}(\alpha)},\end{aligned}$$

さもなければ $S_j=0$ かつ $S_j\,\mathrm{P}(S_0,\ldots,S_N|N)=0$ である。従って $j=1,2,\ldots$ について

$$\begin{aligned}\mathrm{E}_J(S_j|N) &= \sum_{\boldsymbol{S}\in\mathcal{S}(N)} S_j\,\mathrm{P}_J(S_0,\ldots,S_N|N)\\ &= (\frac{K_{j-1/2}(\alpha)}{j!})^1\frac{N!K_{N-j-1/2}(J\alpha)}{(N-j)!J^jK_{N-1/2}(J\alpha)}\frac{1}{K_{-1/2}(\alpha)}\\ &\times\sum_{\boldsymbol{S}\in\mathcal{S}(N)}(S_0+1)\,\mathrm{P}_J(S_0+1,\ldots,S_{j-1},S_j-1,S_{j+1},\ldots,S_N|N-j)\,I(S_j\geq 1)\\ &= (\frac{K_{j-1/2}(\alpha)}{j!})^1\frac{N!K_{N-j-1/2}(J\alpha)}{(N-j)!J^jK_{N-1/2}(J\alpha)}\frac{1}{K_{-1/2}(\alpha)}\\ &\times\sum_{\boldsymbol{S}\in\mathcal{S}(N-j)}(S_0)\,\mathrm{P}_J(S_0,\ldots,S_{j-1},S_j,s_{j+1},\ldots,S_N|N-j)\\ &= (\frac{K_{j-1/2}(\alpha)}{j!})^1\frac{N!K_{N-j-1/2}(J\alpha)}{(N-j)!J^jK_{N-1/2}(J\alpha)}\frac{1}{K_{-1/2}(\alpha)}\times\mathrm{E}_J(S_0|N-j)\end{aligned}$$

という関係が成立する。

4.9.3 母数推測

CIGP 分布の母数推測では、(G)IGP 分布に関する議論で用いた技法がかなり使える。なお Hoshino[83] が指摘するように、CIGP 分布は個体のラベルに依存しない。故に命題 2.1 より、分

布 (4.48) に従う大きさ N の母集団から n 個の標本を非復元単純無作為抽出した場合、標本分布は以下の様になる。

$$\mathrm{P}(s_0,\ldots,s_n)=(\frac{2\alpha}{\pi})^{\frac{J-1}{2}}\frac{J!n!}{J^{n+1/2}K_{n-1/2}(J\alpha)}\prod_{i=0}^{n}\{\frac{K_{i-1/2}(\alpha)}{i!}\}^{s_i}\frac{1}{s_i!}.$$

対数尤度は以下の様に表される。

$$L=\frac{J-1}{2}\log(2\alpha)-\log K_{n-1/2}(J\alpha)+\sum_{i=0}^{n}s_i\log K_{i-1/2}(\alpha)+\mathrm{const.}$$

ここで

$$\begin{aligned}\frac{dL}{d\alpha} &= \frac{J-1}{2\alpha}-\{-R_{n-1/2}(J\alpha)+\frac{n-1/2}{J\alpha}\}J+\sum_{i=0}s_i\{-R_{i-1/2}(\alpha)+\frac{i-1/2}{\alpha}\}\\ &= JR_{n-1/2}(J\alpha)-\sum_{i=0}^{n}s_iR_{i-1/2}(\alpha),\end{aligned}$$

ただし $R_\gamma(\alpha)=K_{\gamma+1}(\alpha)/K_\gamma(\alpha)$ となる。最尤推定量は尤度方程式 $dL/d\alpha=0$ の解である。

二次の微分係数は

$$\frac{d^2L}{d\alpha^2}=J^2\{R^2_{n-1/2}(J\alpha)+\frac{2n}{J\alpha}R_{n-1/2}(J\alpha)\}-\sum_{i=0}s_i\{R^2_{i-1/2}(\alpha)+\frac{2i}{\alpha}R_{i-1/2}(\alpha)\}-J^2+J$$

となる。

モメント推定については、変形ベッセル関数の評価が問題になる。すなわち厳密な推定量は計算上不便である。従って Hoshino[83] は、IGP 分布のモメント推定量を利用して近似的モメント推定量を提案した。(4.40) から θ の推定量を消去すれば、近似的推定量

$$\tilde{\alpha}=\frac{n\sqrt{n(2Jv-n)}}{J(Jv-n)}$$

を得る。ただし v は標本分散 (4.8) である。

また s_0 を用いた IGP 分布の推定量 (4.41) に対応して

$$\bar{\alpha}=-\frac{1}{2}(\log s_0-\log J)(1+\frac{n/J}{n/J+\log s_0-\log J})$$

が提案された。Hoshino[83] による実データへの当てはめでは、後者の近似的推定値が前者より最尤推定値に近い傾向が見られた。

4.10 拡張負の二項モデル

4.10.1 定義等

ガンマ=ポアソン分布モデルでの寸法指標 S_i の期待値 (4.15) は、負の二項分布 (4.10) の $\mathrm{P}(F=i)$ に J をかけたものであった。ここで期待値制約 (4.13) より、$J=1/\gamma\beta$ を代入すると

$$\mathrm{E}(S_i)=\frac{1}{\beta}\frac{\Gamma(i+\gamma)}{\Gamma(1+\gamma)i!}(\frac{1}{N_0\beta+1})^\gamma(\frac{N_0\beta}{N_0\beta+1})^i,\quad i=0.1.2.\ldots, \tag{4.50}$$

を得る。対数級数モデルでは $\gamma \to 0$ の挙動を考察したが、Engen[46] は (4.50) 式の $\Gamma(1+\gamma)$ に着目し、$i \geq 1$ ならば $-1 < \gamma < 0$ と出来ると指摘した（生態学の文脈だったので、S_0 は考慮されていない)。そしてこのようなモデルを、「拡張負の二項（Extended Negative Binomial, ENB）モデル」と呼ぶ。しかし寸法指標の同時分布は与えられず、ENB モデルの着想は次のかたちで利用されてきた。すなわち $-1 < \gamma < 0, 0 < \theta \leq 1$ について、確率関数

$$\mathrm{P}(F = y) = \frac{-\gamma}{1-(1-\theta)^{-\gamma}} \frac{\Gamma(y+\gamma)}{\Gamma(\gamma+1)y!}\theta^y, \quad y = 1, 2, \ldots, \tag{4.51}$$

を持つ分布を、「拡張（切り落とし）負の二項分布」(extended truncated negative binomial distribution）と呼ぶ。確率関数 (4.51) は、$y = i$ ならば ENB モデルの下での S_i の期待値 (4.50) に比例する。この分布は、ENB モデルとは区別しなければならない。より詳しくは Engen[49]、または Johnson et al.[90] の 5.12.2 節を見よ。なお θ が 1 の時、拡張切り落とし負の二項分布は渋谷分布 (Sibuya[163]) である。(4.51) の確率母関数は

$$g(z) = \frac{1-(1-z\theta)^{-\gamma}}{1-(1-\theta)^{-\gamma}}$$

のように書ける。渋谷分布の確率母関数は

$$g_s(z) = 1-(1-z)^{-\gamma}$$

なので、$g(z) = g_s(\theta z)/g_s(\theta)$ という関係が成立する。すなわち拡張切り落とし負の二項分布は、ベキ級数分布化した渋谷分布である。ただし渋谷分布は期待値が発散してしまうので、母集団モデルとして使いづらい。なお γ が正の時、(4.51) は Sampford[150] の切り落とし負の二項分布である。

Sichel[180] によれば、拡張切り落とし負の二項分布の歪度は対数級数分布と対数正規＝ポアソン分布の間になる。従って、多くの観測データを記述するのに都合が良いと言う。経験的に拡張切り落とし負の二項分布が集団を良く記述するという主張は、例えばそもそも Engen[46] にも見られる。Hoshino[82] でも議論されたように、γ が-1 に近づくと $\mathrm{P}(F = 1)$ の割合が増加する。従って小集団が支配的な（$\gamma = 0$ に対応する対数級数曲線よりも歪んだ）データでは、拡張負の二項分布が有力である。

(4.50) で $\theta = N_0\beta/(N_0\beta+1)$ とおいて、$\gamma = -1/2$ とすると

$$\mathrm{E}(S_i) = \frac{1}{\beta}\frac{\Gamma(i-1/2)}{\Gamma(1/2)i!}\frac{1}{\sqrt{1-\theta}}\theta^i = \frac{\Gamma(i-1/2)}{\Gamma(1/2)i!}\frac{N_0\sqrt{1-\theta}}{\theta}\theta^i \tag{4.52}$$

のように書ける。逆ガウシアン＝ポアソンモデルに小数法則を適用すると、各 S_i 独立に平均母数 (4.52) のポアソン分布に従うモデルが得られる (Hoshino[81])。これを ENB モデルの特殊ケースとして考えれば、制約 (4.50) を満たすモデルが自然に得られる。実際、IGP モデルを一般化した GIGP モデルに小数法則を適用すれば、一般の ENB モデルが得られる（Hoshino[82])。

定義 $-1 < \gamma\,(\gamma \neq 0), 0 < \theta < 1$ について、母集団サイズ N が確率変数である拡張負の二項モデルは次のように定義される。

$$\mathrm{P}(S_1, S_2, \ldots) = \prod_{i=1}^{\infty} \frac{\exp(-\tau(i;\gamma.\theta))\tau(i;\gamma.\theta)^{S_i}}{S_i!}, \tag{4.53}$$

ただし

$$\tau(i;\gamma,\theta)=\frac{N_0}{\theta}\frac{(1-\theta)^{\gamma+1}\theta^i\Gamma(i+\gamma)}{\Gamma(\gamma+1)i!}$$

である。なお $\mathrm{E}(N)=N_0$ となる。特に $-1<\gamma<0$ のみ考える場合は、そのように明示する。

ENB モデルでは、3.2 節で考察したように、無限大個のセルが有るとして各度数 F_j が 1 以上という条件付分布が拡張（切り落とし）負の二項分布に従っていると解釈出来る。$\theta=1$ の場合は母集団サイズが有限の期待値を持たないので、母数空間から除いている。

4.10.2 モメント

S_i がポアソン分布に従う事より、非負整数の r について

$$\mathrm{E}(S_i^{(r)})=\tau(i;\gamma,\theta)^r,\quad i=1,2,\ldots,$$

のように書ける。Johnson et al.[90] の 4 章を見よ。なお各 i 独立であり、特に

$$\mathrm{E}(S_i)=\tau(i;\gamma,\theta),$$

$$\mathrm{V}(S_i)=\tau(i;\gamma,\theta).$$

4.10.3 母数推測

ENB モデルでは母集団サイズが確率変数なので、ベルヌーイ抽出が用いられる。抽出率が n_0/N_0 の場合、標本分布は以下の様に表される。

$$\mathrm{P}(s_1,s_2,\ldots)=\prod_{i=1}^{\infty}\frac{\exp(-\tau'(i;\gamma,\theta))\tau'(i;\gamma,\theta)^{s_i}}{s_i!},$$

ただし

$$\tau'(i;\gamma,\theta)=\frac{n_0}{\theta}\frac{(1-\theta)^{\gamma+1}\theta^i\Gamma(i+\gamma)}{\Gamma(\gamma+1)i!}$$

である。従って対数尤度は

$$\begin{aligned}L &= -n_0(1-\theta)^{\gamma+1}\frac{(1-\theta)^{-\gamma}-11}{\theta\gamma}+(n-u)\log\theta+u(\gamma+1)\log(1-\theta)\\ &\quad+\sum_{i=1}^{\infty}s_i(\log\Gamma(i+\gamma)-\log\Gamma(1+\gamma))+\mathrm{const.}\end{aligned}$$

のように書ける。

対数尤度について、一次の微分係数は以下の式で与えられる。

$$\frac{\partial L}{\partial\theta} = -\frac{n_0}{\gamma}\left(\frac{(1-\theta)^{\gamma+1}-1}{\theta^2}+\frac{(1+\gamma)(1-\theta)^{\gamma}}{\theta}\right)+(n-u)\frac{1}{\theta}-u\frac{(\gamma+1)}{1-\theta}.$$

$$\frac{\partial L}{\partial \gamma} = n_0 \frac{1-\theta}{\theta\gamma^2} + \frac{n_0}{\theta}(-\frac{(1-\theta)^{\gamma+1}}{\gamma^2} + \frac{(1-\theta)^{\gamma+1}}{\gamma}\log(1-\theta)) + u\log(1-\theta) + S_1 + \sum_{i=2}^{\infty} s_i \sum_{j=1}^{i-1} \frac{1}{\gamma+j}.$$

また二次の微分係数は以下の式で与えられる。

$$\frac{\partial^2 L}{\partial\gamma\partial\theta} = \frac{n_0((1-\theta)^{\gamma+1}-1)}{\theta^2\gamma^2} - \frac{n_0}{\theta^2\gamma}(1-\theta)^{\gamma+1}\log(1-\theta) + \frac{n_0}{\theta\gamma^2}(1-\theta)^{\gamma} - \frac{n_0}{\theta\gamma}(1-\theta)^{\gamma}\log(1-\theta) - \frac{n_0}{\theta}(1-\theta)^{\gamma}\log(1-\theta) - \frac{u}{1-\theta}.$$

$$\frac{\partial^2 L}{\partial\theta^2} = -\frac{n_0}{\gamma}(\frac{2-2(1-\theta)^{\gamma+1}}{\theta^3} - \frac{2(1+\gamma)(1-\theta)^{\gamma}}{\theta^2} - \frac{(1+\gamma)\gamma(1-\theta)^{\gamma-1}}{\theta}) - \frac{n-u}{\theta^2} - \frac{u(\gamma+1)}{(1-\theta)^2}.$$

$$\frac{\partial^2 L}{\partial\gamma^2} = \frac{n_0}{\theta}(\frac{-2(1-\theta)+2(1-\theta)^{\gamma+1}}{\gamma^3} - \log(1-\theta)\frac{2(1-\theta)^{\gamma+1}}{\gamma^2} + \log^2(1-\theta)\frac{(1-\theta)^{\gamma+1}}{\gamma}) - \sum_{i=2}^{\infty} s_i \sum_{j=1}^{i-1} \frac{1}{(j+\gamma)^2}.$$

最尤推定は、同時方程式 $\partial L/\partial\theta = \partial L/\partial\gamma = 0$ を解けば良い。ただし応用の際は n_0 を n の実現値で置きかえるので、$n = n_0$ となる。この時尤度方程式は簡略化されて、以下の様になる。

$$\frac{\partial L}{\partial\theta} = 0 \Leftrightarrow u = \frac{n_0(1-(1-\theta)^{\gamma})(1-\theta)}{\gamma\theta},$$

$$\frac{\partial L}{\partial\gamma} = 0 \Leftrightarrow \frac{n_0(1-\theta)}{\gamma\theta}\{\frac{1-(1-\theta)^{\gamma}}{\gamma} + \log(1-\theta)\} + \sum_{i=2}^{\infty} s_i \sum_{j=1}^{i-1} \frac{1}{\gamma+j} = 0.$$

単純なモメント推定では、非線型方程式を解く必要が有る。従って利便性の点で益がないので、本稿では考察しない。なお Engen[46] は様々な推定法について考察し、擬似モメント推定を推奨している。

ENB モデルからの非復元単純無作為抽出は、N 所与の条件付分布で考えれば良い。4.12 節でこれを考察する。しかし $\gamma = -1/2$ の場合（極限 CIGP 分布）を除いて、利便性に欠けるように思う。

4.11 極限 CIGP モデル

4.11.1 定義等

Hoshino[81] は、CIGP 分布 (4.48) に小数法則を適用した。すなわち $J\alpha = A$ とおき、$J \to \infty.\alpha \to 0$ とした。ここで得られた分布を、極限 CIGP(LCIGP) 分布と呼ぶ事にする。

定義 母数 $A>0$ について、LCIGP モデルは次のように表される。

$$\mathrm{P}(S_1,\ldots,S_N)=\sqrt{\frac{\pi}{2}}\frac{N!\exp(-A)}{A^{N-U+1/2}K_{N-1/2}(A)}\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{S_i}\frac{1}{S_i!}. \tag{4.54}$$

なお本モデルでは、母集団サイズ N は固定されている。

LCIGP 分布は、拡張負の二項モデル (4.53) で $\gamma=-1/2$ とした時の母集団サイズ条件付き分布でもある。Ewens 分布で $N\to\infty$ とした場合に有限項の対数級数モデルが得られる（命題 4.2）ように、極限 CIGP 分布についても $N\to\infty$ で有限項の拡張負の二項分布モデルが得られる。命題として述べれば次のようになる。

