統計数理(2003)
第 51 巻 第 2 号 297–319

特集「個票開示問題の統計理論」
[総合報告]

超母集団モデルによる個票開示リスク評価

星野 伸明[†]

(受付 2003 年 1 月 31 日；改訂 2003 年 4 月 18 日)

要　　旨

個票データは，各個体が分割表のどのセルに所属するかを示すものである．そして秘匿処理は，分割表の解像度を粗くする操作に他ならない．分割表を細かくすればセル内の個体数(頻度)が減る事を考えると，頻度の頻度(寸法指標)推測は開示リスク評価において，重要な役割を果たす．しかし母集団に何の仮定もおかないと，実用的な推測は不可能である．このため有限母集団解析の定石どおり，超母集団モデルを仮定する．経験的に母集団は Zipf の法則に従うので，右裾が長い分布で混合したポアソン分布を利用する．つまり基本モデルでは，各セルでの頻度を独立同一な混合ポアソン分布とみなす．個体総数(の期待値)を一定のままセル数を無限大とする極限操作を，小数法則と呼ぼう．無限分解可能な混合ポアソン分布の基本モデルに小数法則を適用すると，モデルとして意味のある極限分布が得られる．無限分解可能な混合ポアソン分布は多いので，この法則によって便利なモデルを新規に導出できるかもしれない．なお幅広い母集団の記述の為には，多くのモデルを使い分ける必要が有る．従って新規モデルを提供していく事は重要である．

キーワード： 母集団一意，無限分解可能，複合ポアソン，混合ポアソン，自然数の確率分割．

1. 導入

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」または「個体開示」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．この概念自体は自明かもしれないが，リスク評価の統計学的定式化は必ずしも自明ではない．本稿では超母集団モデルという方法論が，個票開示リスク評価で果たす役割を考察する．本論文の構成は以下のとおりである．1.1 節では計測可能なリスク概念が導入される．1.2 節ではリスク評価における超母集団モデルの必要性を説明する．1.3 節では既存のモデル同士の関係を明らかにする．その中で無限分解可能性と条件付けが，重要な意味を持つ．2 章の各節では，既存のモデルの使用法が解説される．3 章では今後の課題について述べたい．

1.1 個票開示リスクの定式化

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だろう．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，レンジは c_l 個のカテ

[†] 金沢大学 経済学部：〒920–1192 石川県金沢市角間町

ゴリーに分類されるだろう．例えば性別という変数なら，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する事が出来ない．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母集団での個体が非常に少ない(たいてい一意に定まる)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，それらは比較的危険な個体と考えられる．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であろう．

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．そのような秘匿処理(修正)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」という．そうでなければ「局所的」と呼ぶ．以下では実務的に重要な，大域的修正のみ考える．また修正の技術としては，「隠蔽」と「再符号化」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」が別物として調査されていても，再符号化して両者を同一の「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，これをうまく説明する事が出来る．すなわち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．従って考え方としては，秘匿処理を用いて分割表を確定，この表の危険性をセルの大きさで計る，という事になる．特定のセルの大きさではなく危険なセルの総数に興味が有るので，「頻度の頻度」(Good (1953))を数える．すなわち

(定義 1.1)
$$S_i = \sum_{j=1}^{J} I(F_j = i), \quad i = 0, 1, \ldots,$$

ただし $I(\cdot)$ は指示関数であり，

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

なお $F_j, j = 1, 2, \ldots, J,$ は非負整数である．以下では Sibuya (1993)にならい，S_i を「寸法指標」と呼ぶ．これは個票開示リスク評価の分野では，Greenberg and Zayatz (1992)が言う equivalence class の概念に対応する．ここで記法をまとめておこう．空でないセルの総数を

(定義 1.2)
$$U = \sum_{i=1}^{\infty} S_i = J - S_0$$

であらわす．また母集団サイズは

(定義 1.3)
$$N = \sum_{j=1}^{J} F_j = \sum_{i=1}^{\infty} iS_i < \infty$$

とする．標本については，セル内個体数を f_j と書く．標本寸法指標も同様に定義され，s_i で表す．u は標本において空でないセルの総数である．また標本サイズは $n = \sum_{j=1}^{J} f_j$ としよう．

母集団で一意な事が知られている個体のレコードが公開されると，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」と呼ばれ，Marsh et al.（1991）によればリスクの指標として，米国・英国の統計当局が用いている．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ような個体は，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 $\sum_{i=1}^{N} w_i iS_i$, 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.（1990）は

$$\sum_{i=1}^{N} \left(\frac{i}{N}\right)^2 S_i$$

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（1992）では，エントロピーから類推して

$$-\sum_{i=1}^{N} \log\left(\frac{i}{N}\right) \frac{i}{N} S_i$$

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．なお Skinner et al.（1994）によれば，カナダの統計当局が用いているリスクの指標は，標本一意かつ母集団一意な個体数の，標本一意に占める割合である．これは

$$\frac{n}{N \cdot s_1} S_1$$

で推定できる．私見を述べれば，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．

1.2 リスクの推測

全数調査については以上の議論で済む．しかし標本調査では母集団の寸法指標が未知であり，推定の対象となる．ここで利用可能な情報は，標本の寸法指標 $(s_0, s_1, \ldots, s_n)$ である．単純化のために，標本が単純無作為非復元抽出でとられたとしよう．この場合

(1.1)
$$\begin{bmatrix} \mathrm{E}(s_1) \\ \mathrm{E}(s_2) \\ \mathrm{E}(s_3) \\ \vdots \\ \mathrm{E}(s_n) \end{bmatrix} = \frac{1}{\binom{N}{n}} \begin{bmatrix} \binom{1}{1}\binom{N-1}{n-1} & \binom{2}{1}\binom{N-2}{n-1} & \binom{3}{1}\binom{N-3}{n-1} & \cdots & \binom{q}{1}\binom{N-q}{n-1} \\ 0 & \binom{2}{2}\binom{N-2}{n-2} & \binom{3}{2}\binom{N-3}{n-2} & \cdots & \binom{q}{2}\binom{N-q}{n-2} \\ 0 & 0 & \binom{3}{3}\binom{N-3}{n-3} & \cdots & \binom{q}{3}\binom{N-q}{n-3} \\ \vdots & & \ddots & & \vdots \\ 0 & & \cdots & & \binom{q}{q}\binom{N-q}{n-q} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_q \end{bmatrix}$$

という関係が成立し，ここから S_i の推定量を構成できる．Engen (1978)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，

$$\hat{S}_{i\,Engen} = \sum_{k=i}^{n} \left\{ \sum_{j=i}^{k} \frac{\binom{N}{j}\binom{k}{j}\binom{j}{i}}{\binom{n}{j}} (-1)^{j-i} \right\} s_k \tag{1.2}$$

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．

残念ながら推定量(1.2)は，きわめて不安定な事が知られている(例えば渋谷(2002)の数値実験を参照せよ)．理由として Shlosser (1981)は,(1.1)の右辺の行列がほとんど特異に近いと指摘する．また渋谷((2000), p. 159)は推定が不安定である事について，直感的な説明を与えている．このように母集団に何の仮定もおかない推定がうまくいかない時，超母集団という構造を仮定するのが一つの方法である．すなわち母集団の性質を分布で記述する．以下では我々も，この方法論を採用しよう．なお超母集団アプローチについては，Smith (1976)の整理が分かりやすい．

