様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　4 月　16 日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18730145
研究課題名（和文）　疑似多項分布による個票開示リスク評価

研究課題名（英文）　Disclosure risk assessment using the quasi-multinomial distribution

研究代表者
星野　伸明　（HOSHINO NOBUAKI）
金沢大学・経済学経営学系・准教授
研究者番号：00313627

研究成果の概要：まず実務的な成果を説明する。研究代表者は、個票開示リスク評価に用いるモデル族の合理的な構成方法を提示した。疑似多項分布はこの族のメンバーであり、特に個票開示リスク評価に必要な結果を与えた。数理的な成果としては、離散多変数分布の族を提案して性質を評価した。また、確率分割の族を小数法則によって特徴づけした事が重要である。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,200,000	0	1,200,000
2007 年度	1,400,000	0	1,400,000
2008 年度	1,300,000	390,000	1,690,000
年度			
年度			
総　計	3,900,000	390,000	4,290,000

研究分野：統計学
科研費の分科・細目：経済学・経済統計学
キーワード：個票開示リスク、母集団一意、プライバシー、寸法指標

１．研究開始当初の背景

ミクロ計量分析の広がりを受けて、日本でも個票データの需要が高まっている。しかし、調査客体のプライバシー確保はデータの公開に明らかに優先する。プライバシーを確保しつつ、個票データを公開する方策を研究する必要がある。本研究の問題意識は、このようなプライバシーの危険性を正確に評価する事である。

この危険性評価は、分割表解析と密接な関連を持つ。個票と対応する個体は、分割表内の特定セルに所属する。あるセルに所属する個体が少なければ、その中から個体を識別するのは比較的容易である。従ってプライバシーのリスク評価は、小さいセル数の推定に帰着させるのが普通である。一般に、ある大きさのセル数を「頻度の頻度」または「寸法指標」と言い、これを正確に推定する事が大切である。ただこの問題については、通常の有限母集団からの標本抽出理論援用では不十分な結果しか得られない。従って本研究では、母集団を確率モデルとして記述する方法論（超母集団モデル）を採用する。このように考えた場合、母集団を良く記述できる柔軟なモデルが必要である。

Hoshino (2005a)では、任意の複合ポアソン分布から、条件付けと極限操作を組み合わ

せて自然数の確率分割モデルを生成可能と指摘した。このように生成される分布を族とみなして解析する視点は、確率論の文脈でも存在しない。また応用上は、この基本的関係を利用して新規モデルを構成可能である。ここで自然な興味の対称は、解析的操作が容易で当てはまりの良い複合ポアソン分布のモデルである。ここ数年の研究（特に Hoshino (2005b)）からは、ラグランジュアン・ポアソン分布(Consul and Jain, 1973)が有望と判断される。疑似多項分布は、この分布から派生する。これらの分布は、個票開示リスク評価分野で研究代表者以外は用いていない。また統計学の全分野を見渡しても、研究の蓄積が薄い。従って「柔軟なモデル」の候補として、疑似多項分布及びそこから派生するモデルが有力と考えられる。

疑似多項分布自体は 1975 年に提案されており、目新しいものではない。しかしその利用は、限られた範囲にとどまっている。総個体数が所与の場合、多項分布や負の超幾何分布を利用しろと教科書に書いてあるのが理由であろう。にもかかわらず本研究が疑似多項分布を持ち出す理由は、主に二点ある。(1)各セルの個体数が独立に複合ポアソン分布に従うとしてその総和で条件付けたモデルは、個票を秘匿する処置について閉じる（処理後も同分布となる）。この性質は、モデルに対する自然な要求と思われる。疑似多項分布は、このようなモデルの例である。(b)疑似多項分布は、（多項分布や負の超幾何分布と違い）分散をコントロールする母数を持つ。従って疑似多項分布は、幅広いデータを柔軟に記述出来ると期待される。

２．研究の目的

疑似多項分布及びそこから派生する分布について、応用と理論的評価を並行させる。特に、(1)（セルが互換でない）一般の疑似多項分布の応用 (2)極限疑似多項分布を含む分布族の組み合わせ論的解釈、を研究対象とする。(1)では、リスクに比例する母数をセル毎に変えたモデルを実用に供する。(2)については、ベル多項式を母関数とする分布族に共通する性質を理論的に評価する。