命題 4.4 有限で正の整数 m を固定する。*LCIGP*分布 (4.54) の最初の m 要素である $(S_1,S_2,\ldots,S_m)$ は $N\to\infty$ の時 $i=1,2,\ldots,m$ について独立な平均

$$\mu_i=A(\frac{1}{2})^i\frac{(2i-3)!!}{i!}$$

のポアソン分布の同時分布に収束する。

命題 4.4 の μ_i は、ENB モデル (4.53) で $\gamma=-1/2,\theta=1$ とした場合の $\mathrm{E}(S_i)$ に他ならない。本命題は、Sibuya[166] による命題 4.2 の証明と同様に示す事が出来る。4.11.2 節で示されているように、m 項の同時階乗モメント $\mathrm{E}(\prod_{i=1}^{m}S_i^{(r_i)})$ は

$$M(r_1,r_2,\ldots,r_m)=\frac{K_{N-R-1/2}(A)A^{r-R}N!}{K_{N-1/2}(A)(N-R)!}\prod_{i=1}^{m}(\frac{(2i-3)!!}{i!})^{r_i},$$

ただし $r_i,i=1,2,\ldots,m,$ は非負整数、となる。ここで Ismail[87] の公式

$$\lim_{\gamma\to\infty}K_\gamma(A)=2^\gamma\gamma^\gamma\exp(-\gamma)A^{-\gamma}\sqrt{\frac{\pi}{2\gamma}}$$

を用いれば $N\to\infty$ の時

$$M(r_1,r_2,\ldots,r_m)\to\prod_{i=1}^{m}\{\mu_i\}^{r_i}$$

となる。つまりすべての次数 $r_i,i=1,2,\ldots,m,$ について同時階乗モメントが独立なポアソン分布の同時階乗モメントに収束するので、分布収束が示された。

4.11.2 モメント

Hoshino[81] によれば、寸法指標の同時階乗モメントは非負整数の $r_i,i=1,2,\ldots,N,$ について次式で与えられる。

$$\mathrm{E}(\prod_{i=1}^{N}S_i^{(r_i)})=\frac{K_{N-R-1/2}(A)A^{r-R}N!}{K_{N-1/2}(A)(N-R)!}\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{r_i}. \tag{4.55}$$

ただし $r=\sum_{i=1}^{N}r_i,R=\sum_{i=1}^{N}ir_i\,(\le N),n^{(R)}=n(n-1)\cdots(n-R+1)$ である。

特に $i=1,2,\ldots,N$ について

$$\mathrm{E}(S_i)=\frac{K_{N-i-1/2}(A)A^{1-i}(2i-3)!!N!}{K_{N-1/2}(A)i!(N-i)!},$$

$$\begin{aligned}\mathrm{V}(S_i) &= \frac{K_{N-2i-1/2}(A)A^{2-2i}N!}{K_{N-1/2}(A)i!(N-2i)!}(\frac{(2i-3)!!}{i!})^2\\ &+\frac{K_{N-i-1/2}(A)A^{1-i}(2i-3)!!N!}{K_{N-1/2}(A)i!(N-i)!}\\ &-(\frac{K_{N-i-1/2}(A)A^{1-i}(2i-3)!!N!}{K_{N-1/2}(A)i!(N-i)!})^2.\end{aligned}$$

Hoshino[81] は $\mathrm{P}(U)$ を求めているので

$$\mathrm{P}(S_1,S_2,\ldots,S_N|U)=N!\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{S_i}\frac{1}{S_i!}.2^{N-U}\frac{(U-1)!(N-U)!}{(2N-U-1)!} \tag{4.56}$$

と評価できる。ここで条件付のモメントを評価しよう。非負整数の $r_i, i=1,2\ldots,N,$ について

$$\begin{aligned}&\mathrm{E}(\prod_{i=1}^{N}S_i^{(r_i)}|U)\\ &= \sum_{\boldsymbol{S}\in\mathcal{S}(N,U)}N!\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{S_i}\frac{1}{(S_i-r_i)!}.2^{N-U}\frac{(U-1)!(N-U)!}{(2N-U-1)!}I(S_i\geq r_i)\\ &= \frac{N!}{(N-R)!}\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{r_i}\frac{(U-1)!(N-U)!}{(2N-U-1)!}2^{R-r}\frac{(2N-2R-U+r-1)!}{(U-r-1)!(N-R-U+r)!}(N-R)!\times\\ &\quad\sum_{\boldsymbol{S}\in\mathcal{S}(N,U)}\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{S_i-r_i}\frac{1}{(S_i-r_i)!}.2^{N-R-U+r}\frac{(U-r-1)!(N-R-U+r)!}{(2N-2R-U+r-1)!}I(S_i\geq r_i)\\ &= \frac{N!}{(N-R)!}\prod_{i=1}^{N}(\frac{(2i-3)!!}{i!})^{r_i}\frac{(U-1)!(N-U)!}{(2N-U-1)!}2^{R-r}\frac{(2N-2R-U+r-1)!}{(U-r-1)!(N-R-U+r)!},\end{aligned}$$

ただし $r=\sum_{i=1}^{N}r_i(\leq U), R=\sum_{i=1}^{N}ir_i(\leq N)$ であり、

$$\mathcal{S}(N,U)=\{\boldsymbol{S}=(S_1,S_2,\ldots,S_N)|\sum_{i=1}^{N}iS_i=N,\sum_{i=1}^{N}S_i=U\}$$

である。最後の式で $\sum_{\boldsymbol{S}\in\mathcal{S}(N-R,U-r)}\mathrm{P}(\boldsymbol{S}|U-r)=1$ を用いた。特に

$$E(S_i|U)=\frac{N!}{(N-i)!}(\frac{(2i-3)!!}{i!})\frac{(U-1)(N-U)!}{(N-U-i+1)!}2^{i-1}\frac{(2N-U-2i)!}{(2N-U-1)!},\quad i=1,2,\ldots,N,$$

なので

$$\lim_{N\to\infty}E(S_i|U)=\frac{(2i-3)!!}{i!}(U-1)2^{-i}=(U-1)\frac{1}{2\sqrt{\pi}}\frac{\Gamma(i-1/2)}{\Gamma(i+1)},\quad i=1,2,\ldots,N.$$

である。従って N, U が大きい時には

$$\frac{\mathrm{E}(S_i|U)}{U} \doteq \frac{1}{2\sqrt{\pi}} \frac{\Gamma(i-1/2)}{\Gamma(i+1)} = \frac{1}{2\Gamma(1/2)} \frac{\Gamma(i-1/2)}{\Gamma(i+1)}, \quad i = 1, 2, \ldots, N,$$

であり、これは拡張（切り落とし）負の二項分布 (4.51) で $\gamma = -1/2, \theta = 1$ とした場合の確率 $\mathrm{P}(F = i)$ と等しい。先ほど示した $(S_1, S_2, \ldots, S_m)$ の独立ポアソン同時分布への収束と整合的な結果である。また $\mathrm{E}(S_i|U)/U$ の極限は、Pitman 分布 (4.46) で母数 $\alpha = 1/2$ の場合の $\lim_{N\to\infty} \mathrm{E}(S_i)/\mathrm{E}(U)$ と等しい（(4.47) 式を見よ）。実は $\alpha = 1/2$ の Pitman 分布の U 所与の条件付分布 $\mathrm{P}(S_1, S_2, \ldots, S_N|U, N)$ は (4.56) のように書ける。従って S_i/U が一致するのは偶然ではない。これについては、4.12 節でより詳しく議論する。

4.11.3 母数推測

本モデルでは母集団サイズが固定されている。故に命題 2.1 が利用出来る。非復元単純無作為抽出の場合、大きさ n の標本分布は次のように書ける。

$$\mathrm{P}(s_1, \ldots, s_n) = \sqrt{\frac{\pi}{2}} \frac{n! \exp(-A)}{A^{n-u+1/2} K_{n-1/2}(A)} \prod_{i=1}^{n} \left(\frac{(2i-3)!!}{i!}\right)^{s_i} \frac{1}{s_i!}.$$

従って対数尤度 L は次式で表される。

$$L = -A - (n - u + \frac{1}{2}) \log A - \log(K_{n-1/2}(A)) + \text{const.}$$

対数尤度について一次の微分係数は以下の通り。

$$\frac{\partial L}{\partial A} = -1 - (2n - u)\frac{1}{A} + R_{n-1/2}(A),$$

ただし $R_\gamma(\omega) = K_{\gamma+1}(\omega)/K_\gamma(\omega)$ である。変形ベッセル関数の微分に関しては、GIGP 分布（3.4.3 節）での議論を参照の事。最尤推定量は、$\partial L/\partial A = 0$ の解である。これを解くには二次の微分係数

$$\frac{\partial^2 L}{\partial A^2} = R^2_{n-1/2}(A) - \frac{2n}{A} R_{n-1/2}(A) + (2n - u)\frac{1}{A^2} - 1$$

を用いて、ニュートン＝ラフソン法が利用できる。極限 CIGP 分布は指数族に所属するので、最尤推定の挙動は良好である。なお u は母数 A の十分統計量である。

極限 CIGP 分布について厳密なモメント推定量は、変形ベッセル関数の評価を伴い実用的ではない。従って Hoshino[81] は、近似的なモメント推定量

$$\tilde{A} = u/(1 - \sqrt{1 - 4s_2/s_1})$$

を提案した。しかし現実のデータでは $4s_2/s_1$ は 1 より大きくなりうる。この場合推定値は虚数となり、非合理である。そのような場合は、u を A の推定値とすれば良いだろう。

4.12 条件付拡張負の二項モデルと関連する分布について

ENB モデル (4.53) において、γ は-1 から 0 までの値をとりうる。$\gamma = -1/2$ の場合については、N 所与の条件付分布として既に極限 CIGP 分布 (4.54) の性質を議論した。しかし $\gamma \neq -1/2$ につい

ては、拡張負の二項モデルの条件付分布を考察していない。そのような条件付分布は、極限CIGP分布を特殊ケースとして含む一般化された分布になるはずである。またそのような分布は、小数法則の極限として表れるはずである。故に $-1 < \gamma < 0$ に対応する、一般化されたCIGP分布も考えられる。だとすれば、一般化されたCIGP分布を条件付分布とする、一般化されたIGP分布モデルはどのように書けるだろうか（図3.8, 4.1参照の事）。

実はそのような一般化IGPモデルは、複数有りえる。一般に $\mathrm{E}(N)$ を固定して $J \to \infty$ という制約を満たす系列は、極限が同一だとしても出発点は複数有りえる。一例を挙げよう。4.4節で検討したGIGP分布モデル (4.27) は、小数法則を適用する事で拡張負の二項モデルが得られる。しかし小数法則の適用前と後で、N の分布が変化する。ここで N の分布が変化しないとすれば、GIGP分布とは別の一般化されたIGP分布を得ることが出来る。すなわち、拡張負の二項分布をクラスター分布とする複合ポアソンである。

Hoshino[82] の Proposition 1 によれば、そのような一般化されたIGP分布の確率母関数は次式で表される。

$$G(z) = \exp(\alpha\{(1-\theta)^{-\gamma} - (1-z\theta)^{-\gamma}\}), \tag{4.57}$$

ただし $0 < \alpha$, $0 < \theta < 1$, $-1 < \gamma < 0$ である。この分布を本稿ではKatti and Gurland[96]にならい、ポアソン・パスカル分布と呼ぶ（例3.2も参照の事）。もし γ が正ならば、この分布は負の二項分布をクラスター分布とする複合ポアソンであり、Skellam[183]はこれを一般化Polya-Aeppli分布と呼ぶ。ここで実は $-1 < \gamma < 0$ と出来る事を指摘したのは、Willmot[225]の貢献である。

$F_j, j = 1, 2, \ldots, J,$ が互いに独立に (4.57) で定義される分布に従うとする。ここで N の分布が固定されているならば、$J \to \infty$ ($\alpha \to 0$) の極限で拡張負の二項モデルが得られる。もし $\theta = 1$ ならば、(4.57) は離散安定分布の確率母関数である。3.1節で議論したように、離散安定分布は $\gamma = -1$ の場合を除き有限の期待値を持たない（つまり $\mathrm{E}(F_j) = \infty$）。この場合は母集団モデルとして役に立たないので、$\theta = 1$ を母数空間から除いてある。IGP分布が指数1/2の離散安定分布を特殊ケースとして含む事に着目すれば、(4.57) は一般の離散安定分布に対応する自然な一般化である。

なおポアソン・パスカル分布 (4.57) について、$\gamma \to -1$ の極限で

$$G(z) \to \exp(\alpha\theta(z-1))$$

となる。これは平均 $\alpha\theta$ のポアソン分布の確率母関数である。なお (4.57) は $\gamma = -1$ の時もプロパーな確率母関数だが、ENBモデルでは $\gamma = -1$ と出来ない為に母数空間から $\gamma = -1$ を除いている。また $\gamma \to -0$ の場合

$$(1-\theta)^{-\gamma} - (1-z\theta)^{-\gamma} = \gamma(\log(1-z\theta) - \log(1-\theta)) + O(\gamma^2)$$

なので、$\alpha\gamma = \nu$ が負の定数になるように $\alpha \to +\infty$ ($\gamma \to -0$) とすれば

$$\begin{aligned} G(z) &\to \exp(\nu(\log(1-z\theta) - \log(1-\theta))) \\ &= \exp(\nu \log(1-\theta)(g(z)-1)), \end{aligned}$$

ただし

$$g(z) = \frac{\log(1-z\theta)}{\log(1-\theta)}$$

となり、極限分布は負の二項分布である。

問題は、(4.57) の確率関数が $\gamma = -1, -1/2, 0$ の場合を除いて解析的に取り扱いにくい事である。同様に一般化 CIGP 分布、一般化極限 CIGP 分布の確率関数もこれらの場合を除いて複雑である。Hoshino[82] の Proposition 2 によれば、確率変数 F がポアソン・パスカル分布 (4.57) に従う場合、

$$\mathrm{P}(F=i) = \frac{B(i;\alpha,\gamma)\theta^i}{\exp(\alpha(1-(1-\theta)^{-\gamma}))}, \quad i=0,1,2,\ldots,$$

ただし $B(i;\alpha,\gamma)$ は漸化式で定義される。すなわち $B(0;\alpha,\gamma)=1$ かつ

$$B(i+1;\alpha,\gamma) = \frac{\alpha}{i+1}\sum_{j=0}^{i}\frac{-\gamma\Gamma(i+1-j+\gamma)}{\Gamma(\gamma+1)(i-j)!}B(j;\alpha,\gamma)$$

となる。一般化 CIGP 分布は

$$\mathrm{P}(F_1,F_2,\ldots,F_J|N) = \prod_{j=1}^{J}B(F_j;\alpha,\gamma)/B(N;J\alpha,\gamma)$$

または

$$\mathrm{P}_J(S_0,S_1,\ldots,S_N|N) = \frac{J!}{B(N;J\alpha,\gamma)}\prod_{i=0}^{N}\frac{B(i;\alpha,\gamma)^{S_i}}{S_i!} \tag{4.58}$$

である。そして一般化極限 CIGP 分布、すなわち拡張負の二項モデルの N 所与の条件付分布は $A>0$ について

$$\mathrm{P}(S_1,S_2,\ldots,S_N|N) = \{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^U\frac{1}{B(N;A,\gamma)}\prod_{i=1}^{N}\{\frac{\Gamma(i+\gamma)}{i!}\}^{S_i}\frac{1}{S_i!} \tag{4.59}$$

と書ける。これらの結果は命題 3.8 から導く事が出来る。拡張負の二項分布がベキ級数分布なのでポアソン・パスカル分布もベキ級数分布となり、N 所与の条件付分布 (4.58),(4.59) はベキ母数 θ に依存しない。

まず一般化 CIGP 分布 (4.58) から考察する。特殊な場合を除いて $B(i;\alpha,\gamma)$ の評価が簡単ではないが、少なくとも階乗モメントは以下のように評価出来る。非負整数 $r_i, i=1,2,\ldots,N$, について $r=\sum_{i=1}^{N}r_i$, $R=\sum_{i=1}^{N}ir_i\,(R\le N)$ とする。なお $n^{(R)}=n(n-1)\cdots(n-R+1)$ である。この時

$$\begin{aligned}
\mathrm{E}(\prod_{i=1}^{N}S_i^{(r_i)}) &= \sum_{\boldsymbol{S}\in\mathcal{S}_{N,J}}\frac{J!}{B(N;J\alpha,\gamma)}\prod_{i=0}^{N}\frac{B(i;\alpha,\gamma)^{S_i}}{(S_i-r_i)!}I(S_i\ge r_i)\\
&= \frac{J!B(N-R;(J-r)\alpha,\gamma)}{(J-r)!B(N;J\alpha,\gamma)}\prod_{i=0}^{N}B(i;\alpha,\gamma)^{r_i}\times\\
&\quad\sum_{\boldsymbol{S}\in\mathcal{S}_{N,J}}\frac{(J-r)!}{B(N-R;(J-r)\alpha,\gamma)}\prod_{i=0}^{N}\frac{B(i;\alpha,\gamma)^{S_i-r_i}}{(S_i-r_i)!}I(S_i\ge r_i),
\end{aligned}$$

ただし

$$\mathcal{S}_{N,J} = \{(S_1,S_2,\ldots,S_N)|\sum_{i=1}^{N}iS_i=N,\ \sum_{i=0}^{N}S_i=J\}$$