なお筆者は超母集団を，「議論を明確にする為の数学的道具に過ぎない」と解釈する．超母集団を，「実世界の確率的機構・過程の記述」とみなすべきではない．Feller (1966)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紹介されている．データが表面的にあてはまるからと言って，背後の構造は一意に特定されない．だとすればモデルの価値は，実機構・過程と切り離されて存在する．なお仮定される超母集団の母数は，データより推定する．すなわち経験ベイズ法が採用される．Lehmann and Casella ((1998), Chapter 4)によれば，経験ベイズ法はモデルの誤特定について比較的頑健である．従って構造の仮定による一般性の喪失を，ある程度緩和する意味がある．とはいえ真の母集団と仮定される超母集団が遠ければ，寸法指標の推定が良いはずがない．観測されている様々な母集団が，頻度的な裏づけとなるような超母集団が望ましい．

実は母集団の寸法指標は，経験的に一定の傾向，右への歪みを示す．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 $r \cdot x_{(r)} \doteq c$ と Zipf (1949), Mandelbrot (1983)等は主張した．様々な分野 —— 個票開示リスク評価も例外ではない —— 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，$S_i \propto 1/i^2$ と主張している(Urzúa (2000))．同様に Pareto (1897)は，所得分布の研究から $S_i \propto 1/i^a$, ただし a は適当な定数，としている．個票の寸法指標実データは，例えば竹村(1998)が与えている．これを見ると，Zipf 等の主張は不当とは言い難い．つまり，このような「逆 J 字」を記述出来る分布がモデルとして望ましい．

「逆 J 字」型分布は，右裾の長い連続分布とポアソン分布を混合して得られる．すなわち，第 j セルの度数 F_j が平均 λ のポアソン分布に従い，λ がガンマ分布や対数正規分布などに従うと考える．そしてもっとも単純なのは，F_j が独立に同一の分布に従う場合だろう．この時寸法指標の同時分布は

$$\mathrm{P}(S_0 = t_0, S_1 = t_1, \ldots) = J! \prod_{i=0}^{\infty} \frac{\mathrm{P}(F_j = i)^{t_i}}{t_i!} \tag{1.3}$$

という多項分布で書ける．従って F_j の分布を規定すれば寸法指標の挙動は定まる．また既知の母集団サイズ N_0 を，モデルに取り込む必要がある．ここで二通りの考え方が有りうる．一つ目は母集団サイズを固定し，$N = N_0$ と制約する．二つ目は母集団サイズを確率変数とみなし，$\mathrm{E}(N) = N_0$ と制約する．一番目の方法はよりリアルだが，モデルが複雑になり条件付分布

$\mathrm{P}(F_1, F_2, \ldots, F_J \,|\, N = N_0)$ を評価しなければならない．二番目の場合は独立同一分布を扱うので，評価が単純となる．すなわち $\mathrm{P}(F_1, F_2, \ldots, F_J) = \prod_{j=1}^{J} \mathrm{P}(F_j)$，ただし $\mathrm{E}(F_j) = N_0/J, j = 1, 2, \ldots, J$, と制約する．

次に標本分布を考察する．単純化の為に単純無作為非復元抽出，すなわち

$$\mathrm{P}(f_1 = y_1, f_2 = y_2, \ldots, f_J = y_J \,|\, F_1, F_2, \ldots, F_J) = \frac{\binom{F_1}{y_1}\binom{F_2}{y_2}\cdots\binom{F_J}{y_J}}{\binom{N}{n}},$$

およびベルヌーイ抽出，すなわち

$$\mathrm{P}(f_1 = y_1, f_2 = y_2, \ldots, f_J = y_J \,|\, F_1, F_2, \ldots, F_J) = \prod_{j=1}^{J} \binom{F_j}{y_j} \left(\frac{n_0}{N_0}\right)^{y_j} \left(\frac{N_0 - n_0}{N_0}\right)^{F_j - y_j}$$

のみ使用する．なお本稿では，周辺分布 $\mathrm{P}(f_1, f_2, \ldots, f_J)$ を「標本分布」と呼ぶ．モデル $\mathrm{P}(F_1, F_2, \ldots, F_J) = p_N$ が所与の時，抽出方法に応じて標本分布を得たい．もし標本分布が N_0 を既知の標本サイズ n_0 に置き換えるだけで得られる，すなわち $\mathrm{P}(f_1, f_2, \ldots, f_J) = p_n$ と表されるなら議論は簡潔である．言い換えれば，超母集団から直接標本を抽出する場合の分布と，超母集団から抽出された母集団より標本を抽出した場合の分布が同じ状況が望ましい．まず母集団サイズが固定されたモデルの場合，Takemura (1999)がそのようになる条件を示した．すなわち個体のラベルにモデルが依存しない時，p_N から単純無作為非復元抽出された標本分布は p_n となる．また母集団サイズが確率変数の場合，本稿で考察する独立同一混合ポアソン分布モデル p_N は，ベルヌーイ抽出をすれば標本分布が p_n となる．何故なら Feller ((1957), p. 268) によれば，平均 $N_0\lambda$ のポアソン分布に従う確率変数から抽出率 n_0/N_0 でベルヌーイ抽出をすれば，標本は平均 $n_0\lambda$ のポアソン分布に従う．

以上の議論を踏まえると，適当な標本抽出法を仮定すれば，寸法指標の推測は次のように行うことが出来る．まず仮定されたモデルの母数を標本から推測する($\hat{\theta}$)．そして推測された母数の下で，寸法指標の期待値 $\mathrm{E}_{\hat{\theta}}(S_i)$ を評価する．これを寸法指標の推定値 $\hat{S}_i$ と考える．

1.3 混合ポアソン分布と小数法則

使用するモデルが決まれば，モデルに応じてリスクの推定値が得られる．ではいかなるモデルを用いるか．出来るだけ多様な母集団を記述出来る，柔軟なモデルが望ましい．しかし私見では，単独のモデルで全ての母集団をカバーする必要は無い．多くのモデルを用意し，データに依存して AIC 等の基準でモデルを選択すれば良い．そして選択されたモデルによるリスク評価を，採用すべきである．従って多くのモデルを提供する事は，記述出来る母集団の範囲を広げる為に重要である．まず既存の超母集団モデルの挙動をふまえた上で，モデル開発の指針を導きたい．以下では個票開示リスク評価における超母集団モデルの文脈を整理する．そしてモデル同士の関係を議論しよう．なお本節ではモデルの名前に言及するにとどめ，その詳細は次節で記述される．

本分野では Bethlehem et al. (1990)がガンマ＝ポアソンモデル(またはポアソン・ガンマモデル)を超母集団モデルとして利用したのが嚆矢となる．これは独立同一な $F_j, j = 1, \ldots, J$, の分布をガンマ分布で混合したポアソン分布，すなわち負の二項分布とみなす．佐井(1998)は J と $\mathrm{E}(N)$ 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(2000)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 (1991)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，有意に当てはまりを欠くと報告している．Skinner et al. (1994)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 (2001)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 (1998)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(2000)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 (1993)は，Fisher et al. (1943)の対数級数分布および対数正規＝ポアソン分布を，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 (1998)は，対数正規＝ポアソンモデル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 (1998)は Anscombe (1950)の対数級数モデルを再発見し，母数推測を議論した．また Takemura (1999)は，ディリクレ＝多項(多変量負の超幾何分布)モデルを提案した．Omori (1999)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 モデルは Takemura (1999)で言及されているが，同分布の使用は Samuels (1998)にも見られる．Hoshino (2001)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 (2003)では Conditional Inverse Gaussian-Poisson (CIGP)モデルを，佐井・竹村(2000)が用いた 1997 年の労働力調査データにあてはめている．