３．研究の方法

まず一般の疑似多項分布によるリスク評価が出来るような方法を主な研究課題とする。既に述べたとおり、まず分割表のセルを安全／危険という二群に分け、それぞれの郡内で各セルが互換というモデルを考察する。以下の議論は明らかに二以上の多群に拡張出来るが、それは後年度の課題とする。二群への分類は応用上の問題意識に沿うものであり、特に意味がある事を強調しておく。セルの総数が J として、安全群のセル数を K とすれば危険群のセル数は(J-K)と定まる。安全群のセルの母数を p_1、危険群のセルの母数を p_2 とすれば、K*p_1+(J-K)*p_2=1 という制約より、J,K 所与で p_1 を与えれば p_2 も定まる。問題は、観測された寸法指標を二群に分離する方法である。標本の個体は、どちらの群に所属するか観測出来ない。ただ危険群の方がセルに属する個体数が小さめという事から、p_1>p_2 でなければならない。この順序制約を使って識別する方法を考えたい。このアイディアは、有限混合分布の考え方と近いと思われる。有限混合分布は例えば二群混合の場合、Pr(X=1)=p_1*Pr(X=1;theta_1)+p_2*Pr(X=1;theta_2)のようにみなす。このような考え方は古くから用いられており（例えば Sundberg (1974), Scandinavian Journal of Statistics)、その研究成果を調べる事で識別性の問題を解決できると期待される。この問題が解決出来れば、プログラムを書いて数値的評価に進む。

次に理論的研究について述べる。各セルの度数 F_1,F_2,…,F_J がそれぞれ独立に複合ポアソン分布に従うとしよう。ここで複合ポアソン分布は正の母数 theta について、確率母関数 G(z)=exp(theta(g(z)-1))、ただし g(z):=Σ_{i=1}^{∞}z^iq_i は正の整数上の分布{q_i}_{i=1}^{∞}の確率母関数、で定義される。この g(z)は、いかなる正則な分布の確率母関数でも良い。応用面では解析的に操作が容易なモデルが得られる g(z)が興味の対象だが、理論的な性質評価は一般の g(z)について行うべきだろう。もし g(z)がボレル分布の確率母関数なら、G(z)はラグランジュアン・ポアソン分布の確率母関数となり、本プロジェクトの興味の対象となる。なお本研究の枠組みでは、j 番目のセル度数 F_j が、確率母関数 exp(theta_j(g(z)-1))で定義される複合ポアソン分布に従うとする。この場合、度数の総和 N:=F_1+F_2+…+F_J は、確率母関数 exp((theta_1+…+theta_J)(g(z)-1))で定義される複合ポアソン分布に従う。従って、条件付き分布 Pr(F_1,F_2,…,F_J|N)を計算する際の困難がかなり回避出来る事となる。疑似多項分布は、F_j がラグランジュアン・ポアソン分布に従う場合の条件付き分布 Pr(F_1,F_2,…,F_J|N) である。Hoshino (2005a)では theta_1=theta_2=…=theta_J の場合を考察した。この場合寸法指標の分布 Pr(S_1,…,S_N|N)を容易に評価出来る。ただし、S_i は大きさ i のセル数である。また小数法則に相当する極限操作で自然数の確率分割モデルを得られる。すなわち共通の母数を theta として、J*theta を定数μに固定、J

を無限大とする。ここで J=S_0+S_1+…+S_N なので、セル数Jを無限大にした結果、モデルが S_0 に依存しなくなる。極限分布 Pr(S_1,S_2,…,S_N|N)は g(z)で規定され、自然数Nを確率的に分割する。g(z)を変える事で様々な Pr(S_1,S_2,…,S_N|N)が得られるので、このような分布の集合を族と考える事が出来る。ボレル分布の g(z)から生成されるこの族のメンバーが「極限疑似多項分布」であり、Hoshino (2005b)で考察した。さて、このようなPr(S_1,S_2,…,S_N|N)は、各S_iが平均μ*q_i のポアソン分布に独立に従う場合の同時分布(S_1,S_2,…)のN所与での条件付き分布になっている。