である。ここで

$$\sum_{\boldsymbol{S}\in\mathcal{S}_{N,J}} \frac{(J-r)!}{B(N-R;(J-r)\alpha,\gamma)} \prod_{i=0}^{N} \frac{B(i;\alpha,\gamma)^{S_i-r_i}}{(S_i-r_i)!} I(S_i \geq r_i)$$

$$= \sum_{\boldsymbol{S}\in\mathcal{S}_{N-R,J-r}} \frac{(J-r)!}{B(N-R;(J-r)\alpha,\gamma)} \prod_{i=0}^{N-R} \frac{B(i;\alpha,\gamma)^{S_i}}{S_i!} = 1$$

なので、一般化 CIGP 分布の階乗モメントは

$$\mathrm{E}_{gcigp}(\prod_{i=1}^{N} S_i^{(r_i)}) = \frac{J!B(N-R;(J-r)\alpha,\gamma)}{(J-r)!B(N;J\alpha,\gamma)} \prod_{i=0}^{N} B(i;\alpha,\gamma)^{r_i} \tag{4.60}$$

である。なお $\gamma=-1/2$ の場合は (4.49) 式のように書ける。特に

$$\mathrm{E}_{gcigp}(S_i) = \frac{JB(N-i;(J-1)\alpha,\gamma)}{B(N;J\alpha,\gamma)} B(i;\alpha,\gamma),$$

$$\begin{aligned}\mathrm{V}_{gcigp}(S_i) &= \frac{J(J-1)B(N-2i;(J-2)\alpha,\gamma)}{B(N;J\alpha,\gamma)} B(i;\alpha,\gamma)^2 \\ &\quad + \frac{JB(N-i;(J-1)\alpha,\gamma)}{B(N;J\alpha,\gamma)} B(i;\alpha,\gamma) \\ &\quad + \{\frac{JB(N-i;(J-1)\alpha,\gamma)}{B(N;J\alpha,\gamma)} B(i;\alpha,\gamma)\}^2.\end{aligned}$$

一般化 CIGP 分布 (4.58) は $\gamma=-1$ の場合、(条件付き) ポアソンモデルになり、寸法指標の同時確率は

$$\mathrm{P}_J(S_0,S_1,\ldots,S_N|N) = \frac{J!N!}{J^N} \prod_{i=0}^{N} (\frac{1}{i!})^{S_i} \frac{1}{S_i!} \tag{4.61}$$

のように書ける。これは等確率多項分布に他ならない。(4.61) 式がディリクレ=多項分布 (4.2) に対するある極限操作の結果として得られる事を確認しておこう。ディリクレ=多項分布は、一般化 CIGP 分布について $\gamma\to 0\,(\alpha\gamma=\nu)$ として得られる。そして更に $-\nu\to\infty$ という極限をとれば、(4.61) になる。なお N がベキ級数母数 $\alpha\theta$ の十分統計量なので、(4.61) 式は α,θ に依存せず、柔軟なモデルではない。ただし寸法指標の階乗モメントは明示的に評価出来て、

$$\mathrm{E}_{mult}(\prod_{i=1}^{N} S_i^{(r_i)}) = \frac{J!N!(J-r)^{N-R}}{(J-r)!(N-R)!J^N} \prod_{i=0}^{N} (\frac{1}{i!})^{r_i} \tag{4.62}$$

となる (等確率多項分布のモメント)。

一般化極限 CIGP 分布 (4.59) は、γ を定数とみなせば指数族に所属し、その十分統計量は U である (Remark 2, Hoshino[82])。また $\gamma=-\alpha$ と置けば、(4.59) は

$$\mathrm{P}(S_1,S_2,\ldots,S_N|N) = \frac{(\alpha A)^U}{B(N;A,-\alpha)} \prod_{i=1}^{N} \{\frac{(1-\alpha)^{[i-1]}}{i!}\}^{S_i} \frac{1}{S_i!} \tag{4.63}$$

と書き直すことが出来る。特に $\alpha = 1/2$ (LCIGP 分布) の場合、

$$\mathrm{P}(S_1,\ldots,S_N|N) = \frac{N!A^{U_N-N}}{\sum_{l=1}^{N} C(N,l,\frac{1}{2})2^N(\frac{-1}{A})^{N-l}} \prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i}\frac{1}{S_i!},$$

ただし $C(N,U,\alpha)$ は C-ナンバー (Charalambides[31]) である。C-ナンバーは一般化されたスターリング数であり、付録 B で簡単に性質が述べられている。このように書けば、LCIGP 分布の定義式 (4.54) における変形ベッセル関数は、N を分割する組み合わせを数え上げた結果である事が一層明らかとなる。(4.63) 式を Pitman 分布 (4.46) と比較すれば、(4.59) の U 所与の条件付分布が、Pitman 分布のそれと同じ事が分かる。厳密にこれを示すには、U の確率関数を評価すれば良い。Pitman[136]、Yamato et al.[228] が Pitman 分布の下での U の挙動を評価 (詳しくは大和 [226] を見よ) しており、これ等と同様に考えることが出来る。すなわち

$$\frac{(\alpha A)^U}{B(N;A,-\alpha)}\prod_{i=1}^{N}\{\frac{(1-\alpha)^{[i-1]}}{i!}\}^{S_i}\frac{1}{S_i!} = \frac{A^U}{B(N;A,-\alpha)}(-1)^{N-U}\prod_{i=1}^{N}\binom{\alpha}{i}^{S_i}\frac{1}{S_i!}$$

なので、Charalambides and Singh[35] の (3.24) 式 (付録の (B.4) 式を見よ) から

$$\mathrm{P}_{glcigp}(U) = \frac{A^U}{B(N;A,-\alpha)}(-1)^{N-U}\frac{1}{N!}C(N,U,\alpha), \quad U=1,2,\ldots,N. \tag{4.64}$$

特に $\gamma = -1/2$ の場合、Hoshino[82] によれば

$$\begin{aligned}\mathrm{P}_{lcigp}(U) &= \sqrt{\frac{\pi}{2A}}\frac{\exp(-A)}{K_{N-1/2}(A)}(\frac{1}{2A})^{N-U}\frac{(2N-U-1)!}{(U-1)!(N-U)!}\\ &= \frac{1}{\sum_{l=1}^{N}(\frac{-1}{A})^{N-l}C(N,l,\frac{1}{2})}(\frac{-1}{A})^{N-U}C(N,U,\frac{1}{2}), \quad U=1,2,\ldots,N,\end{aligned}$$

となる。また $\gamma \to 0$ の場合は Ewens モデルであり、この場合

$$\mathrm{P}_{Ewens}(U) = |s(N,U)|\frac{A^U}{A^{[N]}}, \quad U=1,2,\ldots,N,$$

ただし $|s(N,U)|$ は符号なし第一種スターリング数である (Ewens[51])。

結局、一般化極限 CIGP 分布 (4.59) の下では

$$\mathrm{P}(S_1,S_2,\ldots,S_N|U,N) = \frac{N!}{C(N,U,\alpha)}\prod_{i=1}^{N}\binom{\alpha}{i}^{S_i}\frac{1}{S_i!}, \tag{4.65}$$

ただし $\alpha = -\gamma$ となる。特に $\alpha = 1/2$ の場合は (4.56) 式で与えられた。参考の為に、Pitman 分布 (4.46) の下での U の分布は Pitman[136] によれば

$$\mathrm{P}_{Pitman}(U) = \frac{\theta^{[U:\alpha]}}{\theta^{[N]}}(-1)^{N-U}C(N,U,\alpha)\alpha^{-U}, \quad U=1,2,\ldots,N,$$

と書ける。従って Pitman 分布の下でも (4.65) が成立する。ちなみに $\sum_{U=1}^{N}\mathrm{P}(U)=1$ を利用すれば、(4.64) から

$$B(N;A,\gamma) = \sum_{U=1}^{N}A^U(-1)^{N-U}\frac{1}{N!}C(N,U,-\gamma) = \sum_{U=1}^{N}A^U\frac{1}{N!}c(N,U,-\gamma), \tag{4.66}$$

ただし $c(N,U,-\gamma)$ は符号なしの C-ナンバーである（付録の (B.6) 式参照)。ここで

$$C_{N\alpha}=\sum_{U=0}^{N}C(N,U,\alpha)=\sum_{U=1}^{N}C(N,U,\alpha)$$

と書けば、

$$N!B(N;-1,-\alpha)=\sum_{U=1}^{N}(-1)^{U}c(N,U,\alpha)=(-1)^{N}C_{N\alpha}$$

である。Charalambides[32][33] は $C_{N\alpha}$ について $N\to\infty$ と $\alpha\to\pm\infty$ の極限を明らかにしている。しかしこれらの極限は、いずれも我々の問題意識では興味の対象にならない。なお $B(i;\alpha,\gamma)$ は、一般化された IGP 分布 (4.57) をベキ級数分布と見て、ベキ級数母関数

$$\eta(\theta)=\exp(\alpha(1-(1-\theta)^{-\gamma}))$$

を θ について展開した係数である。すなわち

$$\exp(\alpha(1-(1-\theta)^{-\gamma}))=\sum_{i=0}^{\infty}B(i;\alpha,\gamma)\theta^{i}.$$

このように $\eta(\theta)$ を $B(i;\alpha,\gamma)$ の母関数とみなした展開は、Carlitz[28] の (3.13) 式に見られる。また Charalambides and Singh[35] の (3.19) 式（付録の (B.5) 式を見よ）を利用すれば、直接 (4.66) を得られる。

一般化極限 CIGP 分布 (4.59) の下で、寸法指標の階乗モメントは一般化 CIGP 分布の場合と同様に評価出来る。非負整数の $r_i, i=1,2,\ldots,N$, について $r=\sum_{i=1}^{N}r_i$, $R=\sum_{i=1}^{N}ir_i(\le N)$, とする。この時

$$\begin{aligned}\mathrm{E}(\prod_{i=1}^{N}S_i^{(r_i)}) &= \sum_{\boldsymbol{S}\in\mathcal{S}_N}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^{U}\frac{1}{B(N;A,\gamma)}\prod_{i=1}^{N}\{\frac{\Gamma(i+\gamma)}{i!}\}^{S_i}\frac{1}{(S_i-r_i)!}I(S_i\ge r_i)\\ &= \frac{B(N-R;A,\gamma)}{B(N;A,\gamma)}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^{r}\prod_{i=1}^{N}\{\frac{\Gamma(i+\gamma)}{i!}\}^{r_i}\times\\ &\quad\sum_{\boldsymbol{S}\in\mathcal{S}_N}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^{U-r}\frac{1}{B(N-R;A,\gamma)}\prod_{i=1}^{N}\{\frac{\Gamma(i+\gamma)}{i!}\}^{S_i-r_i}\frac{1}{(S_i-r_i)!}I(S_i\ge r_i),\end{aligned}$$

ただし

$$\mathcal{S}_N=\{(S_1,S_2,\ldots,S_N)|\sum_{i=1}^{N}iS_i=N\}$$

である。ここで

$$\begin{aligned}&\sum_{\boldsymbol{S}\in\mathcal{S}_N}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^{U-r}\frac{1}{B(N-R;A,\gamma)}\prod_{i=1}^{N}\{\frac{\Gamma(i+\gamma)}{i!}\}^{S_i-r_i}\frac{1}{(S_i-r_i)!}I(S_i\ge r_i)\\ &=\sum_{\boldsymbol{S}\in\mathcal{S}_{N-R}}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^{U}\frac{1}{B(N-R;A,\gamma)}\prod_{i=1}^{N-R}\{\frac{\Gamma(i+\gamma)}{i!}\}^{S_i}\frac{1}{(S_i)!}=1\end{aligned}$$

なので一般化 LCIGP 分布の階乗モメント

$$\mathrm{E}_{glcigp}(\prod_{i=1}^{N} S_i^{(r_i)}) = \frac{B(N-R;A,\gamma)}{B(N;A,\gamma)}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^r \prod_{i=1}^{N}\{\frac{\Gamma(i+\gamma)}{i!}\}^{r_i}$$

が示された。ここで $\gamma=-1/2$ の場合は (4.55) 式を見よ。特に

$$\mathrm{E}_{glcigp}(S_i) = \frac{B(N-i;A,\gamma)}{B(N;A,\gamma)}\frac{-\gamma A}{\Gamma(\gamma+1)}\frac{\Gamma(i+\gamma)}{i!},$$

$$\begin{aligned}\mathrm{V}_{glcigp}(S_i) &= \frac{B(N-2i;A,\gamma)}{B(N;A,\gamma)}\{\frac{-\gamma A}{\Gamma(\gamma+1)}\}^2\{\frac{\Gamma(i+\gamma)}{i!}\}^2 \\ &+\frac{B(N-i;A,\gamma)}{B(N;A,\gamma)}\frac{-\gamma A}{\Gamma(\gamma+1)}\frac{\Gamma(i+\gamma)}{i!} \\ &+\{\frac{B(N-i;A,\gamma)}{B(N;A,\gamma)}\frac{-\gamma A}{\Gamma(\gamma+1)}\frac{\Gamma(i+\gamma)}{i!}\}^2.\end{aligned}$$

一般化極限 CIGP 分布 (4.59) は $\gamma=-1$ の場合、退化してしまう。拡張負の二項モデルで $\gamma=-1$ と出来なかった事に注意すべきだろう。一般化 CIGP 分布で $\gamma=-1$ とおいた場合の多項分布 (4.61) は

$$\frac{J!}{(J-U)!J^N}N!\prod_{i=1}^{N}(\frac{1}{i!})^{S_i}\frac{1}{S_i!}$$

のように書き直す事が出来る。ここで $J\to\infty$ とすれば、$U=N$ でない限り退化する事が分かる。

U 所与の条件付一般化極限 CIGP 分布または条件付 Pitman 分布 (4.65) について階乗モメントを確認しておこう。$\alpha=1/2$ の場合を既に 4.11.2 節で求めており、同様に評価出来る。すなわち非負整数の $r_i, i=1,2,\ldots,N$, について $r=\sum_{i=1}^{N} r_i(\le U)$, $R=\sum_{i=1}^{N} ir_i(\le N)$ とする。この時

$$\mathrm{E}(\prod_{i=1}^{N} S_i^{(r_i)}|U,N) = \frac{N!C(N-R,U-r,\alpha)}{(N-R)!C(N,U,\alpha)}\prod_{i=1}^{N}\binom{\alpha}{i}^{r_i}$$

と書ける。特に

$$\mathrm{E}(S_i|U,N) = \frac{N!C(N-i,U-1,\alpha)}{(N-i)!C(N,U,\alpha)}\binom{\alpha}{i}, \quad i=1,2,\ldots,N,$$

となる。

命題 4.2 と命題 4.4 から類推して、一般化極限 CIGP 分布の最初の m 要素である $(S_1,S_2,\ldots,S_m)$ は $N\to\infty$ の時 $i=1,2,\ldots,m$ について独立な平均

$$-\gamma A\frac{\Gamma(i+\gamma)}{\Gamma(1+\gamma)\Gamma(i+1)}$$

のポアソン分布の同時分布に収束する事が予想される。条件

$$\lim_{N\to\infty}\frac{B(N-R;A,\gamma)}{B(N;A,\gamma)} = 1$$

が満たされればこの予想は証明されるが、今後の課題とする。

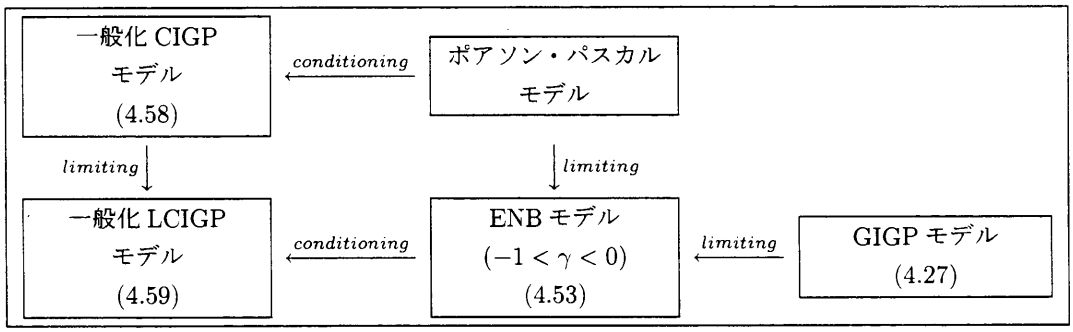

図 4.1: ポアソン・パスカル分布によるモデリング (Hoshino[82])

一般化 CIGP 分布と一般化極限 CIGP 分布に関する母数推定は、未だ議論がなされていない。そして残念ながら、簡潔になりそうにない。特に一般化極限 CIGP 分布は U 所与での挙動が Pitman 分布と一致するので、寸法指標の推定値が Pitman 分布のそれと似る事が予想される。Pitman 分布の最尤推定は簡単なので、苦労して一般化極限 CIGP 分布をあてはめる意味は無いかもしれない。本節の考察は、理論的な興味の範囲を出ない可能性がある。