前節の議論では，モデルとして混合ポアソン分布を用いるのが自然であった．実はこれら既存のモデルは全て，混合ポアソン分布またはその極限と関係付けられる．従って混合ポアソン分布の性質を理解する事で，モデルの意味が一層明確になる．以下では混合ポアソン分布の重要な性質を述べる．なおここで触れられない性質については，Johnson et al. ((1993), Chapter 3)等を見よ．

確率変数 F が平均 λ のポアソン分布に従い，λ の密度関数は $f(\lambda)$ とする．つまり我々は混合ポアソン分布，

$$\mathrm{P}(F=y) = \frac{1}{y!}\int_0^\infty \exp(-\lambda)\lambda^y f(\lambda)d\lambda\,, \quad y=0,1,2,\ldots, \tag{1.4}$$

を考察する．Maceda (1948)によれば，分布(1.4)のモメントが存在するとして，

$$\mathrm{E}(F^{(r)}) = \mathrm{E}(\lambda^r)\,, \tag{1.5}$$

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(1.5)が成立するならば，(1.4)を満たす．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．

次に，$J\to\infty$ という極限操作によるモデル導出を考察する．本分野では J が非常に大きくなるので，このような極限に意味がある．ただし期待値制約が有るので

$$J\to\infty\,, \quad \text{ただし } \mathrm{E}(N)=N_0 \text{ 固定}\,, \tag{1.6}$$

という操作を適用する．仮に $F_j, j=1,\ldots,J,$ が独立で同一な分布に従うとしよう．Engen (1977)によれば，もし

$$\lim_{J\to\infty}\frac{c_i}{J} = \mathrm{P}(F_j=i)\,, \quad i=1,2,\ldots, \tag{1.7}$$

ならば $J \to \infty$ の時，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．つまり多項分布(1.3)の周辺の二項分布は，独立なポアソン分布となる．これはいわゆる「小数法則」に他ならない．$\mathrm{E}(N) = N_0$ という制約が(1.7)を意味するのなら，極限分布として意味のあるモデルが得られる．

では極限操作(1.6)が意味を持つような分布族は存在するのだろうか．少なくとも複合ポアソン分布がそれにあたる．正の整数上の確率変数 X について，$\mathrm{P}(X = i) = q_i, i = 1, 2, \ldots,$ とする．X の確率母関数は

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

で表される．ここで確率母関数

$$G(z) = \exp(a(g(z) - 1)) \,, \quad a > 0 \,, \tag{1.8}$$

で定義される非負整数上の分布を，複合ポアソン分布と言う．さて $F_1, \ldots, F_J$ が互いに独立に同一の分布(1.8)に従うとしよう．この時 N の分布の確率母関数は

$$\begin{aligned} G_N(z) &= G(z)^J \\ &= \exp(Ja(g(z) - 1)) \,, \quad a > 0 \,, \end{aligned} \tag{1.9}$$

である．この事からも分かるように，複合ポアソン分布において母数 a は平均に比例する．従って制約条件 $\mathrm{E}(N) = N_0$ は，適当な定数 A について $Ja = A$ のように書ける．Hoshino ((2002b), Theorem 1)によれば，ここで $J \to \infty\, (a \to 0)$ の時，S_i は独立に平均 $q_i A$ のポアソン分布となる．なお Kemp (1978)によれば，$g(z)$ は(1.8)のゼロ切り落とし分布の $a \to 0$ での極限分布を定義する．この $g(z)$ の二重性は，後述されるように混乱を生んだ．

複合ポアソンの独立同一分布モデルの場合，N の分布は F_j の分布の母数 a を Ja に置き換えれば得られる．従って $N = N_0$ の条件付分布を導出しやすい．$g(z)$ がべき級数分布の場合，Hoshino ((2002b), Theorem 2)は混合ポアソンの条件付分布 $\mathrm{P}(F_1, F_2, \ldots, F_J \,|\, N)$ とその(1.6)による極限 $\mathrm{P}(S_1, S_2, \ldots, S_N \,|\, N)$ を与えている．また $J = 1, 2, \ldots$ について $Ja = A$ と固定される場合，N の分布は極限操作(1.6)を適用しても変化しない．この時(1.9)は，複合ポアソン分布が無限分解可能という事を意味する．逆に 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となる(Feller (1957), Section 12.3 を見よ)．すなわち N の分布が極限操作によって変化しない独立同一分布のモデルは，複合ポアソンに限る．

次に複合ポアソン分布と混合ポアソン分布の関係を考察する．Maceda (1948)は，混合ポアソンかつ複合ポアソンとなる例がいくらでも作れる事を指摘した．また混合ポアソン分布が無限分解可能である必要十分条件は，λ の分布が無限分解可能である事を示した．この時 Lévy の定理より，混合ポアソンは複合ポアソンである．Steutel (1983)によれば，よく使われる分布の多くが無限分解可能である．従って極限操作(1.6)と条件付け $(N = N_0)$ の組み合わせは，モデル構築の方法としてかなりの一般性を持つ．

但し混合ポアソンかつ複合ポアソンとなる分布であっても，解析的に扱いやすいとは限らない．例えば対数正規分布は Thorin (1977)によれば，無限分解可能である．従って対数正規=ポアソン分布は複合ポアソンとなるが，その解析的操作は困難である．本稿では混合ポアソンかつ複合ポアソンであり解析的操作が容易な具体例として，λ がガンマ分布および逆ガウシアン分布に従う場合を紹介する．Anscombe (1950)は極限操作(1.6)を，ガンマ=ポアソンモデルに適用し，対数級数モデルを得た．なお負の二項分布のゼロ切り落とし分布の極限が，対数級数分布になる．これは $g(z)$ の二重性の一例であり，Fisher の記述が曖昧だったので混乱を招いた．Sibuya et al. (1964)によれば，ガンマ=ポアソンモデルを母集団サイズで条件付けると

表 1. 混合ポアソン分布モデル.

	$N = N_0$		$\mathrm{E}(N) = N_0$
S_0 が定義される $(J < \infty)$	$\mathrm{P}(F_1, \ldots, F_J \mid N)$ ディリクレ＝多項 CIGP -	$\leftrightarrow$	$\mathrm{P}(F_1, \ldots, F_J)$ ガンマ＝ポアソン 逆ガウシアン＝ポアソン 対数正規＝ポアソン
	$\downarrow$		$\downarrow$
S_0 が定義されない $(J = \infty)$	$\mathrm{P}(S_1, \ldots, S_N \mid N)$ Ewens LCIGP -	$\leftrightarrow$	$\mathrm{P}(S_1, S_2, \ldots)$ 対数級数 拡張負の二項 -

ディリクレ＝多項モデルを得る．そして Watterson (1976), Takemura (1999)によれば，ディリクレ＝多項モデルに(1.6)を適用すると Ewens モデルが得られる．また対数級数モデルを母集団サイズで条件付けると Ewens モデルが得られる(Watterson (1974), Hoshino and Takemura (1998))．次に逆ガウシアン分布の場合だが，Hoshino (2003)は逆ガウシアン＝ポアソンモデルの条件付き分布として CIGP モデルを提案した．そして Hoshino (2002a)は逆ガウシアン＝ポアソンモデルに小数法則を適用し，Engen (1974)の拡張負の二項分布モデルの特殊ケースを得た．なお逆ガウシアン＝ポアソン分布のゼロ切り落とし分布の極限は,(切り落とし)拡張負の二項分布の特殊な場合である．また Hoshino (2002a)によれば，CIGP モデルに小数法則を適用すると，Limiting CIGP (LCIGP)モデルが得られる．これは拡張負の二項分布モデルの特殊ケースの条件付分布になっている．以上の議論は表 1 のように整理される．

