様式 C-19

科学研究費補助金研究成果報告書

平成２１年　６月　１日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：１９５００２３９
研究課題名（和文）　多カテゴリ離散分布におけるカテゴリ間相関構造の統計的推測への影響評価

研究課題名（英文）　Effects of Category Correlations on Statistical Inference in Discrete Categorical Distributions.

研究代表者
越智　義道（OCHI YOSHIMICHI）
大分大学・工学部・教授
研究者番号：60185618

研究成果の概要：カテゴリを反応として得るような離散データの分析について検討するために，相関をもつ多カテゴリ離散分布に関わる構造上の特性について調査した。多項分布の条件付二項分布分解に基づく混合化から導かれる相関や連続分布の離散化によって得られる分布のカテゴリ間相関について検討を加えた。これらの結果得られた相関の様子をもとに，統計的推論について相関構造の変化の平均構造の推測への影響について検討した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
１９年度	1,400,000	420,000	1,820,000
２０年度	700,000	210,000	910,000
年度			
年度			
年度			
総　計	2,100,000	630,000	2,730,000

研究分野：統計科学
科研費の分科・細目：統計科学
キーワード：離散データ，カテゴリ間相関，多項分布，混合化，統計的推測

１．研究開始当初の背景

離散データの解析において，典型的に用いられる分布として，二項分布や，多項分布がある。ところが,これらの分布を用いてデータの分析を行うと，分析に用いたモデルから得られた，反応の変動の範囲が，十分にデータの変動を取らえることができない現象が起きることが知られている。このようなデータ変動は超過変動と呼ばれている。二項反応のケースで超過変動が生じている場合には，その分布を拡張してモデルによる反応の変動の範囲を広げて分析を行うことが行われるが，その拡張法には典型的に２種類の拡張法がある，それはベータ二項分布のように反応確率に確率的な変動を想定し，分布の混合化を図る方法と観測個体が反応カテゴリのうちどれか１つを反応として選択するものとし，その個体の集合を観測単位としてとらえ，これらの個体間に相関を想定することによって超過変動をモデル化する方法である。当該研究者はこれまで，超過変動の処理について研究してきており，多項反応への拡張を試みてきていた。

多項反応における超過変動の扱い方としては理論的には二項分布の場合と同様に，混合

化と観測個体間相関の観点からモデル化が可能である。混合化に関してはディリクレ多項分布が比較的簡潔な分布表現を与え，超過変動への対応を可能にしている。ところが，一般的には，それ以外の混合化や相関の導入では必ずしも得られた分布は扱いやすいものでなく，一般化線形モデルの拡張あるいは一般化推定方程式によるアプローチが用いられることが多い。基本的にはこれらの分析において超過変動を扱う場合には，ディスパーションパラメータを導入して変動分をとらえることが行われる。ところがこのアプローチは多項分布におけるカテゴリ相関について多項分布の分散共分散構造のそれを想定し，それの定数倍としての構造を基本とした分析となり，その基礎分布としてディリクレ多項分布を想定することに対応する。これは，超過変動の処理としてはあまりに強い制約として考えられるため，超過変動としてのカテゴリ間相関の変化の様相とそれに対応したより柔軟な分析モデルの提案が求められていた。

２．研究の目的

この研究では，これらの問題に応えるために，

(1)多項分布を基礎分布とした際に生じる超過変動の生成機序とそのカテゴリ間相関への影響の範囲について調査すること

(2)ディリクレ多項分布に基づく，カテゴリ相関を越えて，より柔軟な相関構造を許す分析アプローチについて考察し，その相関構造の統計的推測に及ぼす影響について検討すること

を研究目的とした。

３．研究の方法

(1)離散カテゴリ分布の超過変動生成のメカニズムとしては１．で述べたように，反応確率の混合化あるいは反応単位間の相関により構成することが可能である。ここでは多カテゴリを持つ離散反応についてカテゴリ間相関構造の変化に焦点を絞り，次の３通りの方法により検討を行った。

①多項分布の混合化の際に用いられる典型的な分布はディリクレ分布であり，混合化の結果得られる分布はディリクレ多項分布と呼ばれる。このディリクレ多項分布は，多項分布の条件付二項分布分解に対して,分解された二項分布に対して，ベータ分布による混合化したものとしてのベータ二項分布による表現が可能である。この拡張の方法に対して，各二項反応に対して，正規分布・ガンマ分布に拡張について調査した。

一般に多項分布は，r をカテゴリ数，n を観測単位の観測総数，y_l $(l=1,2,\ldots,r)$ を各カテゴリの観測数とすると，

$$P(y_1,\ldots,y_r)=\frac{n!}{\prod_{l=1}^{r} y_l!}\prod_{l=1}^{r} p_l^{y_l} \qquad (1)$$

$$=\prod_{j=1}^{r-1}\binom{S_j}{y_j}\theta_j^{y_j}(1-\theta_j)^{S_j-y_j},\quad \theta_j=\frac{p_j}{1-F_j},$$

$$S_j=N-\sum_{l<j} y_l,\quad F_j=\sum_{l<j} p_l$$

なる分解が可能である。この分解は，各カテゴリにおいて，その右にあるカテゴリを結合してその２つのカテゴリ(群)について２値反応化して反応を考えることに等しい。ディリクレ多項分布では，この第２式の各二項分布表現の各θ_jについて，直接的に(0,1)区間の値をとるベータ分布による混合を考えることになる。この分解に基づく混合化によってカテゴリ間相関の構造変化について検討した。