第5章　超母集団モデルの応用

本節では寸法指標推測の応用問題について考察する。主な応用分野は統計的生態学、計量言語学、統計的開示制限だが、前二者についてはリファレンスとも言うべき文献が存在する。従って統計的生態学（5.1節）と計量言語学（5.2節）に関する記述は、簡単なものにとどまる。それよりも、統計的開示制限の文脈を詳しく説明しよう（5.3節）。また本節の最後（5.4節）では、実データにモデルを当てはめた数値例を示す。

5.1　統計的生態学

統計的生態学では、種の "abundance" が興味の対象となる。これは必ずしも個体数を意味せず、Tokeshi[204] によれば、ニッチ（生態学的地位）やバイオマス等にも用いられる。この "abundance" を記述するモデルとして、混合ポアソン分布が重要な役割を果たしてきた。ただし近年、統計的生態学において abundance モデルは余り注目を浴びていないようだ。比較的新しい教科書（Ludwig and Reynolds[111] や Young and Young[230]）によれば、abundance を計る究極の目的は生態学的共同体の構造に関する仮説を立てる事にある。しかし個体数の表面的挙動から実体的構造を同定する能力が不足している（例えば混合ポアソンかつ複合ポアソンとなる分布ではどちらが構造式だろうか?）ので、abundance モデルをあてはめて得られる洞察は限られる。従って、実体への興味が優越する以上、計数データのモデリングと異なるアプローチが重要性を増すのは自然な事である。なお群集生態学の問題意識を知る為に、木本・武田 [101] を参考にした。

生態学分野における混合ポアソン分布の解釈を確認しよう。まず伝統的にポアソン分布は、単位時間あたりの標本抽出機構の記述とみなされる。J 種が存在するとして、j 番目の種の abundance を λ_j と書く。観測された第 j 種の標本数 f_j は、平均 λ_j のポアソン分布からの実現値とみなす。そして $(\lambda_1, \ldots, \lambda_J)$ は、連続な確率分布に従うと考える。

ただ abundance の分布は、連続ではおかしいという議論が有る。Holgate[77] 等の主張によれば、連続分布が表す無限母集団からの抽出というモデリングは、種の数が有限だという事実に「矛盾」する。ただモデルを実機構の記述とみなさなければ、このような批判は意味を持たない。また λ で表される abundance を、生物群集の利用する資源量（ニッチ）のように母集団個体数を間接的に規定する量と考えれば、連続分布で記述しても構わないだろう。しかし有限の集団を無限とみなす事への批判は無視できない。λ は一単位時間当たりの観測数の平均なので、t 期間観測すれば、$t\lambda$ が観測個体数の平均とされる。しかし長く観測すれば、追加的に観測される標本は徐々に減ってゆくのではないだろうか。つまり t が大きい時、観測個体数の平均は $t\lambda$ を下回ると思われる。このような食い違いは、有限母集団を無限母集団で近似しているから起きる。ポアソン分布を抽出メカニズムと解釈して良いのは、限られた場合だろう。

本稿の方法論では、混合ポアソンの λ の分布に実機構もしくはニッチのような理論上の解釈は無い。そしてポアソン分布は、二項分布の近似として無限母集団からの抽出に対応する。そして観測の為の抽出をポアソン分布と切り離すことで、例えば非復元無作為抽出と整合的なモデルを用意できる。母集団寸法指標を標本から推測する場合、無限母集団からの抽出とは別に標本設計を考えた方が目的合理である。一方、生態学ではデータの記述が重視されてきたように見える。

なお生態学では繰り返し標本を取る事が珍しくないので、推測を $\lambda_1, \ldots, \lambda_J$ 所与ですべきか否かが問題になる。つまり標本が、$\mathrm{P}(f_j|\lambda_j)$ か $\mathrm{P}(f_j)$ のどちらに従うかが問題になる。Engen[48] によれば、生態学の論文ではどちらの場合も有るが、条件付の方が現実に近いと主張されている。ただ本稿では、複数標本からの推測は議論しない。

Fisher et al.[56] がマレー半島の蝶のデータに対数級数分布を当てはめて以来、統計的生態学では様々な分布が abundance の記述に用いられてきた。なお古典的応用例については、Engen[49] が詳しい。中でも重要なモデルは Preston[138] の対数正規＝ポアソン分布である。Dennis and Patil[42] は、対数正規分布の生態学での応用に関するレビューを与えている。一般に大標本では、対数級数分布よりも対数正規曲線が良くあてはまると考えられているようだ。なお Kempton and Taylor[98] は対数級数分布と対数正規＝ポアソン分布を様々な生物データにあてはめて比較した。そこでは種の構成が経年変化しない静態的環境では対数級数分布が、動態的環境では対数正規＝ポアソン分布のあてはまりが良かった事を報告している。Myers and Pepin[122] は、対数正規、ワイブル、ガンマ分布の生物データへのあてはまりを尤度比統計量、残差 χ^2 統計量で比較した。彼らのデータでは、ワイブルやガンマが良いケースも多い。特に昆虫の先行研究で、対数正規が必ずしも良くないケースをあげている。また同論文では、観測値が 0 の種を含むような対数正規分布（Δ-分布）について、観測値非零の種が対数正規性を持つという仮定に関する頑健性をシミュレーションで評価している。具体的には $q-p$ 個の対数正規標本と p 個の他分布標本を混合、q 個の標本から推定を行う。ここで言う他分布とは、所与の対数正規分布と同じ期待値・分散を持つワイブル、ガンマ、または標準正規乱数 u を変換した

$$\theta[u\sigma + \sqrt{(u\sigma)^2+1}]^2$$

の分布である。ただし σ, θ は正の母数とする。これらは対数正規の母数空間の大部分で形状が似ているが、より歪みが少ないので選択された。推定量の標本平均と標本分散について相対バイアスと効率を評価した結果、対数正規を前提にした推定量は仮定からの乖離に大変敏感であった。

他に比較的新しい応用を挙げておく。Ord and Whitmore[129] は、（切り落とし）IGP 分布をいくつかの生物データセットにあてはめている。また Sichel[180] は、（切り落とし）GIGP 分布のあてはまりが良好と主張している。

なお近年の abundance 推測の議論については、Schwarz and Seber[155] のレビュー論文が詳しい。生態学分野の標本設計（確率抽出法）は複雑な場合が多く、同論文のまとめ方は標本設計別に文脈を整理するというものである。なお超母集団モデルによる推測についてはほとんど触れられていない。また新しい文献として、Haas and Stokes[69] を挙げておく。彼らはジャックナイフを使って U の推定を試みている。ただ推定量を簡素化する為に母集団に関する仮定（$S_{N/U} = U$）を用いており、分析の意義に疑問が残る。

5.2 計量言語学

計量文献学及び計量言語学等においては、特定の単語が文章中に何度現れたかという計数データを取り扱う。これらの分野の問題意識は、竹内他 [200] における解説が研究史も含めて分かり易い。なお初期の議論は、Yule[232] に尽きる。また近年の成果についても、Baayen[9] が議論をまとめている。従って本節で新しく述べるべき事は、ほとんど無い。

これらの分野での寸法指標推測応用例を説明しておこう。ある言語の語彙数が J 語であるとする。ここで第 j 語が、抽出されたある作家の文章中に f_j 個出てくると考える。例えば母集団は、抽出された文章を含む作品全体と考えられる。一度だけ使われる単語（文法用語で「臨時語」）は "hapax legomenon (pl. -na)" と呼ばれるが、その母集団中の総数 S_1 は興味の対象となる。なお二度使われる言葉は "dis legomena"、三度は "tris legomena" である。そして作家が長く書けば知っている全ての言葉を使うとして、その語彙数は $\lim_{N\to\infty} U$ である。

統計的生態学と計量言語学の関連は古くから意識されていた。故に混合ポアソン分布の使われ方も似ている。本分野での混合ポアソン分布の平均母数 λ_j の解釈は、文章の単位長あたりの第 j 語出現頻度の平均である。文章長 N は、無限にいくらでも近づける事が出来るだろう。従って生態学のような母集団の有限性はあてはまらない。しかし語彙数 U は有限と考えられるので、$\lim_{N\to\infty} \mathrm{E}(U)$ が有限のモデルが望ましいようだ。

Williams[221] は、1900 年以前の単語長分布に関する先駆的な研究（Mendenhall[118] 等）を再発見して報告している。しかし言語データの統計学的分析は、Yule[231] がその端緒に見える。なお言語データは、右裾が重いという特徴は共有するが、モードが 1 ではない例も見られる。GIGP 分布は、このような例も記述出来ると主張されている。例えば Sichel の一連の論文 [177][178] では、GIGP 分布による語彙データの記述を議論している。また Sichel[179] は、ある分野の論文が雑誌に掲載された頻度の分布に、最小 χ^2 法を用いて 0 切り落とし GIGP 分布を当てはめている。この場合の最尤推定については、基本的に Stein et al.[192] の方法に従って、Heller[74] が検討している。また Price[139] は、語彙数の学習による増加曲線を、GIGP 分布で記述している。

5.3 統計的開示制限

官庁統計などで、集計される前の調査票を個票と呼ぶ。今のところ日本で一般公開されている官庁統計は、集計結果のみである。しかし集計により調査は本来持っている情報を一部失う為、分析の目的によっては個票データが必須となる。ただし個票レベルの分析には、「目的外申請」をする必要がある。これは非常に煩雑な手続きを要求され、誰もが実質的に可能というわけではない。また分析は、申請した範囲でしか行えない。この場合特に問題なのは、探索的データ解析が不可能な事である。このように日本の官庁統計個票データについては、利用の制限が大きい。

しかしいかなる集団にも同じデータを提供して独自解釈を保証する事は、民主主義的政治過程に不可欠だという主張が存在する（Dale et al.[41]. p.8）。この立場からは、個票への差別的アクセス制限は許容されない。また統計は利用される事に存在意義がある。多様なデータ分析を促進する事で、新しい知見を誘発する事が期待される。このような社会的利益が有るからこそ、多額の公費により統計は維持されているのである。逆に税金で賄われているからこそ、データへの要求に答えるべきである。そして統計の利用を通してデータの質に関心を持つ者が増えれば、フィードバックを

産み統計の改善にもつながるだろう。このような好循環を阻む現行制度は、問題が有ると言わざるを得ない。

個票を公開していない究極的な要因は、匿名性が確保できないと、調査が存立しえないという所にある。法的な問題もあるが、調査客体が調査に不信感を抱くと、回答拒否や虚偽回答につながる。このようなデータは利用価値がない。従って匿名性を危うくする事は、万人の利益に反する。故に匿名性を事実上確保できる範囲で、個票を公開する事を考えたい。集計結果のみ公表する事が、調査の匿名性の必要条件ではないのである。実際アメリカ・カナダでは、国勢調査の一部個票データが一般に提供されている（ヨーロッパでは国毎に対応が違うが、詳しくは Dale et al.[41] を見よ）。また各国統計当局で、個票提供について研究が積み重ねられている。これらの成果をまとめた Willenborg and de Waal[217] [219] は、基本的な文献である。他に関係論文を編集した Doyle et al.[43] も挙げておこう。邦文の解説では竹村 [197] が参考になる。より詳しくは松田他 [117] を参照すべきだろう。

ある個票が、特定の調査客体のものだと明らかになる事を、「個体識別 (identification)」または「個体開示 (disclosure)」という。「(個票) 開示リスク」とは、そのような危険性の事である。統計的開示制限分野での寸法指標推測問題は、個票開示リスクの評価に関わる。

ある統計調査で、個体の同定・識別に役立つ情報を含む項目が L 個有るとしよう。そのような項目を「キー変数」という。調査項目の中には、個体の識別に使えないものも存在する。例えば事業所の特定品目在庫などは、外からはうかがい知れない。リスク評価については、そのような項目は無視して良い。第 $l, l = 1, 2, \ldots, L$, 変数について、実現値は c_l 個のカテゴリーに分類されるだろう。例えば性別なら、男と女の二分類 ($c = 2$) になる。個票データセットは第 l キー変数が第 l フィールドを構成し、調査客体（統計単位）毎にレコードをつくる。調査客体は通常個人や世帯、事業所である。

ではどのようなデータセットが危険と考えられるか。極端なケースとして、名前や住所を含むようなデータセットが挙げられる。この場合調査客体の特定は容易であり、このようなリスクは明らかに許容出来ない。何故「名前」や「住所」をデータとして含んではならないのかというと、特定の名前や住所という条件を満たす個体が母集団において非常に少ないからである。小さいグループの中から個体を特定するのは、比較的容易かもしれない。一般に考えて、キー変数の特定の組み合わせ条件を満たす個体が少なければ、その個体群は比較的危険と考えられる。キー変数について組み合わせの総数は各カテゴリー数の積、すなわち $c_1 \times \cdots \times c_L$ であり、これを J と書く。この組み合わせのそれぞれを「セル」と呼び、各個体はいずれかのセルに所属する。母集団において第 j セルに所属する個体数を F_j と書く ($j = 1, 2, \ldots, J$)。ここで F_j が小さければ小さい程、そのセルに所属する個体データの公開はより危険という事である。

なお J 個の中には度数 F_j が必ず 0 となるセルも有り、「構造的ゼロ (structural zero)」と呼ばれる。例えば年齢 10 で配偶者有りという条件を満たす個体は存在しない。従って、真の J はナイーブに決まる J から構造的ゼロセルを引いた数だと主張される場合が有る。これを J に関する不確実性と考えれば、J に依存しないモデリングが推奨される。

直感的に考えて、調査客体の情報を粗くすればそのデータセットはより安全になる。そのようなデータ修正 (editing) については、いくつか選択肢が有る。任意の変数の組について、全てのレコードに同じ修正をする場合「大域的 (global)」という。そうでなければ「局所的 (local)」と呼ぶ。以下では大域的修正のみ考える。また修正の技術としては、「隠蔽 (suppression)」と「再符号

化 (recoding)」を考える。前者は変数の組を、公開対象から外してしまう。「名前」などは大域的隠蔽の自明な対象である。再符号化を説明しよう。これは変数のカテゴリーを再編成する事を言う。「学歴」を例にとる。「大卒」と「大学院卒」が分けて調査されていても、再符号化して両者を同一のカテゴリー、すなわち「大卒以上」で扱える。また再符号化には、数値の丸め等も含まれる。隠蔽はその変数のカテゴリー数を 1 にする事と解釈出来るので、再符号化の極端なケースと同等である。なお再符号化を進める事で、危険性は単調に減少するはずである。セルの大きさをリスクの指標とすれば、これをうまく説明出来る。つまり再符号化によるカテゴリー数 c_l の減少は、J の減少を意味する。この時、セルあたりの個体数は増える方向で変化する。

以上の議論から、セルの大きさはリスクの指標としてふさわしい性質を持っている事が理解される。従って個票データセットの各レコードが所属するセルの大きさは、開示リスクに関して重要な情報と考えられている。しかし F_j は標本調査では未知であり、標本データから推測する必要がある。標本でのセル内個体数を f_j と書く。標本サイズと母集団サイズは、それぞれ

$$\sum_{j=1}^{J} f_j = n, \quad \sum_{j=1}^{J} F_j = N$$

とする。本分野では、n, N は所与で固定して考えるのが自然である。

ここまで述べれば寸法指標推測問題との接続は明らかだろう。一章（例 5）で示したように、小さな F_j の推定は難しい。故に本分野では J 個のセル内で、どの F_j が小さいかを問題にしない。そのかわり母集団寸法指標を推測し、小さなセルの数が多ければ危険と考える。特に母集団で一意な事が知られている個体のデータが公開されると、理屈としては個体識別が可能である。故に S_1 は「母集団一意 (population unique)」と呼ばれ、リスクの指標として米国・英国の統計当局が用いている。例えば Marsh et al.[116] を見よ。また、二意な事が知られている個体のレコードが公開されたとしよう。このレコードで表される属性を持つ二者は、自分が公開されているのでなければ他方が公開されている事が分かる。論理的帰結として識別が可能なケースは、この二つである。しかし母集団で一意や二意な事が知られている個体は、非常に稀である。従って一意や二意の個体の公開が、即危険という事にはならない。最低限言えるのは、S_1 や S_2 が大きいようなデータはより危険だという事である。また S_3 以下なら即安全とも言い切れない。

このように考えると、ある母集団のリスクを寸法指標の重み付き和で表すのは自然である。すなわちリスクを

$$\sum_{i=1}^{N} w_i i S_i, \tag{5.1}$$

ただし $w_i, i = 1, \ldots . N.$ は非負、のように考えれば計測可能な概念となる。例えば Bethlehem et al.[15] は

$$\sum_{i=1}^{N} (\frac{i}{N})^2 S_i$$

の逆数を "resolution" と呼び、リスクの指標として用いている。また Greenberg and Zayatz[64] では、エントロピーから類推して

$$-\sum_{i=1}^{N} \log(\frac{i}{N}) \frac{i}{N} S_i$$

が用いられた。母集団一意のみ考慮する場合は、$w_1 = 1, w_2 = w_3 = \cdots = 0$ である。またデータが公開される個体数を m としよう。これはサブサンプリングの結果、標本サイズ（n）より小さくなる場合も考えている。この時、母集団で大きさ $F = 1$ のセルに所属する標本中の個体数は、$S_1 m/N$ と考えるのが尤もらしい。一般に、母集団で大きさ i のセルに所属する標本個体数は

$$i\frac{m}{N}S_i$$

で推定できる。カナダの統計当局が用いているリスクの指標は、母集団一意な個体の標本一意に占める割合であり（例えばSkinner et al.[188]を見よ）、同様に考えて

$$\frac{n}{N \cdot s_1}S_1$$

と推定される。私見を述べれば、リスク (5.1) の重み w_i の選択は利便性の観点から行えば良い。そして許容できるリスクの範囲は、重みの付け方に依存して決まる。推定問題として重要なのは、母集団の寸法指標を与える事である。