2. 超母集団モデル各論

本章ではこれまでに言及したモデルを紹介する．なお全てのモデルについて，定義式の引数は(定義 1.1)から(定義 1.3)を満たす．また本章では対数尤度関数を，標本寸法指標 $(s_1, s_2, \ldots, s_n)$ の関数 L で表す．

2.1 ディリクレ＝多項モデル

本節の議論は Johnson et al. (1997)の 35.13.1 節，及び Takemura (1999)による．

定義. ディリクレ＝多項モデルでは母集団サイズ N が固定されている.(セル同士が交換可能な場合の)モデルは $\gamma > 0$ について

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

のように表される．この時寸法指標に関しては次式で表される．

(定義 2.1)
$$\mathrm{P}(S_0 = t_0, \ldots, S_N = t_N) = \frac{N!J!\Gamma(J\gamma)}{\Gamma(J\gamma + N)} \prod_{i=0}^{N} \left(\frac{\Gamma(\gamma + i)}{\Gamma(\gamma) i!} \right)^{t_i} \frac{1}{t_i!}.$$

リスクの推定

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

(2.1)
$$\mathrm{E}\left(\prod_{j=1}^{J} F_j^{(r_j)} \right) = \frac{N^{(R)} \prod_{j=1}^{J} \gamma^{[r_j]}}{(J\gamma)^{[R]}},$$

但し $R=\sum_{j=1}^{J} r_j$, $x^{(r)}=x(x-1)\cdots(x-r+1)$, $x^{[r]}=x(x+1)\cdots(x+r-1)$ である．寸法指標の期待値については

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

母数 γ の推定については，Mosimann (1962), Takemura (1999)等で議論されている．単純無作為非復元抽出で，大きさ n の標本が取られたとする．標本分布は(定義 2.1)において，N を n で置き換えて得られる．

最尤法から考察しよう．Levin and Reeds (1977)は，尤度関数が γ に関して単峰と証明した．最尤推定量は $dL/d\gamma=0$ の解である．

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

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

より，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．

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

の時，Takemura (1999)は

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

を推定量としている．また標本分散を

$$s^2=\frac{\sum_{i=0}s_i(i-n/J)^2}{J-1} \tag{2.2}$$

と書く．この時近似的に Bethlehem et al. (1990)の推定量

$$\hat{\gamma}_{Bethlehem}=\frac{n}{J(Js^2/n-1)} \tag{2.3}$$

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．

2.2 ガンマ＝ポアソンモデル

Greenwood and Yule (1920)によれば,(1.4)で λ がガンマ分布に従う場合,

$$\mathrm{P}(F=y)=\frac{\Gamma(y+\gamma)}{\Gamma(\gamma)y!}p^{\gamma}\cdot q^{y}\,,\quad y=0,1,2,\ldots, \tag{2.4}$$

但し $0<\gamma, 0<p<1, q=1-p$ 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. (1993)の Chapter 5 を見よ．

定義. ガンマ＝ポアソンモデルでは，母集団サイズが確率変数である．すなわち正の γ,β について，

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

ただし $p=1/(N_0\beta+1), q=1-p$ のように表される．なお寸法指標で表せば

$$\mathrm{P}(S_0=t_0,S_1=t_1,\ldots)=J!\prod_{i=0}^{\infty}\left(\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)i!}p^{\gamma}\cdot q^{i}\right)^{t_i}\frac{1}{t_i!} \tag{定義 2.2}$$

となる．なお期待値制約 $\mathrm{E}(N) = N_0$ を満たすため，

$$\gamma\beta = 1/J\,. \tag{2.5}$$

リスクの推定

周辺の負の二項分布(2.4)については，階乗モメント

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

が評価できる．寸法指標の期待値は

$$\mathrm{E}(S_i) = \sum_{j=1}^{J} \mathrm{P}(F_j = i) = J\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)i!}\left(\frac{1}{N_0\beta+1}\right)^{\gamma}\cdot\left(\frac{N_0\beta}{N_0\beta+1}\right)^{i}. \tag{2.6}$$

負の二項分布に関する母数の推定については Johnson et al.（1993)の 5.8 節，Engen（1978）の 3.4 節等が詳しい．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前提とする．この時(定義 2.2)について N を n で置き換えれば，標本分布が得られる．

ここで最尤法を考えよう．制約式(2.5)の下では

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

を数値的に解けば良い．なお

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

である．なおモメント推定量(2.3)は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．

2.3 対数正規＝ポアソンモデル

Preston（1948)のアイデアに基づき,(1.4)において λ が対数正規分布に従うとする．この時

$$\mathrm{P}(F = y) = \frac{1}{y!\sqrt{2\pi V}}\int_0^{\infty}\lambda^{y-1}\exp(-\lambda-(\log\lambda-M)^2/2V)d\lambda,\quad y = 0,1,2,\ldots, \tag{2.7}$$

のようになる．これを対数正規＝ポアソン分布と呼ぼう．なお本節の議論は，Shaban（1988）に大部分含まれている．

定義. 対数正規＝ポアソンモデルにおいて，母集団サイズは確率変数である．すなわち

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

ただし $V > 0$, $\infty > M > -\infty$ である．寸法指標について表せば

(定義 2.3) $\qquad \mathrm{P}(S_0 = t_0, S_1 = t_1,\ldots)$

$$= J! \prod_{i=0}^{\infty} \left\{ \frac{1}{i!\sqrt{2\pi V}} \int_0^{\infty} \lambda^{i-1} \exp\left(-\lambda - \frac{(\log\lambda - M)^2}{2V} \right) \right\}^{t_i} \frac{1}{t_i!}$$

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

$$M = \log N_0 - \log J - V/2\,.$$

リスクの推定

階乗モメントは Bulmer (1974)が示している．すなわち

$$\mathrm{E}(F^{(r)}) = \exp\left(rM + \frac{1}{2} r^2 V \right) \tag{2.8}$$

である．寸法指標の期待値は(2.7)を利用して

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

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．母集団分布が(定義 2.3)で表される時，標本分布は N を n で置き換えて得られる．なおこの場合の期待値制約を

$$m = \log n_0 - \log J - V/2 \tag{2.9}$$

と書く．

最尤法から考察しよう．

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

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

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 (1974)によれば期待値制約(2.9)が無いものとして，

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

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

となる．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(2.9)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．

モメント推定量としては，

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

を得る事が出来る．ただし s^2 は標本分散(2.2)である．

2.4 逆ガウシアン＝ポアソンモデル

逆ガウシアン(Inverse Gaussian)分布については，Seshadri (1999)のモノグラフが応用も含めて網羅的である．他に Seshadri (1993)，Johnson et al. (1994)も参考になる．

(1.4)において λ が IG 分布に従う時，Holla (1966)によれば

$$\mathrm{P}(F=y) = \sqrt{\frac{2\alpha}{\pi}} \exp(\alpha\sqrt{1-\theta}) \frac{(\alpha\theta/2)^y}{y!} K_{y-1/2}(\alpha)\,, \quad y = 0, 1, 2, \ldots, \tag{2.10}$$

ただし $0<\theta\le 1, \alpha>0$ となる．ここで $K_{y-1/2}(\cdot)$ は $y-1/2$ 次第三種変形ベッセル関数であり，

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

また $K_{-1/2}(\alpha)=K_{1/2}(\alpha)$ となる．本稿では(2.10)を IG ＝ポアソン分布と呼ぶ．本分布は Seshadri ((1999), Section 7.1)において，研究史も含めて詳しく解説されている．また Johnson et al. ((1993), Section 11.15)も参考になる．