②多項分布の直接的な混合化として，多変量正規分布による混合化の方法とそのカテゴリ間相関への効果について調査した。

③多変量正規分布において，変量間に相関構造を想定し，各変量を離散化し，その変量の総和を反応と考えることによって，カテゴリ間の相関構造を変化させ，①の結果とも比較検討した。

(2)４．に示す(1)の検討から，種々のモデルから導出されるカテゴリ間相関の構造から，超過変動の処理の際に用いられる，分散共分散行列は，ディリクレ多項分布を基礎とする場合のように分散共分散構造のスケール変化としては捉えられないことが分かった。したがって,通常の一般化線形モデルや一般化推定方程式で用いられるような超過変動パラメータの導入では分析が適切でなくなる可能性が確認できた。このために，ここでは，より一般的な推定方程式構造による分析の方式について検討を加えた。

４．研究成果

(1)離散カテゴリ分布のカテゴリ間相関の変化について

①この混合化の分布パラメータを変化する

と同時に，このθ_jのロジット変換に関して正規分布，対数変換に対してガンマ分布を適用して，混合化分布の，特にその分散の変化に対するカテゴリ間相関の挙動について検討することにした。

表１：二項分解における混合化の場合の３項反応の反応分散共分散

多項分布の場合		
2.223637	-1.11243	-1.11121
-1.11243	2.224133	-1.11171
-1.11121	-1.11171	2.222918
二項分解での正規混合化		
4.04263	-2.0213	-2.02133
-2.0213	3.73573	-1.71443
-2.02133	-1.71443	3.735754
二項分解でのガンマ混合化		
4.18752	-2.11538	-2.07214
-2.11538	3.757656	-1.64228
-2.07214	-1.64228	3.714414

表１はその結果の一部を示したものであるが３カテゴリ，観測単位での観測総数 10,反応確率はそれぞれ 1/3 に調整し，分散については，二項反応の分散相当の調整を行ったときの結果である。正規混合化，ガンマ混合化いずれも多項反応のカテゴリ間分散共分散からのずれを見て取ることができる。一方でガンマ混合化と正規混合化との相違は大きなものではない。

②多項分布の直接的な混合化については，多変量正規分布を利用し，各変量の逆ロジスティック変換により，反応確率の確率変数化に基づいて混合化することとした，このときの基礎変数間の相関と，反応確率における相関との関連について調査した。

表２：３変量正規分布(逆ロジスティック変換)混合化の場合の３項反応の反応分散共分散

相関係数	３変量正規分布による混合化		
0.1	4.138778	-2.06585	-2.07293
	-2.06585	4.129901	-2.06405
	-2.07293	-2.06405	4.136979
0.5	3.422402	-1.69655	-1.72585
	-1.69655	3.417449	-1.7209
	-1.72585	-1.7209	3.446751
0.8	2.772587	-1.3712	-1.40139
	-1.3712	2.758698	-1.3875
	-1.40139	-1.3875	2.788881

表２で示した結果は多変量正規分布の分散共分散構造はスプリットプロットタイプとしたときの結果の一部である。これから分かるように，相関係数が大きくなると反応確率の変動が少なくなることから，分散が抑えられる傾向になることが分かる。さらにこれらの相関構造は比較的ディリクレ多項反応のそれに近いものが得られた。この結果は，変量間の相関，つまり無拘束変量の相関の程度よりはその変量の総和が１と固定されることによってもたらされる拘束から生じる相関の方が強く影響し，結果的にディリクレ分布の構造に近い結果が得られているものと推測される。

③多変量正規分布において，変量間に相関構造を想定し，各変量を離散化し，その変量の総和を反応と考えることによって，カテゴリ間の相関構造を変化させることが可能である。この形態での相関の範囲について検討した。

結果の一例を示すと次のような反応分散共分散が得られることになった。

表３：３項反応における反応分散共分散

Correlation			
0	2.222222	-1.11111	-1.11111
	-1.11111	2.222222	-1.11111
	-1.11111	-1.11111	2.222222
0.1	3.419387	-1.13233	-2.28705
	-1.13233	2.264436	-1.1321
	-2.28705	-1.1321	3.419155
0.5	8.707037	-1.7834	-6.92364
	-1.7834	3.550555	-1.76716
	-6.92364	-1.76716	8.690803
0.8	13.82058	-3.61803	-10.2026
	-3.61803	7.306142	-3.68811
	-10.2026	-3.68811	13.89066

表３は表２と同様には多変量正規分布においてスプリットプロット型の分散共分散構造を仮定し，相関係数を０から 0.8 まで変化させたとき，３カテゴリの生起確率を等しく設定したうえで，観測総数 10 の場合の反応の分散共分散を求めたものである。相関係数が０の場合が多項分布あるいはディリクレ多項分布の基礎構造部に相当するが相関係数が上昇するにつれてその相関構造からのずれが顕著になる様子が分かる。