個票データセット公開の際は、データを編集してリスクを評価するという試行錯誤の結果として、データの粗さとリスクのバランスを取る事になる。なおリスクの許容度は、データの配布形態、情報の価値・有用性、外部データベースの充実度等、様々な要因を考慮に入れて決定される。個別の状況にひどく依存する事もあり、本稿ではこれ以上議論しない。

次に個票開示リスク推測における超母集団モデルの利用について、文脈を整理しておく。以下の議論は星野 [84] のサーベイを再録した。なお応用については S_1、すなわち母集団一意の推定が大部分をしめる。

本分野ではBethlehem et al.[15]がガンマ＝ポアソンモデル (4.11) を超母集団モデルとして利用したのが嚆矢となる。佐井 [147] は J と N_0 の大小関係について場合分けをし、ガンマ＝ポアソンモデルの挙動を解析した。また佐井 [148] では、ガンマ＝ポアソンモデルを利用してリスク評価を議論している。しかし、ガンマ＝ポアソンモデルの評判は芳しくない。Zayatz[233]はガンマ＝ポアソンモデルをコルモゴロフ＝スミルノフ検定にかけたが、有為に当てはまりを欠くと報告している。Skinner et al.[188]はガンマ=ポアソンモデルが S_1 を過小推定するとし、モデルの限界ではないかと指摘した。小さなセルが支配的なデータセットは、経験的にガンマ＝ポアソンモデルで記述しきれないようである。Hoshino[80] は Pitman モデルの含意として、ガンマ＝ポアソンモデルは安全 (つまり $S_1/U \doteq 0$) なデータセットしか記述できないだろう、と述べている。一意な個体割合が無視できないデータセットが興味の対象だとすれば、ガンマ＝ポアソンモデルが不十分という評価を受けているのは納得できる。それからガンマ＝ポアソンモデルを修正する試みを紹介する。確率変数 X が負の二項分布に従う時、Chen and Keller-McNulty[36] は $X + 1$ の分布を Slide Negative Binomial と呼ぶ。彼らは F_j の分布を SNB と仮定、S_1 の推定が改善されたと報告する。佐井・竹村 [149] では、セル間に相関が有る場合を考察した。

他に当然、ガンマ＝ポアソン以外のモデルも使用されている．Skinner and Holmes[185] は、対数級数分布 (4.42) および対数正規＝ポアソン分布 (4.17) を、米国及びイタリアのセンサスデータへ当てはめている．ここでは対数正規＝ポアソン分布の当てはまりが良好であった．なお Skinner and Holmes[186] は、対数正規＝ポアソンモデル (4.19) の母数を対数線形モデルで記述する事を考察している．Hoshino and Takemura[85] は対数級数モデル (4.43) を再発見し、母数推測を議論した．また Takemura[199] は、ディリクレ＝多項モデル (4.2) を提案した．Omori[126] はディリクレ＝多

項モデルを仮定し、母集団一意の事後確率を議論した. Ewens モデル (4.44) は Takemura[199] で言及されているが、同分布の使用は Samuels[151] にも見られる。Hoshino[80] は対数正規＝ポアソンモデル、Ewens モデル、ディリクレ＝多項モデル、Pitman モデル (4.46) を 1995 年の労働力調査のデータにあてはめた。AIC による比較では、Pitman モデルが優越する結果となった。Hoshino[83] では CIGP モデル (4.48) を、佐井・竹村 [149] が用いた労働力調査データ（詳細は例 5.3 を見よ）にあてはめている。

5.4 数値例

本節では統計的生態学、計量言語学、個票開示リスク評価分野の実データに対し、超母集団モデルを当てはめてみる。また同一母集団から繰り返し標本を抽出し、母集団一意 (S_1) の推定値の変動を数値的に確認しよう。そして実験結果に対する考察を、本論文の結語に代える。

例 5.1 しらみのデータ (Williams[222])

Williams[222] は、一人の頭にしらみが何匹観測されたかという計数データを、寸法指標の形で与えている。観測対象の人数は $J = 1083$ であり、しらみが一匹以上観測された人数は $u = 461$、観測されたしらみの総数は $n = 7442$ であった。

Stein et al.[192] は、このデータに IGP 分布を当てはめた。Hoshino[83] は、同データに関する CIGP 分布のあてはまりが IGP 分布と同等であると報告した。しかしこれらのあてはまりは悪く見える。

同データの s_0 を無視して ENB モデルを最尤法によってあてはめた所、$\hat{\gamma}_{ENB} = 0.003801, \hat{\theta}_{ENB} = 0.985454$ となった。AIC は 383.53 である。推定された γ_{ENB} はわずかながら正であり、$\gamma_{ENB} = -1/2$ の場合に相当する IGP 分布や CIGP モデルのあてはまりが悪い事が理解できる。次に $\gamma_{ENB} \to 0$ の場合に相当する Ewens モデルをあてはめた。最尤法により $\hat{\theta}_{Ewens} = 108.56$ を得た。AIC は 366.55 となり、ENB モデルより好ましい。これらの推定値の下で計算した $\mathrm{E}(s_i), i = 1, 2, \ldots, 30,$ と元データを、表 5.1 にまとめてある。ENB モデルと Ewens モデルのあてはまりは、かなり近く見える。

このデータで試しに Pitman モデルの最尤解を求めようとしたが、$\alpha > 0$ の領域で探索して収束しなかった。おそらく ENB モデルを当てはめた場合の $\hat{\gamma}$ が負になるデータでなければ、$\alpha > 0$ に解が無いと思われる。なおディリクレ＝多項モデルをあてはめると、母数の推定値 $\hat{\gamma}_{DM}$ は無限大となり、多項分布が選択される。—

例 5.2 名詞の出現回数 (Yule[232])

Thomas à Kempis による "*De Imitatione Christi*" というテキスト（ラテン語）において、i 回出現した名詞の数を Yule[232] (p.10) は寸法指標 $s_i, i = 1, 2, \ldots, 418,$ で与えている。名詞の出現回数の総和 n は 8225、出現した名詞の総数 u は 1168 であった。

まず ENB モデルを最尤法であてはめた所、母数の推定値として $\hat{\gamma}_{ENB} = -0.380079, \hat{\theta}_{ENB} = 0.987631$ を得た。AIC は 377.37 である。γ_{ENB} の推定値が負なので、Pitman モデルを当てはめてみよう。最尤法の結果、$\hat{\alpha}_{Pitman} = 0.379796. \hat{\theta}_{Pitman} = 103.083368$ を推定値として得た。AIC

i	s_i	Ewens	ENB	i	s_i	Ewens	ENB	i	s_i	Ewens	ENB
1	106	107.01	106.53	11	3	8.43	8.46	21	3	3.82	3.84
2	50	52.74	52.69	12	10	7.61	7.64	22	4	3.60	3.61
3	29	34.66	34.68	13	8	6.93	6.95	23	4	3.39	3.40
4	33	25.62	25.66	14	6	6.34	6.37	24	4	3.20	3.21
5	20	20.21	20.25	15	3	5.83	5.86	25	3	3.03	3.04
6	14	16.60	16.64	16	6	5.39	5.41	26	4	2.87	2.88
7	12	14.03	14.07	17	7	5.00	5.02	27	6	2.73	2.74
8	18	12.10	12.14	18	4	4.66	4.67	28	2	2.59	2.60
9	11	10.60	10.64	19	7	4.35	4.36	29	4	2.47	2.47
10	11	9.40	9.44	20	7	4.07	4.09	30	1	2.35	2.36

表 5.1: しらみのデータ (例 5.1)

i	s_i	Pitman	ENB	i	s_i	Pitman	ENB	i	s_i	Pitman	ENB
1	520	540.39	540.18	11	10	12.38	12.36	21	5	4.43	4.42
2	174	165.51	165.36	12	9	10.83	10.80	22	1	4.10	4.09
3	111	88.29	88.19	13	11	9.56	9.54	23	1	3.81	3.80
4	70	57.12	57.05	14	5	8.51	8.49	24	7	3.54	3.53
5	37	40.35	40.79	15	4	7.63	7.61	25	2	3.31	3.30
6	33	31.07	31.02	16	7	6.89	6.87	26	1	3.09	3.08
7	20	24.64	24.60	17	7	6.25	6.23	27	4	2.90	2.89
8	28	20.13	20.10	18	4	5.70	5.69	28	3	2.72	2.71
9	11	16.84	16.81	19	5	5.22	5.21	29	2	2.56	2.55
10	14	14.34	14.31	20	2	4.80	4.79	30	3	2.41	2.41

表 5.2: 名詞の出現回数 (例 5.2)

は 363.79 となり、ENB モデルのそれを優越する。理論通り、$\hat{\gamma}_{ENB} \doteq -\hat{\alpha}_{Pitman}$ となっている。これらのモデルのあてはまりと元データ（一部）は、表 5.2 にまとめてある。やはり ENB モデルと Pitman モデルが、同等の $\mathrm{E}(s_i)$ を生ずるように見える。—

例 5.1 と 5.2 では、ENB モデルが N 所与のモデルに AIC で負けている。N 所与のモデルでは裾が S_N で打ち切られるのに対し、畳み込みポアソン分布モデルでは $\{S_i\}_{i=N+1}^{\infty}$ の尤度への寄与も無視できないのではないか。

例 5.3 労働力調査 (佐井・竹村 [149])

佐井・竹村 [149] は 1997 年 12 月に集められた労働力調査のデータを匿名化し、寸法指標データを算出した。中でも山形県のデータについて、標本数は $n = 908$、総セル数は $J = 5.644 \times 10^{12}$ であった。

既に述べたように、Hoshino[83] が CIGP モデルを同データにあてはめている。そこでは言及されていないが、AIC は 39.35 であった。今回 s_0 を無視して Pitman モデルをあてはめたところ、最尤推定値として $\hat{\alpha}_{Pitman} = 0.813124, \hat{\theta}_{Piman} = 650.343448$ を得た。AIC は 31.76 となる。従って母数の数の違いを考慮しても、Pitman モデルが CIGP モデルに優越する。元データとこれらのモデルの最尤推定値の下での $\mathrm{E}(s_i), i = 1, 2, \ldots, 6$, および $\sum_{i=7}^{908} \mathrm{E}(s_i)$ が、表 5.3 にまとめられている。確かに Pitman モデルが、より望ましいように見える。—

例 5.1 から 5.3 までは、各モデルの標本データへのあてはまりを比較してみた。これらの例では（Ewens モデルを特殊ケースとして含む）Pitman モデルが良くあてはまるように見える。今度は Pitman モデルを用いて、超母集団モデルによる母集団寸法指標の推測を、数値的に評価してみよう。例 5.2 のデータを、仮に母集団と考える（つまり $N = 8225$ とする）。そして非復元単純無作為抽出で得た標本寸法指標データから、母集団寸法指標 $S_1 = 520$ を推測してみる。標本数は $n = 100, 400, 800. 1600, 6400$ とし、それぞれ 1000 回ずつ抽出を行って Pitman モデルによる S_1 の推定値 $(\hat{S_1})$ の分布を見る。表 5.4 に、それぞれ五数要約値を示す。図 5.1 から 5.5 は、各度数分布のヒストグラムである。

これらの実験で推定値 $\hat{S_1}$ は、真の値 520 よりも大きめに分布しているようだ。そもそも全数調査の場合でさえ、Pitman モデルは 540.39 と過大な推定値を与える。ただしバイアスは、標本数の増加につれて縮小している。現実の母集団の構造がモデル通りならばバイアスは存在しないが、そのような事は有り得ない。この実験が示唆するのは、応用の際、標本抽出率に依存して推定バイアスが変化するかもしれないという事である。また標本抽出率が低いと、推定はモデルを用いても不安定に見える。これは抽出による標本寸法指標の変動が非常に大きい事が原因だろう。モデルによる推定は、標本寸法指標の変動が小さければ安定的（標本数が 6400 の場合を見よ）である。しかし標本寸法指標自体が母集団寸法指標に比べて「極端」ならば、大きく外れるように見える。

1.2 節で議論したように、Neyman は確率抽出を母集団に対する偏見からの防御と位置づけた。しかし寸法指標推測問題では、母集団に関する先見情報を利用しない限り、実用的な結果が得られない。故に予測アプローチを採用し、先見情報を超母集団モデルで記述するという事であった。そこでは尤度としてモデルの良し悪しが表れる。一方伝統的な有限母集団解析では、標本設計の工夫が興味の焦点となる。そして標本設計上の工夫は、母集団を推定する際に尤度の差として現れない。その為予測アプローチと親和的な尤度原理の立場からは、そのような工夫は無意味である。

しかしこのように一度否定された標本設計は、その重要性をやはり否定できない。数値実験で観測されたように、「極端」な標本からの推測は極端な結論につながる。母集団寸法指標の構造を保持した平均的標本を得る為の設計は、極めて大切である。

超母集団モデルを用いて母集団寸法指標を推測する場合、標本は所与とした。もし標本が所与ならば、抽出機構に注意が払われなくて当然である。しかし標本が所与でなく、標本と推定値の同時分布を議論の基礎とするならば、標本設計は大きな影響を及ぼす。例えば非復元単純無作為抽出よりも「良い」設計が用いられていれば、本節の数値実験で $\hat{S_1}$ のレンジは狭くなっていたはずだ。（ただし $\hat{S_1}$ の導出は、非復元単純無作為抽出に依存していた事に注意すべきだろう。厳密に考えれば、用いられた設計に応じた $\hat{S_1}$ を求めるべきかもしれない。しかしこの場合、利便性は大きく損なわれる。）

既に考察したように、尤度原理の立場では全ての情報は確率モデルとして記述されるべきである。しかしモデルは便宜的思考であるのに、利便性を損なうような複雑化は現実的だろうか。標本設計は、確率モデルとして表現しにくい情報を利用する手段として意味がある。特に寸法指標推測問題では、標本の情報が極めて不足している。従って良い推定の為には、どのような情報ももらさず利用するしかない。標本設計の段階からコントロール出来るとして、「良い」標本が得られるような設計をすれば十分報われる事であろう。では、寸法指標推測問題において「良い」設計とは何だろうか。面白い問題と思うが、もはや展開する余力が無い。今後の研究課題とする。

i	1	2	3	4	5	6	7+	u
s_i	771	46	3	6	1	0	1	828
CIGP	760.94	56.65	8.43	1.57	0.33	0.07	0.02	828.00
Pitman	771.34	41.98	9.67	3.08	1.14	0.46	0.37	828.04

表 5.3: 労働力調査 (例 5.3)

n	100	400	800	1600	6400
最小値	77.88	283.90	73.79	397.39	519.62
第 1 四分位点	538.75	596.98	588.63	588.38	549.40
メディアン	857.53	700.73	647.35	624.27	557.93
第 3 四分位点	1396.18	823.11	713.37	659.83	566.96
最大値	6086.79	4494.14	2852.30	1195.47	600.67

表 5.4: $\hat{S}_1$ の分布 (1000 回繰り返し)