定義． 逆ガウシアン＝ポアソンモデルでは母集団サイズが確率変数であり，以下のように書ける．

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

ただし $0<\theta\le 1, \alpha>0$ である．寸法指標については

(定義 2.4)
$$\begin{aligned}&\mathrm{P}(S_0=t_0,S_1=t_1,\ldots)\\&\quad=J!\prod_{i=0}^{\infty}\left\{\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^{i}}{i!}K_{i-1/2}(\alpha)\right\}^{t_i}\frac{1}{t_i!}\end{aligned}$$

となる．期待値制約 $\mathrm{E}(N)=N_0$ は，次式と同値である．

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

リスクの推定

IG ＝ポアソン分布の r 次階乗モメントは，IG 分布の r 次モメントである．すなわち

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

寸法指標の期待値は次のように書ける．

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

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．この時母集団分布が(定義 2.4)ならば，標本分布は N を n で置き換えて得られる．

最尤法から考察する．期待値制約が無いものとして，

$$\frac{\partial L}{\partial\alpha}=\frac{2n+J}{2\alpha}+J\sqrt{1-\theta}+\sum_{i=0}^{\infty}s_i\frac{K_{i-1/2}(\alpha)'}{K_{i-1/2}(\alpha)},$$

$$\frac{\partial L}{\partial\theta}=-\frac{J\alpha}{2\sqrt{1-\theta}}+\frac{n}{\theta}.$$

実は良く知られているように，

$$\frac{K_\gamma(\alpha)'}{K_\gamma(\alpha)}=-R_\gamma(\alpha)+\frac{\gamma}{\alpha},$$

ただし $R_\gamma(\alpha)=K_{\gamma+1}(\alpha)/K_\gamma(\alpha)$ と書ける(例えば Seshadri (1999), p. 125 を見よ)．これを利用して，尤度方程式を数値的に解く事が出来る．期待値制約

$$\alpha=\frac{2n_0\sqrt{1-\theta}}{J\theta}$$

が有る場合はラグランジュ乗数法を用いれば良い．もしくは制約式を代入して α を消去すれば，微分を必要としない最大化アルゴリズムも使える．

期待制約無しの場合，モメント推定は以下の通りになる．

$$\hat{\theta} = 1 - \frac{1}{2\mathrm{V}(x)/\bar{x} - 1}, \quad \hat{\alpha} = \frac{2\bar{x}\sqrt{1-\hat{\theta}}}{\hat{\theta}}.$$

Anscombe (1950)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 (1973)はこの議論を受けて，以下の推定量を提案した．

$$\hat{\theta} = 1 - \left(\frac{-\log \hat{\phi_0}}{2\bar{x} + \log \hat{\phi_0}} \right)^2, \quad \hat{\alpha} = -\frac{1}{2} \log \hat{\phi_0} \left(1 + \frac{\bar{x}}{\bar{x} + \log \hat{\phi_0}} \right), \tag{2.11}$$

ただし $\bar{x}$ は標本頻度の平均，$\hat{\phi_0}$ は観測された頻度 0 の標本割合である．

2.5 対数級数モデル

負の二項分布(2.4)から度数 0 を切り落とした

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

について，$\gamma \to 0$ の極限を適用すると

$$\mathrm{P}(F = y) = \frac{c\, q^y}{y}, \quad y = 1, 2, \ldots,$$

ただし $c = -\{\log(1-q)\}^{-1}$，を得る．これを「対数級数分布」と言う．我々はこれとは区別して，Anscombe (1950)の「対数級数モデル」を考察する．

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

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots) = \prod_{i=1}^{\infty} \frac{\lambda_i^{t_i} \exp(-\lambda_i)}{t_i!} \tag{定義 2.5}$$

のように定義される．この時 $\mathrm{E}(N) = N_0$ となっている．

リスクの推定

対数級数モデルでは，S_i は平均 λ_i のポアソン分布に従う．従って

$$\mathrm{E}(S_i^{(r)}) = {\lambda_i}^r$$

と書ける．Johnson et al. (1993)の 4 章を見よ．なお各 i 独立であり，特に $\mathrm{E}(S_i) = \lambda_i$.

母数の推定は Hoshino and Takemura (1998)で議論された．抽出率 n_0/N_0 のベルヌーイ抽出を考えよう．1.2 節で確認したように，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．なお応用上は $n = n_0$ とみなす事に注意．

対数尤度 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．

モメント推定量は，例えば

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

等が考えられる．

2.6 Ewens モデル

Ewens (1972)が導入した分布を Ewens 分布，または Ewens Sampling Formula と呼ぶ．本分布については Johnson et al. (1997)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

(定義 2.6)
$$\mathrm{P}(S_1=t_1,S_2=t_2,\ldots,S_N=t_N)=\frac{\theta^U}{\theta^{[N]}}\frac{N!}{\prod_{i=1}^N i^{t_i}t_i!},$$

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

リスクの推定

寸法指標の同時階乗モメントは Sibuya (1993)が与えている．すなわち

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

ただし $r=r_1+\cdots+r_N, s=\sum_{i=1}^N ir_i\le N$．特に

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

Ewens 分布は個体のラベルに依存しない．故に大きさ N の母集団から n 個の標本を単純無作為非復元抽出する場合，標本分布は(定義 2.6)について N,U を n,u で置き換えて得られる．

母数 θ の最尤推定量は渋谷(1991)が示したように，

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

の解である．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．なお渋谷(1991)が指摘するように，Ewens 分布は指数分布族に所属する．従って最尤推定は一意な解を持つ．

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．

2.7 Pitman モデル

Pitman (1995)は Ewens 分布を二母数へ拡張し, Pitman 分布を得た．これは Pitman Sampling Formula とも呼ばれる．Ewens 分布と同様に寸法指標の同時確率を与える為，モデルとして使用できる．

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,

$$\text{(定義 2.7)} \qquad \mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_N = t_N) = N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=1}^{N} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{t_j} \frac{1}{t_j!},$$

ただし $\theta^{[U:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(U-1)\alpha)$, $\theta^{[N]} = \theta(\theta+1)\cdots(\theta+N-1)$.

もし α が 0 ならば,(定義 2.7)は Ewens モデル(定義 2.6)と一致する．また $\alpha < 0$ の場合 $\theta = -J\alpha > 0$, $\gamma = -\alpha > 0$ とおけば,(定義 2.7)はディリクレ＝多項モデル(定義 2.1)となる．

リスクの推定

Yamato and Sibuya (2000)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

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

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

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

Pitman 分布は個体のラベルに依存しない．故に大きさ N の母集団から n 個の標本を単純無作為非復元抽出した場合，標本分布は母集団分布(定義 2.7)について N, U を n, u で置き換えて得られる．

最尤推定量は，以下の同時方程式の解である．

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

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

これは数値的に評価する必要が有る．例えば Hoshino (2001)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．

次に近似的モメント推定量を紹介する．

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

ただし $c = s_1(s_1-1)/s_2$．導出は Hoshino (2001)の Appendix を見よ．

2.8 条件付逆ガウシアン＝ポアソン(CIGP)モデル

Hoshino (2003)は逆ガウシアン＝ポアソンモデル(定義 2.4)について，母集団サイズ N を固定した条件付分布を CIGP 分布と呼び，その性質を議論した．これをモデルとして用いる．

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．

(定義 2.8)
$$\mathrm{P}(S_0 = t_0, \ldots, S_N = t_N) = \left(\frac{2\alpha}{\pi}\right)^{\frac{J-1}{2}} \frac{J!N!}{J^{N+1/2}K_{N-1/2}(J\alpha)} \prod_{i=0}^{N} \left\{\frac{K_{i-1/2}(\alpha)}{i!}\right\}^{t_i} \frac{1}{t_i!}.$$