研究のアイディアは、g(z)がベキ級数分布族に入るとする事である。この場合 Pr(N=n)は、偏ベル多項式の重み付き和で書ける。またN所与の分布Pr(S_1,S_2,…,S_N|N)についても、偏ベル多項式の重み付き和を基準化定数とする分布になる。またこの分布について、空でないセル総数 U=S_1+…+S_N の分布も偏ベル多項式の重み付き和を分解すれば求められる。なお多くの実用的な分布がベキ級数分布のメンバーなので、本研究の制約は厳しいとは言えない。偏ベル多項式の性質は組み合わせ論の分野で研究されているはずなので、その成果を利用して漸近論等を展開できると思われる。組み合わせ論の専門図書を、サーベイする。

４．研究成果

一般の疑似多項分布については、安全群・危険群モデルであっても母数の識別性の問題が残る可能性が見えてきた。従って、疑似多項分布を含む分布族の性質評価をする事とした。この族を条件付き複合ポアソン(CCP)分布族と呼ぶ。CCP 分布族を分割表のモデルとして考えた時、これがセルの併合について閉じる事、セル確率に相当する母数を持つ事、過分散になる事、などをあきらかにした。疑似多項分布は CCP 分布の例になっているので、これらの結果は疑似多項分布についても成立する。そして特に疑似多項分布については、周辺分散の簡潔な表現や、母数推定法を明らかにし、良くあてはまるデータセットの実例を報告した。

極限疑似多項分布を含む確率分割のクラスを、極限条件付き複合ポアソン(LCCP)分布族と呼ぶ。LCCP 分布族については、まずこれが正の実数上の無限分解可能分布の離散化と言える事を指摘した。例として極限条件付き逆ガウスポアソン分布を取り上げ、性質を評価した。そしてここで示された性質をベル多項式で置き換えれば、族の性質評価が可能となる事を明らかにした。また LCCP 分布族は、いわゆる小数法則で特徴づけられる事を示した。この事から、他の確率分割族との関連が明白となる。

これらの成果を基礎に、個票開示リスク評価実務を理論的に整理する事が出来た。結果として、これまで曖昧であった点をいくつか明確化する事となった。特に、リスク評価に用いる超母集団モデル族の構成方法について、合理的な解決を見た。また、局所的匿名化手法や摂動的な匿名化手法のリスク評価を、既存のリスク測度と整合的に説明可能となった。

本研究課題の成果は理論的に新しいアイディアを多く含んでおり、公表される雑誌論文の審査の過程で、査読者に好評であった。今後は、ベル多項式による分布族の性質解析を更に進め、これを官庁統計実務の改善へフィードバックする予定である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2 件）

(1) Nobuaki HOSHINO, The quasi-multinomial distribution as a tool for disclosure risk assessment, To appear in Journal of Official Statistics, 2009, 査読有.

(2) Nobuaki HOSHINO, A discrete multivariate distribution resulting from the law of small numbers, Journal of Applied Probability, Vol. 43, 852-866, 2006, 査読有.

〔学会発表〕（計　6 件）

(1) Nobuaki HOSHINO, On a class of random partitioning distributions, Computational Algebraic Statistics, Theories and Applications (CASTA 2008), 2008/12/10, 京都大学会館

(2) 星野伸明, 複合ポアソンから生成される確率分割族について, 統計関連学会連合大会, 2008/9/9, 慶應義塾大学

(3) 星野伸明, 疑似多項分布による個票開示リスク評価（続）, 統計関連学会連合大会, 2007/9/8, 神戸大学

(4) Nobuaki HOSHINO, A family of distributions closed under anonymization, The 56th session of the ISI, 2007/8/23, Lisbon:Portugal

(5) 星野伸明, ギブス分割と複合ポアソンモデリング, 統計関連学会連合大会, 2006/9/6, 東北大学

(6) 星野伸明, 特許取得の大学間格差-確率分割の応用, 応用経済学会, 2006/6/10, 福岡大学

〔その他〕

ホームページアドレス：

http://stat.w3.kanazawa-u.ac.jp/owner/papers.html

６．研究組織
(1)研究代表者

星野　伸明 (HOSHINO NOBUAKI)
金沢大学・経済学経営学系・准教授
研究者番号：00313627

(2)研究分担者
なし

(3)連携研究者
なし