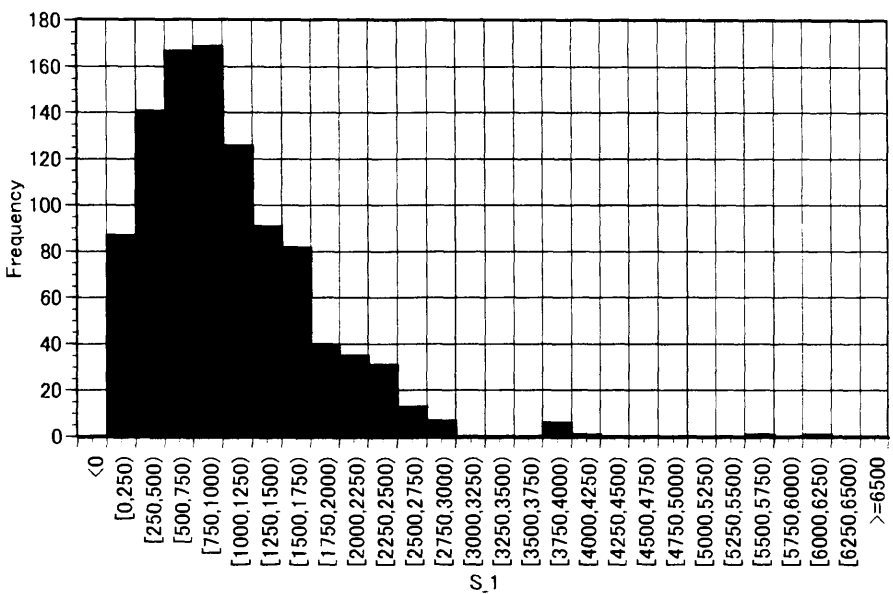

図 5.1: Pitman モデルによる $\hat{S}_1(n = 100)$

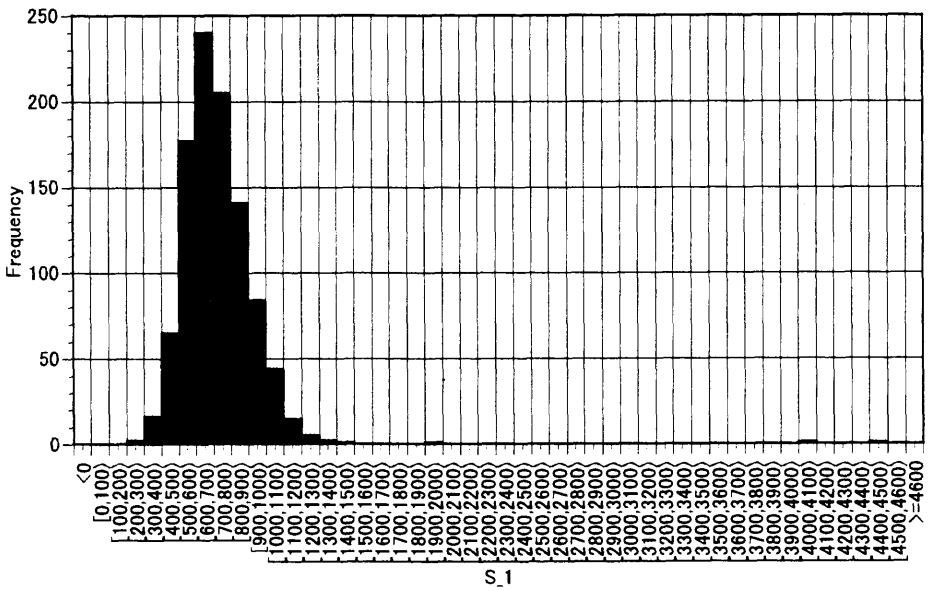

図 5.2: Pitman モデルによる $\hat{S}_1(n = 400)$

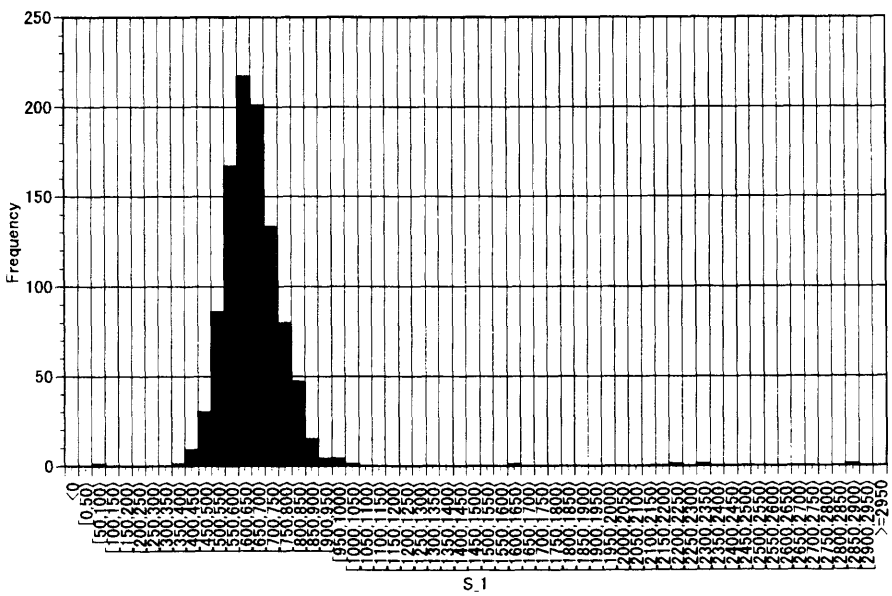

図 5.3: Pitman モデルによる $\hat{S}_1(n = 800)$

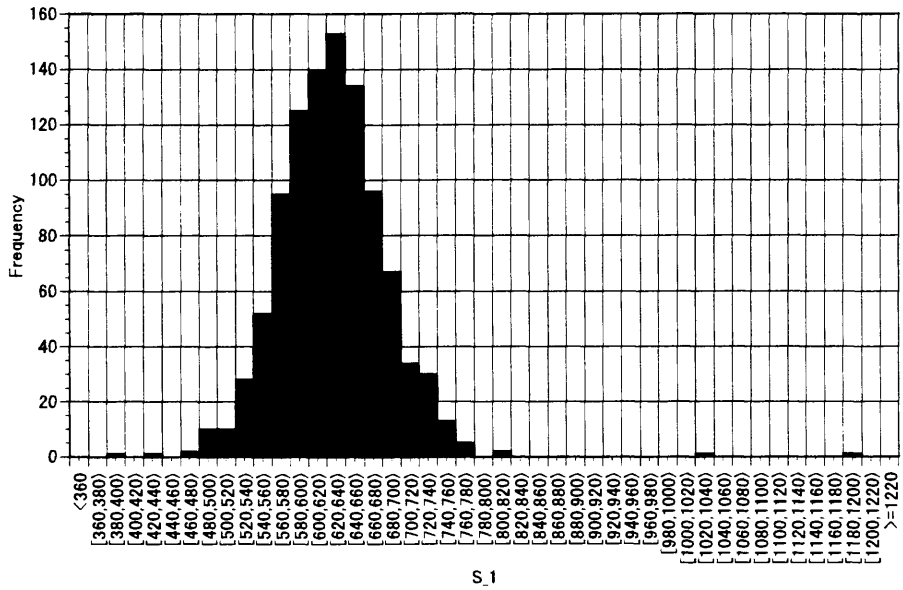

図 5.4: Pitman モデルによる $\hat{S}_1(n = 1600)$

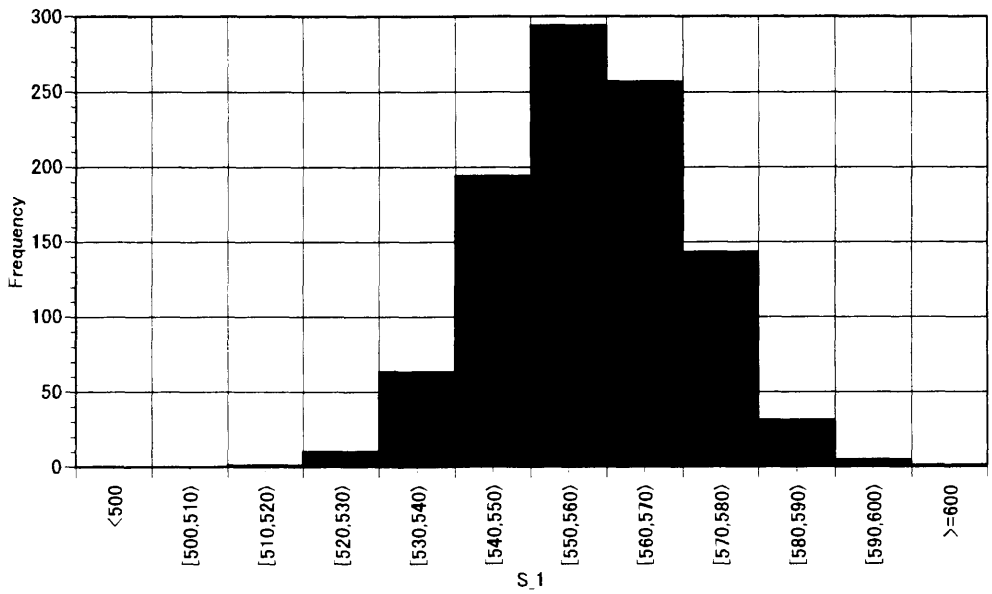

図 5.5: Pitman モデルによる $\hat{S}_1(n = 6400)$

謝辞

本論文の執筆中、竹村彰通教授に多大なる御指導を頂き、渋谷政昭教授、大和元教授からは重大な助言を頂いた。また本論文のかなりの部分は、在外研究でカーネギーメロン大学滞在中に書かれた。その際、S．ファインバーグ教授には大変便宜を図って頂いた。これらを特に記して、謝意を表したい。

付 録A　寸法指標の不偏推定量

Goodman[63] はサイズ既知の有限母集団について、U の不偏推定量 (1.7) 式を示した。以下ではその考え方を説明する。非復元単純無作為抽出を前提とした場合、

$$\mathrm{E}(s_i) = \sum_{l=i}^{q} \frac{\binom{l}{i}\binom{N-l}{n-i}}{\binom{N}{n}} S_l,$$

但し $q = \max_j F_j$ である。ここで

$$a_i = 1 - (-1)^i \frac{(N-n+i-1)^{(i)}}{n^{(i)}}$$

ならば

$$\sum_{i=1}^{j} a_i \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} = 1.$$

左辺は

$$\begin{aligned} & 1 - \frac{\binom{j}{0}\binom{N-j}{n}}{\binom{N}{n}} - \sum_{i=1}^{j} (-1)^i \frac{(N-n+i-1)^{(i)}}{n^{(i)}} \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} \\ = \; & 1 - \sum_{i=0}^{j} (-1)^i \frac{(N-n+i-1)^{(i)}}{n^{(i)}} \frac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} \\ = \; & 1 - \frac{(N-j)!}{(N-n-1)!N!} \sum_{i=0}^{j} (-1)^i \binom{j}{i} \frac{(N-n+i-1)!}{(N-n+i-j)!}. \end{aligned}$$

ここで

$$\sum_{i=0}^{j} (-1)^i \binom{j}{i} \frac{(N-n+i-1)!}{(N-n+i-j)!} = 0$$

を示すには、$a = 0, 1, 2, \ldots$ について

$$\sum_{i=j-a}^{j} (-1)^i \frac{j(N-n+i-1)!}{i!(j-i)!(N-n+i-j)!} = \frac{(N-n-1)^{(a)}(j-1)^{(a)}(N-n+j-1-a)!}{j!a!(N-n)!} (-1)^{j-a} j, \tag{A.1}$$

を示せば良い。帰納的に考えれば、以下の式を示せば良い事になる。

$$\begin{aligned} & \sum_{i=j-a-1}^{j} (-1)^i \frac{j(N-n+i-1)!}{i!(j-i)!(N-n+i-j)!} \\ = \; & \frac{(N-n-1)^{(a+1)}(j-1)^{(a+1)}(N-n+j-2-a)!}{j!(a+1)!(N-n)!} (-1)^{j-a-1} j. \end{aligned}$$

左辺は (A.1) より

$$(-1)^{j-a-1}j\{-\frac{(N-n-1)^{(a)}(j-1)^{(a)}(N-n+j-1-a)!}{j!a!(N-n)!}+\frac{(N-n+j-2-a)!}{(j-a-1)!(a+1)!(N-n-a-1)!}\}$$
$$=\frac{(-1)^{j-a-1}j(N-n+j-2-a)!(j-1)^{(a)}(N-n-1)^{(a)}}{j!(a+1)!(N-n)!}\{(j-1-a)(N-n-a-1)\}$$
$$=\frac{(N-n-1)^{(a+1)}(j-1)^{(a+1)}(N-n+j-2-a)!}{j!(a+1)!(N-n)!}(-1)^{j-a-1}j$$

となって示された。この時、

$$\mathrm{E}(\sum_{i=1}^{n}a_is_i)=\sum_{i=1}^{n}S_i$$

となる。$\sum a_is_i$ はすなわち、U の不偏推定量である。

Engen[49] の定理 2.1 は、同じく非復元単純無作為抽出の場合、寸法指標の不偏推定量を明らかにしている。すなわち、もし $i\leq n$ ならば唯一の不偏推定量が存在して (1.9) 式のように書ける。なお $i>n$ ならば S_i の不偏推定量は存在しない。証明は次の通りである。まず超幾何分布の性質より、

$$\mathrm{E}(\frac{f_j^{(k)}}{n^{(k)}})=\frac{F_j^{(k)}}{N^{(k)}}$$

なので、

$$\sum_{j=1}^{J}\mathrm{E}(\frac{f_j^{(k)}}{n^{(k)}})=\frac{1}{n^{(k)}}\sum_i i^{(k)}\mathrm{E}(s_i)=\frac{1}{N^{(k)}}\sum_i i^{(k)}S_i$$

となる。これは

$$a_k:=\frac{\binom{N}{k}}{\binom{n}{k}}\sum_{i=1}^{n}i^{(k)}\mathrm{E}(s_i)=\sum_{i=1}^{n}i^{(k)}S_i$$

と同値である。ここで解は

$$S_k=\frac{1}{k!}\sum_{j=0}^{n-k}\frac{1}{j!}a_{k+j}(-1)^j$$

で与えられるので、不偏推定量は

$$\hat{S}_k=\frac{1}{k!}\sum_{j=0}^{n-k}\frac{1}{j!}\frac{\binom{N}{k+j}}{\binom{n}{k+j}}\sum_{i=1}^{n}i^{(k+j)}s_i(-1)^j$$

のようになる。これは (1.9) 式と同等である。

付 録B　スターリング数とC-ナンバー

スターリング数は、階乗をベキ乗に展開する際に現れる係数である。有限差分の議論で階乗が占める位置を、連続関数の場合はベキ乗が占める。従ってCharalambides and Singh[35]曰く、スターリング数は有限と無限の間で微積分学の橋渡し役をつとめる。同文献はスターリング数とその一般化に関する詳細なレヴューであり、その中でC-ナンバーも説明されている。ここでは、本論文と密接に関係する部分を紹介する。

任意のtについて

$$t^{(n)} = t(t-1)(t-2)\cdots(t-n+1), \quad n = 1, 2, \ldots, \quad t^{(0)} = 1,$$

の時、

$$t^{(n)} = \sum_{k=0}^{n} s(n,k)t^k, \quad n = 0, 1, 2, \ldots,$$

$$t^n = \sum_{k=0}^{n} S(n,k)t^{(k)}, \quad n = 0, 1, 2, \ldots,$$

と書く。係数$s(n,k)$を第一種スターリング数と呼び、$S(n,k)$を第二種スターリング数と呼ぶ。また

$$t^{[n]} = t(t+1)(t+2)\cdots(t+n-1), \quad n = 1, 2, \ldots, \quad t^{[0]} = 1,$$

ならば、$t^{[n]} = (-1)^n(-t)^{(n)}$に注意すると

$$t^{[n]} = \sum_{k=0}^{n} |s(n,k)|t^k, \quad n = 0, 1, 2, \ldots,$$

である。すなわち

$$|s(n,k)| = (-1)^{n-k}s(n,k).$$

また

$$(t+1)(t+2)\cdots(t+n-1) = \sum_{k=1}^{n} |s(n,k)|t^{k-1}, \quad n = 1, 2, \ldots.$$

より、$i_j \in \{1, 2, \ldots, n-1\}$, $j = 1, 2, \ldots, n-k$, について

$$|s(n,k)| = \sum_{i_1 \neq i_2 \neq \cdots \neq i_{n-k}} i_1 i_2 \cdots i_{n-k},$$

ただし和は、正の整数$\{1, 2, \ldots, n-1\}$の全ての組み合わせ$\{i_1, i_2, \ldots, i_{n-k}\}$についてとられている。また$\{1, 2, \ldots, n-1\} \setminus \{i_1, i_2, \ldots, i_{k-1}\}$が$n-k$個の整数の組になる事に注意すれば、$i_j \in \{1, 2, \ldots, n-1\}$, $j = 1, 2, \ldots, k-1$, について

$$|s(n,k)| = (n-1)! \sum_{i_1 \neq i_2 \neq \cdots \neq i_{k-1}} \frac{1}{i_1 i_2 \cdots i_{k-1}}$$

が成立する。

$$(1+u)^t = \sum_{n=0}^{\infty} \frac{t^{(n)}u^n}{n!} = \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{s(n,k)t^k u^n}{n!} = \exp(t\log(1+u)) = \sum_{j=0}^{\infty} \frac{t^j(\log(1+u))^j}{j!}$$

なので t の係数を比較して

$$\sum_{n=k}^{\infty} s(n,k)u^n/n! = \{\log(1+u)\}^k/k!, \quad k=0,1,2,\ldots,$$

または

$$\sum_{n=k}^{\infty} |s(n,k)|u^n/n! = \{-\log(1-u)\}^k/k!, \quad k=0,1,2,\ldots, \tag{B.1}$$

を得る。同様に

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{S(n,k)t^{(k)}u^n}{n!} = \exp(tu) = ((e^u-1)+1)^t = \sum_{n=0}^{\infty} \frac{t^{(n)}(e^u-1)^n}{n!}$$

なので

$$\sum_{n=k}^{\infty} S(n,k)u^n/n! = (e^u-1)^k/k!, \quad k=0,1,2,\ldots, \tag{B.2}$$

である。(B.1) を書き換えると

$$\sum_{n=k}^{\infty} |s(n,k)|\frac{u^n}{n!} = \frac{\{\sum_{i=1}^{\infty} u^i/i\}^k}{k!}, \quad k=0,1,2,\ldots,$$