リスクの推定

寸法指標の同時階乗モメントは Hoshino (2003)によれば,

$$\mathrm{E}\left(\prod_{j=1}^{N} S_j^{(r_j)}\right) = \left(\frac{2\alpha}{\pi}\right)^{\frac{r}{2}} \frac{N!J!K_{N-R-1/2}((J-r)\alpha)(J-r)^{N-R+1/2}}{(N-R)!(J-r)!J^{N+1/2}K_{N-1/2}(J\alpha)} \prod_{j=1}^{N} \left(\frac{K_{j-1/2}(\alpha)}{j!}\right)^{r_j},$$

ただし $r = \sum_{j=1}^{N} r_j$, $R = \sum_{j=1}^{N} jr_j$, $S_j^{(r_j)} = S_j(S_j-1)\cdots(S_j-r_j+1)$，となる．特に

$$\mathrm{E}(S_i) = \sqrt{\frac{2\alpha}{\pi}} \frac{K_{i-1/2}(\alpha)}{i!} \frac{N!K_{N-i-1/2}((J-1)\alpha)(J-1)^{N-i+1/2}}{(N-i)!J^{N-1/2}K_{N-1/2}(J\alpha)}.$$

CIGP 分布は個体のラベルに依存しない．故に(定義 2.8)に従う大きさ N の母集団から n 個の標本を単純無作為非復元抽出した場合，標本分布は(定義 2.8)の N を n に置き換えたものとなる．

最尤推定量は尤度方程式 $dL/d\alpha = 0$ の解である．ここで

$$\frac{dL}{d\alpha} = JR_{n-1/2}(J\alpha) - \sum_{i=0}^{n} s_i R_{i-1/2}(\alpha),$$

ただし $R_\gamma(\alpha) = K_{\gamma+1}(\alpha)/K_\gamma(\alpha)$ である．また二次の微分係数は

$$\frac{d^2L}{d\alpha^2} = J^2\left\{R_{n-1/2}^2(J\alpha) + \frac{2n}{J\alpha}R_{n-1/2}(J\alpha)\right\} - \sum_{i=0} s_i\left\{R_{i-1/2}^2(\alpha) + \frac{2i}{\alpha}R_{i-1/2}(\alpha)\right\} - J^2 + J$$

で与えられる．

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 (2003)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．標本分散(2.2)を利用すれば,

$$\tilde{\alpha} = \frac{n\sqrt{n(2Js^2 - n)}}{J(Js^2 - n)}$$

が一つの近似的推定量である．また s_0 を用いた IGP 分布の推定量(2.11)に対応して

$$\bar{\alpha} = -\frac{1}{2}(\log s_0 - \log J)\left(1 + \frac{n/J}{n/J + \log s_0 - \log J}\right)$$

が提案された．Hoshino (2003)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．

2.9 拡張負の二項分布モデル

本節の議論は Hoshino (2002b)に詳しい．ガンマ＝ポアソンモデルでの寸法指標 S_i の期待値(2.6)は，期待値制約(2.5)を代入すると

$$\mathrm{E}(S_i) = \frac{1}{\beta}\frac{\Gamma(i+\gamma)}{\Gamma(1+\gamma)i!}\left(\frac{1}{N_0\beta+1}\right)^{\gamma}\left(\frac{N_0\beta}{N_0\beta+1}\right)^{i} \tag{2.12}$$

と書ける．対数級数モデルでは $\gamma \to 0$ の挙動を考察したが，Engen (1974)は(2.12)式の $\Gamma(1+\gamma)$ に着目し，$i \geq 1$ の時 $-1 < \gamma < 0$ と出来ると指摘した．このようなモデルを，「拡張負の二項分布(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，Engen のアイデアは「拡張(切り落とし)負の二項分布」として利用されてきた．確率関数で表すと

$$\mathrm{P}(F=i) = \frac{-\gamma}{1-(1-\theta)^{-\gamma}}\frac{\Gamma(i+\gamma)}{\Gamma(1+\gamma)i!}\theta^i \propto \mathrm{E}(S_i)\,, \quad i=1,2,\ldots,$$

ただし $-1 < \gamma < 0$, $0 < \theta \leq 1$ である．これは ENB モデルとは区別しなければならない．より詳しくは Engen (1978)，または Johnson et al. (1993)の 5.12.2 節を見よ．

定義． $-1 < \gamma\,(\gamma \neq 0)$, $0 < \theta < 1$ について，拡張負の二項分布モデルは次のように定義される．

$$\mathrm{P}(S_1=t_1, S_2=t_2, \ldots) = \prod_{i=1}^{\infty}\frac{\exp(-\tau(i;\gamma,\theta))\tau(i;\gamma,\theta)^{t_i}}{t_i!}\,, \tag{定義 2.9}$$

ただし

$$\tau(i;\gamma,\theta) = \frac{N_0}{\theta}\frac{(1-\theta)^{\gamma+1}\theta^i\Gamma(i+\gamma)}{\Gamma(1+\gamma)i!}$$

である．このモデルを ENB(γ) と表す．期待値制約 $\mathrm{E}(N) = N_0$ を満たす．

$\theta \neq 1$ の時，逆ガウシアン＝ポアソンモデルの極限として得られるのが ENB($-1/2$) である．また逆ガウシアン＝ポアソン分布のゼロ切り落とし分布で $\alpha \to 0$ とすると，拡張(切り落とし)負の二項分布 ($\gamma = -1/2$) を得る．なお Hoshino (2002b)によれば，一般化された逆ガウシアン＝ポアソンモデルの極限として一般の ENB(γ) が得られる．

リスクの推定

S_i がポアソン分布に従う事より，

$$\mathrm{E}(S_i^{(r)}) = \tau(i;\gamma,\theta)^r$$

のように書ける．なお各 i 独立であり，特に $\mathrm{E}(S_i) = \tau(i;\gamma,\theta)$.

母数推定の議論について，詳しくは Hoshino (2002b)を見よ．拡張負の二項分布モデルでは，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(定義 2.9)について N を n で置き換えたものになる．

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．

$$\frac{\partial L}{\partial\theta} = -\frac{n_0}{\gamma}\left(\frac{(1-\theta)^{\gamma+1}-1}{\theta^2} + \frac{(1+\gamma)(1-\theta)^{\gamma}}{\theta}\right) + (n-u)\frac{1}{\theta} - u\frac{(\gamma+1)}{1-\theta}\,,$$

$$\frac{\partial L}{\partial\gamma} = n_0\frac{1-\theta}{\theta\gamma^2} + \frac{n_0}{\theta}\left(-\frac{(1-\theta)^{\gamma+1}}{\gamma^2} + \frac{(1-\theta)^{\gamma+1}}{\gamma}\log(1-\theta)\right)$$

$$+ u\log(1-\theta) + S_1 + \sum_{i=2}^{\infty} s_i \sum_{j=1}^{i-1} \frac{1}{\gamma+j}\,.$$