同様に (B.2) より

$$\sum_{n=k}^{\infty} S(n,k)\frac{u^n}{n!} = \frac{(\sum_{i=1}^{\infty} u^i/i!)^k}{k!}, \quad k=0,1,2,\ldots.$$

ここで u^n の係数を比較すると、

$$|s(n,k)| = \frac{n!}{k!}\sum_{r_1+r_2+\cdots+r_k=n} \frac{1}{r_1 r_2\cdots r_k},$$

および

$$S(n,k) = \frac{n!}{k!}\sum_{r_1+r_2+\cdots+r_k=n} \frac{1}{r_1!r_2!\cdots r_k!},$$

ただし $r_i \geq 1, i=1,2,\ldots,k,$ を得る。ここで $(r_1,r_2,\ldots,r_k)$ の組み合わせは、自然数 n を k 個の自然数の和で表す全ての場合を尽くしている。この事に注意して寸法指標で読み替えると、

$$|s(n,u)| = \sum_{\boldsymbol{s}\in\mathcal{S}_{n,u}} \frac{n!}{s_1!s_2!\cdots s_n!}\prod_{i=1}^{n}(\frac{1}{i})^{s_i},$$

$$S(n,u) = \sum_{\boldsymbol{s}\in\mathcal{S}_{n,u}} \frac{n!}{s_1!s_2!\cdots s_n!}\prod_{i=1}^{n}(\frac{1}{i!})^{s_i}.$$

ただし

$$\mathcal{S}_{n,u} = \{(s_1, s_2, \ldots, s_n) | \sum_{i=1}^{n} s_i = u, \sum_{i=1}^{n} i s_i = n\}.$$

を得る。

C-ナンバーは、$k = 0.1, 2, \ldots, n$, $n = 0, 1, 2, \ldots$, について実数 s（複素数でも可）をパラメータとする数であり、一般化された階乗の係数である。$C(n, k, s)$ と表記され、任意の t について

$$st^{(n)} = \sum_{k=0}^{n} C(n, k, s) t^{(k)}, \quad n = 0, 1, 2, \ldots,$$

で定義される。明らかに

$$C(0, 0, s) = 1, \quad C(n, k, s) = 0, \quad k = n, n+1, \ldots,$$

である。特に Toscano[205], Yamato et al.[228] によれば

$$C(n, k, \frac{1}{2}) = (-1)^{k-n} \frac{(2n-k-1)!}{(k-1)!(n-k)!} (\frac{1}{2})^{2n-k}$$

である。定義より

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} C(n, k, s) t^{(k)} \frac{u^n}{n!} = (1+u)^{st}.$$

この関係を利用して

$$\sum_{n=k}^{\infty} \frac{u^n}{n!} C(n, k, s) = \frac{\{(1+u)^s - 1\}^k}{k!}, \quad k = 0, 1, 2, \ldots, \tag{B.3}$$

を得る。何故なら

$$\begin{aligned}
\frac{\{(1+u)^s - 1\}^k}{k!} &= \sum_{i=0}^{k} \frac{(1+u)^{s(k-i)}}{i!(k-i)!} (-1)^i \\
&= \sum_{i=0}^{k} \frac{(-1)^i}{i!} \sum_{n=0}^{\infty} \frac{u^n}{n!} \sum_{j=0}^{k-i} \frac{C(n, k-i-j, s)}{j!} \\
&= \sum_{n=0}^{\infty} \frac{u^n}{n!} \sum_{l=0}^{k} C(n, l, s) \sum_{i=0}^{k-l} \frac{(-1)^i}{i!(k-l)!} \\
&= \sum_{n=0}^{\infty} \frac{u^n}{n!} C(n, k, s).
\end{aligned}$$

(B.3) 式の右辺において、$(1+u)^s - 1 = \sum_{i=1} \binom{s}{i} u^i$ と展開、u^n の係数を比較すれば

$$C(n, k, s) = \frac{n!}{k!} \sum_{r_1 + r_2 + \cdots + r_k = n} \binom{s}{r_1} \binom{s}{r_2} \cdots \binom{s}{r_k},$$

ただし $r_i \geq 1. i = 1, 2, \ldots, k$ を得る。この式を寸法指標で読み替えると

$$C(n. u. \alpha) = \sum_{\boldsymbol{s} \in \mathcal{S}_{n.u}} n! \prod_{i=1}^{n} \frac{1}{s_i!} \binom{\alpha}{i}^{s_i} \tag{B.4}$$

である。また (B.3) より

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} C(n,k,s)t^k\frac{u^n}{n!} = \exp(t\{(1+u)^s - 1\}) \tag{B.5}$$

が言える。

なお

$$c(n,k,s) := (-1)^{n-k}C(n,k,s) \tag{B.6}$$

と書けば、スターリング数の符号を消した議論と同様に

$$(st)^{[n]} = \sum_{k=1}^{n} c(n,k,s)t^{[k]}$$

が成立する。Carlitz[28] は $c(n,k,s)s^{-k}$ の挙動の他、多くの関連する話題を議論している。

Charalambides[34] によれば、任意の $s \neq 0$ について漸化式

$$C(n+1,k,s) = (sk-n)C(n,k,s) + sC(n,k-1,s),$$

ただし $C(0,0,s) = 1, C(0,n,s) = 0\,(n>0), C(0,k,s) = 0\,(k>0)$ が成立する。または

$$c(n+1,k,s) = (n-ks)c(n,k,s) + sc(n,k-1,s),$$

ただし $c(0,0,s) = 1, c(n,0,s) = 0\,(n>0), c(0,k,s) = 0\,(k>0)$、とも書ける。

（一般化）スターリング数の典型的な統計学的応用として、（切り落とし）ベキ級数分布の畳み込みの問題が有る。ベキ級数分布 (3.15) に確率変数 $F_1, F_2, \ldots, F_J$ が従う場合、母数 θ の完備十分統計量は $N = F_1 + F_2 + \cdots + F_J$ である。従って θ の一様最小分散不偏推定量の構築は、N の分布の導出に帰着する（例えば Lehmann and Casella[108] を見よ）。そして確率関数を導出する際、ベキ級数の母関数 $\eta(\theta)^J$ を θ について無限級数展開する事になる。この場合、θ^i の係数として自然に（一般化）スターリング数が現れる。詳しくは Cacoullos and Charalambides[26]、及び Charalambides and Singh[35] の 7 節を参照のこと。

参考文献

[1] Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions*. Dover Publications, New York.

[2] Aitchison , J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**, 901–908.

[3] Aitchison, J. and Brown, J.A. (1957). *The lognormal distribution*. Cambridge University Press, Cambridge.

[4] Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Anlysis*. Cambridge University Press, Cambridge.

[5] Aitchison, J. and Ho, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, **76**, 4, 643–53.

[6] Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**, 358–382.

[7] Atkinson, A.C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413–418.

[8] Atkinson, A.C. and Lam Yeh (1982). Inference for Sichel's compound Poisson distribution. *Journal of the American Statistical Association*, **77**, 153–158.

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

[10] Barnard, G.A., Jenkins, G.M. and Winsten, C.B. (1966). Likelihood inference and time series. *Journal of the Royal Statistical Society*, A, **125**, 321–372.

[11] Barton, D.E. and David, F.N. (1956). Some notes on ordered random intervals. *J. Roy. Stat. Soc., B*, **18**, 79–94.

[12] Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā*, A, *31*, 441–454.

[13] Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, 203–242, Holt, Rinehart and Winston, Toronto.

[14] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer, New York.

[15] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

[16] Birnbaum, A. (1962). On the foundation of statistical inference. *Journal of the American Statistical Association*, **57**, 269–306.

[17] Bjørnstad, J.F. (1990). Predictive likelihood: A review. *Statistical Science*, **5**, 242–265.

[18] Bliss, C.I. and Fisher, R.A. (1953). Fitting the negative binomial distribution to biological data, and note on the efficient fitting of the negative binomial. *Biometrics*, **9**, 176–200.

[19] Bolfarine, H. and Zacks, S. (1992). *Prediction Theory for Finite Populations*, Springer, New York.

[20] Boswell, M.T. and Patil, G.P. (1971). Chance mechanisms generating the logarithmic series distributions used in the analysis of number of species and individuals, in *Statistical Ecology (G.P. Patil, E.C. Pielou, and W.E. Waters, Eds.)*, Vol. 1, Pennsylvania State University Press, University Park, PA.

[21] Brown, G. and Sanders, J.W. (1981). Lognormal genesis. *Journal of Applied Probability*, **18**, 542–547.

[22] Brown, S. and Holgate, P. (1971). Tables of the Poisson lognormal distribution. *Indian Journal of Statistics*, **33**, B, 235–258.

[23] Bulmer, M.G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, **30**, 101–110.

[24] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**, 364–373.

[25] Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion). *Journal of the Royal Statistical Society*, B, **41**, 279–312.

[26] Cacoullos, T. and Charalambides, C.A. (1975). On minimum variance unbiased estimation for truncated binomial and negative binomial distributions. *Annals of Institute of Statistical Mathematics*, **27**, 235–244.

[27] Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

[28] Carlitz, L. (1979). Degenerate Stirling, Bernoulli and Eulerian numbers. *Utilitas Mathematica*, **15**, 51–88.

[29] Cassel, C., Särndal, C. and Wretman, H.H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

[30] Cassie, R.M. (1962). Frequency distribution models in the ecology of plankton and other organisms. *Journal of Animal Ecology*, **31**, 65–92.

[31] Charalambides, C.A. (1974a). The generalized Stirling and C numbers. *Sankhyā*, A, **36**, 419–436.

[32] Charalambides, C.A. (1974b). A sequence of exponential numbers. *Bulletin of Greek Mathematical Society*, **15**, 52–58.

[33] Charalambides, C.A. (1976). The asymptotic normality of certain combinatorial distributions. *Annals of the Institute of Statistical Mathematics*, **28**, A, 499–506.

[34] Charalambides, C.A. (1977). A new kind of numbers appearing in the n-fold convolution of truncated binomial and negative binomial distributions. *SIAM Journal of Applied Mathematics*, **33**, 279–288.

[35] Charalambides, C.A. and Singh, J. (1988). A review of the Stirling numbers, their generalizations and statistical applications. *Communications in Statistics*, Thoer. Meth., **17**, 2533–2595.

[36] Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*, **14**, 79–95.

[37] Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, **17**, 164–177.

[38] Cochran, W.G. (1953). *Sampling Techniques*, 1st ed. Wiley, New York. (3rd ed., 1977)

[39] Cramér, H. (1947). Problems in probability theory. *Annals of Mathematical Statistics*, **18**, 165–193.

[40] Crow, E.L. and Shimizu, K. (1988). *Lognormal Distributions: Theory and Applications*, Marcel Dekker, New York.

[41] Dale, A., Fieldhouse, E. and Holdsworth, C. (2000). *Analyzing Census Microdata*, Arnold, London.

[42] Dennis, B. and Patil, G.P. (1988). Applications in Ecology. *Lognormal Distributions: Theory and Applications*, Crow, E.L. and Shimizu, K. Eds., Marcel Dekker, New York, 303–330.

[43] Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, L.V. (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier, Amsterdam.

[44] Dubey, S.D. (1966). Graphical tests for discrete distributions. *American Statistician*, **20**, 23–25.

[45] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *Regional Conference Series in Applied Mathematics*, No. 38, SIAM, Philadelphia.

[46] Engen, S. (1974). On species frequency models. *Biometrika*, **61**, 263–270.

[47] Engen, S. (1975). A note on the geometric series as a species frequency model. *Biometrika*, **62**, 697–699.

[48] Engen, S. (1977). Comments on two different approaches to the analysis of species frequency data. *Biometrics*, **33**, 205–213.

[49] Engen, S. (1978). *Stochastic Abundance Models.* Chapman and Hall, London.

[50] Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. *Journal of the Royal Statistical Society*, B, **31**, 195–233.

[51] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

[52] Ewens, W.J. (1990). Population genetics theory – the past and the future. in *Mathematical and Statistical Development of Evolutionary Theory*, S. Lessard ed., 177–227, Kluwer. Dordrecht.

[53] Feller, W. (1943). On a general class of "contagious" distributions. *Annals of Mathematical Statistics*, **14**, 389–400.

[54] Feller, W. (1957). *An Introduction to Probability Theory and its Applications, Vol. 1*, 2nd ed., Wiley, New York.

[55] Feller, W. (1966). *An Introduction to Probability Theory and its Applications, Vol. 2*, Wiley, New York.

[56] 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.

[57] Folks, J.L. and Chhikara, R.S. (1978). The inverse Gaussian distribution and its statistical application–A review. *Journal of the Royal Statistical Society*, B, **40**, 263–289.

[58] Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, B, **17**, 268–278.

[59] Godambe, V.P. (1966). A new approach to sampling from finite populations. I. Sufficiency and linear estimation. *J. Roy. Statist. Soc.*, B. **28**, 310–319.

[60] Godambe, V.P. and Patil, G.P. (1975). Some characterisations involving additivity and infinite divisibility and their applications to Poisson mixtures and Poisson sums., *Statistical Distributions in Scientific Work*, **3**, G.P. Patil, S. Kotz and J.K. Ord ed., 339–351. Reidel, Dordrecht.

[61] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

[62] Good, I.J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, Massachusetts.

[63] Goodman, L.A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*, **20**, 572–579.

[64] Greenberg, B.V. and Zayatz, L.V. (1992). Strategies for measuring risk in public use microdata file. *Statistica Neerlandica*, **46**, 33–48.

[65] Greenwood, M. and Yule, G.U. (1920). An inquiry into the nature of frequency distribution representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, **83**, 255–279.

[66] Grundy, P.M. (1951). The expected frequencies in a sample of an animal population in which the abundances of species are log-normally distributed. *Biometrika*, **38**, 427–434.

[67] Gupta, R.C. (1974). Modified power series distributions and some of its applications. *Sankhyā*, B, **35**, 288–298.

[68] Gurland, J. (1957). Some interrelations among compound and generalized distributions. *Biometrika*, **44**, 265–268.

[69] Haas, P.J. and Stokes, L. (1998). Estimating the number of classes in a finite population. *Journal of the American Statistical Association*, **93**, 1475–1487.

[70] Haight, F. (1967). *Handbook of the Poisson Distribution.* Wiley, New York.

[71] Hall, P., Peng, L. and Tajvidi, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika*, **86**, 871–880.

[72] Halmos, P.R. (1944). Random alms. *Annals of Mathematical Statistics*, **15**, 182–189.

[73] Harris, I.R. (1989). Predictive fit for natural exponential families. *Biometrika*, **76**, 675–684.

[74] Heller, G.Z. (1997). Estimation of the number of classes. *South African Statistical Journal*, **31**, 65–90.

[75] Herdan, G. (1958). The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, **45**, 222–228.

[76] Herdan, G. (1961). A critical examination of Simon's model of certain distribution functions in linguistics. *Applied Statistics*, **10**, 65–76.

[77] Holgate, P. (1969). Species frequency distributions. *Biometrika*, **56**, 651–660.

[78] Holgate, P. (1970). The modality of some compound Poisson distribution. *Biometrika*, **57**, 666–667.

[79] Holla, M.S. (1966). On a Poisson-inverse Gaussian distribution. *Metrika*, **11**, 115–121.

[80] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, **17**, 499–520.

[81] Hoshino, N. (2002a). On limiting random partition structure derived from the conditional inverse Gaussian-Poisson distribution. *Technical Report CMU-CALD-02-100*, School of Computer Science, Carnegie Mellon University.

[82] Hoshino, N. (2002b). Engen's extended negative binomial model revisited. *Discussion Paper No. 2002-1*, Faculty of Economics, Kanazawa University.

[83] Hoshino, N. (2003a). Random clustering based on the conditional inverse Gaussian-Poisson distribution. *Journal of the Japan Statistical Society*, **33**, 105–117.

[84] 星野伸明 (2003b). 「超母集団モデルによる個票開示リスク評価」統計数理, To appear.

[85] Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, **28**, 2, 125–134.

[86] Ishii, G. and Hayakawa, R. (1960). On the compound binomial distribution. *Annals of the Institute of Statistical Mathematics*, **12**, 69–80 and Errata, p. 208.

[87] Ismail, M.E.H. (1977). Integral representations and complete monotonicity of various quotients of Bessel functions. *Canadian Journal of Mathematics*, **29**, 1198–1207.

[88] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Multivariate Distributions*. Vol. 1, 2nd ed., Wiley, New York.

[89] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

[90] Johnson, N.L., Kotz, S. and Kemp, A.W. (1993). *Univariate Discrete Distributions*, 2nd ed., Wiley, New York.

[91] Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution.* Lecture Notes in Statistics 9, Springer, New York.

[92] Karlin, S. and McGregor, J. (1958). Linear growth, birth and death processes. *J. Math. Mech.*, **7**, 643–662.

[93] Karlin, S. and McGregor, J. (1967). The number of mutant forms maintained in a population. *Proc. 5th Berkeley Symp. Math. Satatist. Prob.*, **9**, 415–438.

[94] Karlin, S. and McGregor, J. (1972). Addendum to a Paper of W. Ewens. *Theoretical Population Biology*, **3**, 113–116.