また二次の微分係数は以下の式で与えられる．

$$\begin{aligned}
\frac{\partial^2 L}{\partial\gamma\partial\theta} &= \frac{n_0((1-\theta)^{\gamma+1}-1)}{\theta^2\gamma^2} - \frac{n_0}{\theta^2\gamma}(1-\theta)^{\gamma+1}\log(1-\theta) \\
&\quad + \frac{n_0}{\theta\gamma^2}(1-\theta)^{\gamma} - \frac{n_0}{\theta\gamma}(1-\theta)^{\gamma}\log(1-\theta) - \frac{n_0}{\theta}(1-\theta)^{\gamma}\log(1-\theta) - \frac{u}{1-\theta}\,. \\
\frac{\partial^2 L}{\partial\theta^2} &= -\frac{n_0}{\gamma}\left(\frac{2-2(1-\theta)^{\gamma+1}}{\theta^3} - \frac{2(1+\gamma)(1-\theta)^{\gamma}}{\theta^2} - \frac{(1+\gamma)\gamma(1-\theta)^{\gamma-1}}{\theta}\right) \\
&\quad - \frac{n-u}{\theta^2} - \frac{u(\gamma+1)}{(1-\theta)^2}\,. \\
\frac{\partial^2 L}{\partial\gamma^2} &= \frac{n_0}{\theta}\left(\frac{-2(1-\theta)+2(1-\theta)^{\gamma+1}}{\gamma^3} - \log(1-\theta)\frac{2(1-\theta)^{\gamma+1}}{\gamma^2} + \log^2(1-\theta)\frac{(1-\theta)^{\gamma+1}}{\gamma}\right) \\
&\quad - \sum_{i=2}^{\infty} s_i \sum_{j=1}^{i-1} \frac{1}{(j+\gamma)^2}\,.
\end{aligned}$$

最尤推定は，同時方程式 $\partial L/\partial\theta = \partial L/\partial\gamma = 0$ を解いて得られる．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．

$$\frac{\partial L}{\partial\theta} = 0 \Leftrightarrow u = \frac{n_0(1-(1-\theta)^{\gamma})(1-\theta)}{\gamma\theta}\,,$$

$$\frac{\partial L}{\partial\gamma} = 0 \Leftrightarrow \frac{n_0(1-\theta)}{\gamma\theta}\left\{\frac{1-(1-\theta)^{\gamma}}{\gamma} + \log(1-\theta)\right\} + \sum_{i=2}^{\infty} s_i \sum_{j=1}^{i-1} \frac{1}{\gamma+j} = 0\,.$$

γ が所与でない限り，モメント推定であっても数値的に方程式を解く事になる．簡便な近似推定の余地は有ると思われるが，議論はなされていない．なお γ が -1 に近い程，小さなセルが支配的となる．

2.10 極限 CIGP モデル

Hoshino (2002a)は CIGP 分布(定義 2.8)に極限操作(1.6)を適用した．ここで得られた分布を Limiting CIGP (LCIGP)分布と呼び，モデルとして用いる事が出来る．

定義. 母数 $A > 0$ について，LCIGP モデルは次のように表される．

(定義 2.10)
$$\begin{aligned}
&\mathrm{P}(S_1 = t_1, \ldots, S_N = t_N) \\
&\qquad = \sqrt{\frac{\pi}{2}}\frac{N!\exp(-A)}{A^{N-U+1/2}K_{N-1/2}(A)}\prod_{i=1}^{N}\left(\frac{(2i-3)!!}{i!}\right)^{t_i}\frac{1}{t_i!}\,.
\end{aligned}$$

なお本モデルでは母集団サイズが固定されている．

リスクの推定

Hoshino (2002a) によれば，寸法指標の同時階乗モメントは次式で与えられる．

$$\mathrm{E}\left(\prod_{i=1}^{N} S_i^{(r_i)}\right) = \frac{K_{N-R-1/2}(A)A^{r-R}N!}{K_{N-1/2}(A)(N-R)!}\prod_{i=1}^{N}\left(\frac{(2i-3)!!}{i!}\right)^{r_i}\,,$$

ただし $r = \sum_{i=1}^{N} r_i, R = \sum_{i=1}^{N} i r_i, n^{(R)} = n(n-1)\cdots(n-R+1)$ である．

特に $i = 1, 2, \ldots, N$ について

$$\mathrm{E}(S_i) = \frac{K_{N-i-1/2}(A)A^{1-i}(2i-3)!!N!}{K_{N-1/2}(A)i!(N-i)!}\,.$$

単純無作為非復元抽出の場合，大きさ n の標本分布は母集団分布（定義 2.10）において N, U を n, u で置き換えて得られる．

対数尤度について一次の微分係数は次のとおり．

$$\frac{\partial L}{\partial A} = -1 - (2n-u)\frac{1}{A} + R_{n-1/2}(A)\,,$$

ただし $R_\gamma(\omega) = K_{\gamma+1}(\omega)/K_\gamma(\omega)$ である．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．LCIGP 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．

LCIGP 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（2002a）は，以下の近似的モメント推定量を提案した．

$$\tilde{A} = \begin{cases} u/(1-\sqrt{1-4s_2/s_1}\,), & \text{もし } 4s_2/s_1 \le 1\,, \\ u\,, & \text{その他の場合}\,. \end{cases}$$

3. 展望

経験的に $s_1 - s_2$ は，$s_i - s_{i+1}, i = 2, 3, \ldots,$ に比べて変化率が大きい．つまり寸法指標は滑らかに変化しないので，s_1 の「跳び」を記述出来る分布が望ましい．そして「跳び」という異質性を記述するには，追加の母数が必要と思われる．従って，二母数のモデルが有利だろう．2 章で紹介されたモデルのうち，二母数モデルは Pitman モデルと ENB モデルのみである．故に Pitman モデルを ENB モデルが改善するかどうかは興味深い．ところが Hoshino（2002b）によれば，二つのモデルは N, U 所与の条件付分布が同じになる．この場合，あてはまりは同程度だろう．結局，これらの分布とは別の強みを持つモデルが望まれる．

モデルの開発に試行錯誤は付き物だが，1.3 節で述べた Hoshino（2002b）の一般論を利用すれば，錯誤は減るだろう．基本的には正の整数上の分布 $\mathrm{P}(i), i = 1, 2, \ldots,$ が，s_i/u を記述出来れば良い．この時，$\mathrm{E}(S_i) \propto \mathrm{P}(i), i = 1, 2, \ldots,$ となる独立ポアソン分布のモデルが，小数法則の極限として正当化される．そのようなモデルで N の分布は複合ポアソンなので，$N = N_0$ という条件付けの際に組み合わせ論的困難は存在しない．

このような正の整数上の分布によるモデル構築を考えると，我々は Zipf 以来の古典的議論に回帰した事に気付く．すなわち，計量言語学や統計的生態学においてゼロ切り落とし分布をあてはめる議論である．これら単純なあてはめは発展性が無いため，その議論は忘れられかけている．しかし本稿で示した観点からは，この文脈で用いられた正の整数上の分布が，新たな意味を持つ．例えばべき級数分布族の中でのあてはまりの比較は，Wani and Lo（1986）によって議論された．これは望ましい新規モデルについての指針を与える．

形式的に本分野の超母集団モデルは離散多変数分布である．Johnson et al.（1997）の序文で “less has been done” と述べられているように，実用的な離散多変数分布はそれ程多く知られていない．従って超母集団モデルの開発は，確率論的にも価値がある．例えば $\mathrm{P}(F_1, F_2, \ldots, F_J \mid N)$ の下での U の挙動は，壺モデルでの占有問題に対応する．なお U は計量言語学では語彙数，統

計的生態学では種の総数を表す重要な統計量である(Bunge and Fitzpatric (1993)の総合報告を見よ)．また $\mathrm{P}(S_1,\ldots,S_N \mid N)$ は，自然数の確率分割と同等である．つまり小数法則と条件付けを組み合わせる事で，体系的にこれらの離散多変数分布が得られる．

本分野における超母集団モデルの議論が，著者にとっては思いもよらなかった一般性を見せている事を，ここまでで報告したつもりである．本稿がこの分野に興味を持ってもらうきっかけになる事を期待しつつ，とりあえず筆をおくことにする．

謝　辞

本研究は統計数理研究所の共同利用研究プロジェクト「個票データの開示におけるリスクの評価と官庁統計データの公開への応用(代表：佐井至道)」の成果に基づくものであり，科学研究費の援助も受けている．

参 考 文 献

Anscombe, F. J. (1950). Sampling theory of the negative binomial and logarithmic series distributions, *Biometrika*, **37**, 358–382.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata, *J. Amer. Statist. Assoc.*, **85**, 38–45.

Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data, *Biometrics*, **30**, 101–110.

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

Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata, *Journal of Official Statistics*, **14**, 79–95.

Engen, S. (1974). On species frequency models, *Biometrika*, **61**, 263–270.

Engen, S. (1977). Comments on two different approaches to the analysis of species frequency data, *Biometrics*, **33**, 205–213.

Engen, S. (1978). *Stochastic Abundance Models*, Chapman and Hall, London.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theoretical Population Biology*, **3**, 87–112.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd ed., Wiley, New York.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York.

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

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters, *Biometrika*, **40**, 237–264.

Greenberg, B. V. and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata file, *Statist. Neerlandica*, **46**, 33–48.

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 of or repeated accidents, *J. Roy. Statist. Soc.*, **83**, 255–279.

Holla, M. S. (1966). On a Poisson-inverse Gaussian distribution, *Metrika*, **11**, 115–121.

Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment, *Journal of Official Statistics*, **17**, 499–520.

Hoshino, N. (2002a). On limiting random partition structure derived from the conditional inverse Gaussian-Poisson distribution, Tech. Report, CMU-CALD-02-100, School of Computer Science, Carnegie Mellon University, Pittsburgh.

Hoshino, N. (2002b). Engen's extended negative binomial model revisited, Discussion Paper, No. 2002-1, Faculty of Economics, Kanazawa University, Kanazawa.

Hoshino, N. (2003). Random clustering based on the conditional inverse Gaussian-Poisson distribution, *J. Japan Statist. Soc.*, **33**, 105–117.

Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment, *J. Japan Statist. Soc.*, **28**(2), 125–134.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1993). *Univariate Discrete Distributions*, 2nd ed., Wiley, New York.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Multivariate Distributions*, Vol. 1, 2nd ed., Wiley, New York.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, Wiley, New York.

Kemp, A. W. (1978). Cluster size probabilities for generalized Poisson distributions, *Comm. Statist. Theory Methods*, **7**, 1433–1438.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.

Levin, B. and Reeds, J. (1977). Compound multinomial likelihood functions are unimodal: Proof of a conjecture of I. J. Good, *Ann. Statist.*, **5**, 79–87.

Maceda, E. C. (1948). On the compound and generalized Poisson distributions, *Ann. Math. Statist.*, **19**, 414–416.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*, W. H. Freeman and Company, New York.

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, *J. Roy. Statist. Soc. Ser. A*, **154**, 305–340.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions, *Biometrika*, **49**, 65–82.

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.

Pareto, V. (1897). *Cours d'Économie Politique*, F. Rouge, Lausanne.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions, *Probab. Theory Related Fields*, **102**, 145–158.

Preston, F. W. (1948). The commonness, and rarity, of species, *Ecology*, **29**, 254–283.

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.

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

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.

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.

Seshadri, V. (1993). *The Inverse Gaussian Distribution*, Clarendon Press, Oxford.

Seshadri, V. (1999). *The Inverse Gaussian Distribution*, Lecture Notes in Statist., **137**, Springer,

New York.
Shaban, S. A. (1988). Poisson-lognormal distributions, *Lognormal Distributions*: *Theory and Applications* (eds. E. L. Crow and K. Shimizu), 195–210, Marcel Dekker, New York.
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.
渋谷政昭(1991). あるクラスタ数分布と，その同音語の解析への応用，応用統計学, **20**, 139–153.
Sibuya, M. (1993). A random clustering process, *Ann. Inst. Statist. Math.*, **45**, 459–465.
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, 統計調査制度とミクロ統計の開示(松田芳郎，濱砂敬郎，森 博美 編著), 145–167，日本評論社，東京.
渋谷政昭(2002). 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研究成果報告書「調査データの公有化における理論的問題」(課題番号 11480055), 25–34.
Sibuya, M., Yoshimura, M. and Shimizu, R. (1964). Negative multinomial distribution, *Ann. Inst. Statist. Math.*, **16**, 409–426.
Sichel, H. S. (1973). The density and size distribution of diamonds, *Bull. Int. Statist. Inst.*, **45**, 420–427.
Skinner, C. J. and Holmes, D. J. (1993). Modelling population uniqueness, *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin, 175–199.
Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata, *Journal of Official Statistics*, **14**, 361–372.
Skinner, C. J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata, *Journal of Official Statistics*, **10**, 31–51.
Smith, T. M. F. (1976). The foundations of survey sampling: A review, *J. Roy. Statist. Soc. Ser. A*, **139**, 183–195.
Steutel, F. W. (1983). Infinite divisibility, *Encyclopedia of Statistical Sciences*, Vol. 4, 114–116, Wiley, New York.
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研究成果報告書「統計データの個票開示における開示制限の決定理論的評価」(課題番号 09206102), 95–104.
Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques, *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.
Thorin, O. (1977). On the infinite divisibility of the lognormal distribution, *Scand. Actuarial J.*, 121–148.
Urzúa, C. M. (2000). A simple and efficient test for Zipf's law, *Economics Letters*, **66**, 257–260.
Wani, J. K. and Lo, H. P. (1986). Selecting a power-series distribution for goodness of fit, *Canad. J. Statist.*, **14**, 347–353.
Watterson, G. A. (1974). Models for the logarithmic species abundance distributions, *Theoretical Population Biology*, **6**, 217–250.
Watterson, G. A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model, *J. Appl. Probab.*, **13**, 639–651.
Yamato, H. and Sibuya, M. (2000). Moments of some statistics of Pitman sampling formula, *Bull. Inform. Cybernet.*, **32**, 1–10.
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, Washington, D.C.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, Massachusetts.

Proceedings of the Institute of Statistical Mathematics Vol. 51, No. 2, 297–319 (2003) 319

Microdata Disclosure Risk Evaluation with Superpopulation Models: A Review

Nobuaki Hoshino

(Faculty of Economics, Kanazawa University)

Microdata identify each record's position in the corresponding contingency table. Hence, the anonymization of microdata is nothing but coarsening the resolution of a contingency table. Because a finer table results in the decrement of individuals in a cell, the frequency of cells of the same frequency of individuals plays an important role in the evaluation of disclosure risk. However, the estimation of the frequencies of frequencies is practically impossible without an assumption on a population. Based on the standard theory of finite population analysis, we employ superpopulation models as an assumption for the estimation. Then Zipf's law empirically validates the use of a Poisson distribution mixed by a distribution with a heavy tail. A basic population model assumes that the frequency of individuals in a cell is subject to an independent identical mixed Poisson distribution. Let the law of small numbers indicate a limiting argument that lets the number of cells be infinity, where (the expectation of) the total number of individuals is fixed. A proper model arises by applying the law of small numbers to a basic model that consists of infinitely divisible mixed Poisson distributions. Because there are many kinds of infinitely divisible mixed Poisson distributions, it may be possible to derive useful new models with the law. In order to describe various populations, developing new models is of great importance.

Key words: Population unique, infinite divisibility, compound Poisson, mixed Poisson, random partitioning of the natural numbers.