[95] Karunamuni, R.J. and Quinn, T.J.II (1995). Bayesian Estimation of Animal Abundance for Line Transect Sampling. *Biometrics*, **51**, 1325–1337.

[96] Katti, S. and J. Gurland. (1961). The Poisson Pascal distribution. *Biometrics*, **17**, 527–538.

[97] Kemp, A.W. (1978). Cluster size probabilities for generalized Poisson distributions. *Communications in Statistics*, Thoer. Meth., **7**, 1433–1438.

[98] Kempton, R.A. and Taylor, L.R. (1974). Log-series and log-normal parameters as diversity discriminants for the Lepidoptera. Journal of Animal Ecology, **43**, 381–399.

[99] Khatri, C.G. and Patel, I.R. (1961). Three classes of univariate discrete distributions. *Biometrics*, **17**, 567–575.

[100] Khmaladze, E.V. (1987). The statistical analysis of a large number of rare events. *Technical Report Report MS-R8804*, Department of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

[101] 木元新作・武田博清 (1989). 『群集生態学入門』 共立出版, 東京.

[102] Kimura, M. and Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

[103] Kingman, J.F. (1978a). Random partitions in population genetics. *Proceedings of the Royal Society of London*, A, **361**, 1–20.

[104] Kingman, J.F. (1978b). The representation of partition structures. *Journal of London Mathematical Society*, **18**, 374–380.

[105] Kingman, J.F. (1982). The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

[106] Konishi, S. and Kitagawa. G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

[107] Kupper, J. (1963). Some aspects of cumulative risk. *ASTIN Bulletin*, **3**, 85–103.

[108] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.

[109] Levin, B. and Reeds, J. (1977). Compound multinomial likelihood functions are unimodal: Proof of a conjecture of I. J. Good. *Annals of Statistics*, **5**, 79–87.

[110] Levy, P.S. and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*, Wiley, New York.

[111] Ludwig, J.A. and Reynolds, J.F. (1988). *Statistical Ecology – A Primer on Methods and Computing*, Wiley, New York.

[112] MacArthur, R.H. (1957). On the relative abundance of bird species. *Proceedings of the national academy of sciences of the USA*, **43**, 293–295.

[113] Maceda, E.C. (1948). On the compound and generalized Poisson distributions. *Annals of Mathematical Statistics*, **19**, 414–416.

[114] Madow, W.G. and Madow, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, **15**, 1–24.

[115] Mandelbrot, B.B. (1983). *The Fractal Geometry of Nature*. W. H. Freeman and Company, New York.

[116] Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991). The case for a sample of anonymised records from the 1991 census. *Journal of the Royal Statistical Society*, A, **154**, 305–340.

[117] 松田芳郎・濱砂敬郎・森博美編 (2000). 『統計調査制度とミクロ統計の開示』日本評論社, 東京.

[118] Mendenhall, T.C. (1887). The characteristic curves of composition. *Science*, **9** (214, supplement), 237–249.

[119] Mosimann, J.E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika*, **49**, 65–82.

[120] Mosteller, F. and Wallace, D.L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, **58**, 275–309.

[121] Mudholkar, G.S. and Natarajan, R. (1999). Approximations for the inverse Gaussian probabilities and percentiles. *Communications in Statistics*, Simula., **28**, 1051–1071.

[122] Myers, R.A. and Pepin, P. (1990). The Robustness of Lognormal-Based Estimators of Abundance. *Biometrics*, **46**, 1185–1192.

[123] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–625.

[124] Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Annals of Mathematical Statistics*, **10**, 35–57.

[125] Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, **21**, 127–132.

[126] Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. *Statistical Data Protection - Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 edition*, 59–76, Office for Official Publications of the European Communities, Luxembourg.

[127] Ong, S.H. (1998). A note on the mixed Poisson formulation of the Poisson-inverse Gaussian distribution. *Communications in Statistics*, Simula., **27**, 67–78.

[128] Ord, J.K. (1967). Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society*, A, **130**, 232–238.

[129] Ord, J.K. and Whitmore, G.A. (1986). The Poisson-inverse Gaussian distribution as a model for species abundance. *Communications in Statistics*, Theor. Meth., **15**, 853–871.

[130] Ottestad, P. (1944). On certain compound frequency distributions. *Skandinavisk Aktuarietidskrift*, **27**, 32–42.

[131] Pareto, V. (1897). *Cours d'Économie Politique*. F. Rouge, Lausanne.

[132] Paul, S.R. and Plackett, R.L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika*, **65**, 591–602.

[133] Pennington, M. (1983). Efficient estimators of abundance, for fish and plankton surveys. *Biometrics*, **39**, 281–286.

[134] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[135] Pitman, J. (1996a). Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, **28**, 525–539.

[136] Pitman, J. (1996b). Notes on the two parameter generalization of Ewens' random partition structure. *Unpublished manuscript*.

[137] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, **25**, 855–900.

[138] Preston, F.W. (1948). The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

[139] Price, T. (1997). Vocabulary growth functions. *South African Statistical Journal*, **31**, 39–63.

[140] Raj D. (1958). On the Relative Accuracy of Some Sampling Techniques. *Journal of the American Statistical Association*, **53**, 98–101.

[141] Rao P.S.R.S. (2000). *Sampling Methodologies with Applications*. Chapman and Hall /CRC, New York.

[142] Reid, D.D. (1981). The Poisson lognormal distribution and its use as a model of plankton aggregation. *Statistical Distributions in Scientific Work*, C. Taillie, G.P. Patil and B. Baldessari Ed., **6**, Proceedings of the NATO Advanced Study Institute, 303–316, D. Reidel Publishing Company, Dordrecht.

[143] Rider, P.R. (1962). Expected values and standard deviations of the reciprocal of a variable from a decapitated negative binomial distribution. *Journal of the American Statistical Association*, **57**, 439–445.

[144] Royall, R.M. and Herson, J. (1973a). Robust estimation in finite populations. I *Journal of the American Statistical Association*, **68**, 880–889.

[145] Royall, R.M. and Herson, J. (1973b). Robust estimation in finite populations. II *Journal of the American Statistical Association*, **68**, 890–893.

[146] Royall, R.M. (1988). Finite populations, sampling from. *Encyclopedia of Statistical Sciences*, Vol. 3, 96–101, Wiley, New York.

[147] 佐井至道 (1998). 「個票データにおける個体数とセル数との関係」応用統計学, **27**, 127–145.

[148] 佐井至道 (2000). 「予測個体数の期待値に基づく個票データのリスク評価」統計数理, **48**, 229-251

[149] 佐井至道・竹村彰通 (2000). 「個票データにおける分類の併合モデル」応用統計学, **29**, 63–82.

[150] Sampford, M.R. (1955). The truncated negative binomial distribution. *Biometrika*, **42**, 58–69.

[151] Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics*, **14**, 373–383.

[152] Sankaran, M. (1968). Mixtures by the inverse Gaussian distribution. *Sankhyā*, **30**B, 455–458.

[153] Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

[154] Satterthwaite, F.E. (1942). Generalized Poisson distribution. *Annals of Mathematical Statistics*, **13**, 410–417.

[155] Schwarz, C.J. and Seber, G.A.F. (1999). Estimating animal abundance: Review III. *Statistical Science*, **14**, 427–456.

[156] Scott, A.J., Brewer, K.R.W. and Ho, E.W.H. (1978). Finite Population Sampling and Robust Estimation, *Journal of the American Statistical Association*, **73**, 359–361.

[157] Seber, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters.* 2nd ed., Charles Griffin and Company, London.

[158] Seshadri, V. (1993). *The Inverse Gaussian Distribution.* Clarendon Press, Oxford.

[159] Seshadri, V. (1999). *The Inverse Gaussian Distribution.* Lecture Notes in Statistics 137. Springer, New York.

[160] Shaban, S.A. (1981). Computation of the Poisson-inverse Gaussian Distribution. *Communications in Statistics*, Theor. Meth., **A10**, 1389–1399.

[161] Shaban, S.A. (1988). Poisson-lognormal distributions. *Lognormal Distributions: Theory and Applications*, Crow, E.L. and Shimizu, K. Eds., Marcel Dekker, New York, 195–210.

[162] Shlosser, A. (1981). On estimation of the size of the dictionary of a long text on the basis of a sample. *Engineering Cybernetics*, **19**, 97–102.

[163] Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distribution. *Annals of the Institute of Statistical Mathematics*, **31**, 373–390.

[164] Sibuya, M. (1991). A cluster-number distribution and its application to the analysis of homonyms. *Japanese Journal of Applied Statistics*, **20**, 139–153 (in Japanese).

[165] Sibuya, M. (1992). Numerical calculation of quantiles of the inverse Gaussian distribution, *Japanese Journal of Applied Statistics*, **22**, 113–127 (in Japanese).

[166] Sibuya, M. (1993). A random clustering process. *Annals of the Institute of Statistical Mathematics*, **45**, 459–465.

[167] 渋谷政昭 (2000). 「調査データ公有化における理論的技術的課題」松田他編『統計調査制度とミクロ統計の開示』, 145–167, 日本評論社, 東京.

[168] 渋谷政昭 (2002a). 「母集団と標本で孤立している個体数」文部科学省科学研究費補助金（課題番号 11480055）研究成果報告書, 25–34.

[169] 渋谷政昭 (2002b). 「「星野伸明 2002-06-22 研究会資料」メモ」2002 年 8 月研究会資料.

[170] Sibuya, M., Yoshimura, M. and Shimizu, R. (1964). Negative multinomial distribution. *Annals of the Institute of Statistical Mathematics*, **16**, 409–426.

[171] Sichel, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. *Proceedings of the Third Symposium on Mathematical Statistics* (N.F. Laubscher, ed.). S.A. C.S.I.R., Pretoria, 51–97.

[172] Sichel, H.S. (1973). The density and size distribution of diamonds. *Bull. Int. Statist. Inst.*, **45**, 420–427.

[173] Sichel, H.S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society*, A, **137**, 25–34.

[174] Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547.

[175] Sichel, H.S. (1982a). Asymptotic efficiencies of three methods of estimation for the inverse Gaussian-Poisson distribution. *Biometrika*, **69**, 467–472.

[176] Sichel, H.S. (1982b). Repeat-Buying and the Generalized Inverse Gaussian-Poisson Distribution. *Applied Statistics*, **31**, 193–204.

[177] Sichel, H.S. (1986a). Parameter estimation for a word frequency distribution based on occupancy theory. *Communications in Statistcs*, Theor. Meth., **15**, 935–949.

[178] Sichel, H.S. (1986b). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, **11**, 45–72.

[179] Sichel, H.S. (1992). Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing and Management*, **28**, 5–17.

[180] Sichel, H.S. (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. *South African Statistical Journal*, **31**, 13–37.

[181] Siegel, A.F. and Sugihara, G. (1983). Moments of particle size distributions under sequential breakage with applications to species abundance. *Journal of Applied Probability*, **20**, 158–164.

[182] Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, **42**, 425–440.

[183] Skellam, J.G. (1952). Studies in statistical ecology. *Biometrika*, **39**, 346–362.

[184] Skinner, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, **46**, 21–32.

[185] Skinner, C.J. and Holmes, D.J. (1993). Modelling Population Uniqueness. *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin, 175–199.

[186] Skinner, C.J. and Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 361–372.

[187] Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1990). Disclosure avoidance for census microdata in Great Britain. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington D.C., 131–143.

[188] Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, **10**, 31–51.

[189] Smith, T.M.F. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society*, A, **139**, 183–195.

[190] Smith, S.J. (1988). Evaluating the efficiency of the δ-distribution mean estimator. *Biometrics*, **44**, 485–493.

[191] Spiegel, M.R. (1963). *Theory and Problems of Advanced Calculus*, Shaum Publishing Co., New York.

[192] Stein, G.Z., Zucchini, W. and Juritz, J.M. (1987). Parameter Estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, **82**, 938–944.

[193] Steutel, F.W. (1979). Infinite divisibility in theory and practice. *Scandinavian Journal of Statistics*, **6**, 57–64.

[194] Steutel, F.W. (1983). Infinite divisibility. *Encyclopedia of Statistical Sciences*, Vol. 4, 114–116, Wiley, New York.

[195] Steutel, F.W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability*, **7**, 893–899.

[196] Sugihara, G. (1980). Minimal community structure: an explanation of species abundance patterns. *The American Naturalist*, **116**, 6, 770-787.

[197] 竹村彰通 (1997).「個票データ開示の理論」文部科学省科学研究費補助金（課題番号 08209102）研究成果報告書, 1–25.

[198] 竹村彰通 (1998).「労働力調査に見られるサイズインデックス」文部科学省科学研究費補助金（課題番号 09206102）研究成果報告書, 95–104.

[199] Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. in *Statistical data protection - Proceedings of the conference, Lisbon, 25 to 27 March 1998 - 1999 edition*, 45–58, Office for Official Publications of the European Communities, Luxembourg.

[200] 竹内啓他編 (1989).『統計学辞典』東洋経済, 東京.

[201] Tam, S.M. (1988). Some Results on Robust Estimation in Finite Population Sampling. *Journal of the American Statistical Association*, **83**, 242–248.

[202] Thompson, M.E. (1988). Superpopulation Models. *Encyclopedia of Statistical Sciences*, Vol. 9, 93–99, Wiley, New York.

[203] Thorin, O. (1977). On the infinite divisibility of the lognormal distribution. *Scand. Actuarial J.*, 121–148.

[204] Tokeshi, M. (1993). Species abundance patterns and community structure. *Advances in ecological research*, **24**, 111–186.

[205] Toscano, L. (1978). Some results for generalized Bernoulli, Euler, Stirilng numbers. *Fibonacci Quarterly*, **16**, 103–112.

[206] Tweedie, M.C.K. (1957). Statistical properties of inverse Gaussian distribution. I. *Annals of Mathematical Statistics*, **28**, 362–377.

[207] Uchaikin, V.V. and Zolotarev, V.M. (1999). *Chance and Stability*, VSP, Utrecht.

[208] Urzúa, C.M. (2000). A simple and efficient test for Zipf's law. *Economics Letters*, **66**, 257–260.

[209] Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, Wiley, New York.

[210] Vetterling, W.T., Teukolsky, S.A. and Press, W.H. (1992). *Numerical Recipes: Example Book (C)*, 2nd ed., Cambridge University Press, Cambridge.

[211] Wani, J.K. and Lo, H.P. (1986). Selecting a power-series distribution for goodness of fit. *The Canadian Journal of Statistics*, **14**, 347–353.

[212] Warde, W.D. and Katti, S.K. (1971). Infinite divisibility of discrete distributions, II. *Annals of Mathematical Statistics*, **42**, 1088–1090.

[213] Watson, G.N. (1944). *A Treatise on the Theory of Bessel Functions*. 2nd ed., University Press, Cambridge.

[214] Watterson, G.A. (1974a). The sampling theory of selectively neutral alleles. *Adv. Appl. Prob.*, **6**, 463–488.

[215] Watterson, G.A. (1974b). Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, **6**, 217–250.

[216] Watterson, G.A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *Journal of Applied Probability*, **13**, 639–651.

[217] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice* . Lecture Notes in Statistics 111, Springer, New York.

[218] Willenborg, L. C. R. J. (1996). OR in Statistical Disclosure Control. *Internal Report, Statistics Netherlands*, 1–25.

[219] Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics 155, Springer, New York.

[220] Williams, C.B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, **31**, 356–361.

[221] Williams, C.B. (1956). Studies in the history of probability and statistics: IV. A note on an early statistical study of literary style. *Biometrika*, **43**, 248–256.

[222] Williams, C.B. (1964). *Patterns in the Balance of Nature.* Academic Press, London.

[223] Willmot, G.E. (1986). Mixed compound Poisson distributions. *ASTIN Bulletin*, **16**, S59–S79.

[224] Willmot, G.E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 113–127.

[225] Willmot, G.E. (1989). A remark on the Poisson-Pascal and some other contagious distributions. *Statistics and Probability Letters*, **7**, 217–220.

[226] 大和元 (2003). 「ピットマン確率分割と関連する話題」統計数理, To appear.

[227] Yamato, H. and Sibuya, M. (2000). Moments of some statistics of Pitman sampling formula. *Bulletin of Informatics and Cybernetics*, **32**, 1–10.

[228] Yamato, H., Sibuya, M. and Nomachi, T. (2001). Ordered sample from two-parameter GEM distribution. *Statistics and Probability Letters*, **55**, 19–27.

[229] Yanagimoto, T. (1998). The inverse binomial distribution as a statistical model. *Communications in Statistics*, Theor. Meth., **18**, 3625–3633.

[230] Young, L.J. and Young, J.H. (1998). *Statistical Ecology – A Population Perspective*, Kluwer, Boston.

[231] Yule, G.U. (1939). On sentence-length as a statistical characteristic of style in prose: with applications to two cases of disputed authorship. *Biometrika*, **30**, 363–390.

[232] Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary.* Cambridge University Press, London.

[233] Zayatz, L.V. (1991). Estimation of the percent of unique population elements in a microdata file using the sample. *Statistical Research Division Report RR-91/08*, U.S. Bureau of the Census.

[234] Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Cambridge, MA.