④①から③までの調査から②の結果ではディリクレ多項分布の反応分散共分散構造に比較的近いものが得られることがあったものの，それ以外では，それからのずれが示されることとなった。つまりカテゴリ間相関は相当に限定的であり，より柔軟な相関構造のモデル化が必要であることが示唆された。

(2)(1)での検討から，カテゴリ間相関の構造は，基本的に超過変動の処理の際に用いられる，ディリクレ多項分布を基礎とする，分散共分散構造のスケール変化としてはとらえられないことが分かった。このことをふまえて,通常の一般化線形モデルや一般化推定方程式で用いられる平均・分散共分散構造とディスパーションパラメータによる方法を拡張した分析方法について考察した。

①観測単位における多項反応を示す確率変数を

$$\mathbf{Y}_i = (Y_{i1},\ldots,Y_{ir})^{\mathrm{T}} \qquad n_i = \sum_{l=1}^{r} Y_{il}$$

その反応を説明するための共変量を

$$\mathbf{x}_i = (x_{i1}, x_{i2},\ldots, x_{iq})^{\mathrm{T}}$$

とし，反応確率をその共変量の関数として考え，

$$\mathbf{p}_i = (p_{i1},\ldots,p_{ir})^{\mathrm{T}}$$
$$= \mathbf{p}(\mathbf{x}_i) = (p_1(\mathbf{x}_i),\ldots,p_r(\mathbf{x}_i))^{\mathrm{T}}$$

とする。観測総数は $N(i=1,\ldots,N)$とする。さらに，多項分布(1)に基づく平均と分散を

$$\mathrm{E}(\mathbf{Y}_i) = n_i\mathbf{p}_i,$$
$$\mathrm{Var}(\mathbf{Y}_i) = n_i(\mathrm{diag}(\mathbf{p}_i) - \mathbf{p}_i\mathbf{p}_i^{\mathrm{T}}) = n_i\Delta(\mathbf{p}_i)$$

と書くことにする。

このとき，反応確率 $\mathbf{P}_i$に関する混合化分布の分散あるいは個体間相関を$\boldsymbol{\Sigma}$書くとすると拡張された分布における平均・分散構造は

$$\mathrm{E}(\mathbf{Y}) = n\mathbf{p},$$
$$\mathrm{Var}(\mathbf{Y}) = n(\Delta(\mathbf{p}) + (n-1)\boldsymbol{\Sigma})$$

と書くことができる。従って，ここでは，一般化推定方程式について，

$$\sum_i \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i^* - n_i\mathbf{p}_i^*) = \mathbf{0}$$
$$\mathbf{y}_i^* = (y_{2i},\ldots,y_{ri})^{\mathbf{T}}, \quad \mathbf{p}_i^* = (p_{2i},\ldots,p_{ri})^{\mathbf{T}}$$
$$\mathbf{V}_i = n_i\mathbf{C}(\mathbf{p}_i^*)\boldsymbol{\Phi}_i\mathbf{C}(\mathbf{p}_i^*)^{\mathrm{T}},$$
$$\boldsymbol{\Delta}(\mathbf{p}_i^*) = \mathbf{C}(\mathbf{p}_i^*)\mathbf{C}(\mathbf{p}_i^*)^{\mathrm{T}},$$
$$\boldsymbol{\Phi}_i = \mathbf{I} + (n_i - 1)\boldsymbol{\Psi}_i,$$
$$\boldsymbol{\Psi}_i = \mathbf{C}(\mathbf{p}_i^*)^{-1}\boldsymbol{\Sigma}^*\left(\mathbf{C}(\mathbf{p}_i^*)^{\mathbf{T}}\right)^{-1}$$

のように考えて，反応確率に関する共変量効果を推定する方法を導出した。ここで，$\mathbf{p}_i$ は共変量$\mathbf{x}_i$から構成される計画行列$\mathbf{A}(\mathbf{x}_i)$による線形予測子$\boldsymbol{\eta} = \mathbf{A}(\mathbf{x})\boldsymbol{\beta}$の関数として表現され，$\mathbf{D}_i$は$\mathbf{p}_i$の$\boldsymbol{\beta}$に関する一階微分を示す行列である。カテゴリ間相関に関わる基本部分はこの反応確率$\mathbf{p}_i$の推定値を通じて，$\Delta(\mathbf{p}_i)$の分解$\mathbf{C}(\mathbf{p}_i)$により基本部が構成され，超過変動に関わる構造は$\boldsymbol{\Psi}$により導入されることになる。この超過変動構造行列の推定については，モーメント推定を用いることとして，無構造化推定，対角化推定，単位行列のスカラ倍の形態での推定方式を考えることとした。最後の単位行列のスカラ倍の構造の場合はちょうどディリクレ多項分布のカテゴリ間相関を想定した分析に対応することになる。推定されたパラメータの推測はその漸近正規性に基づくものとし，分散共分散推定はサンドイッチ推定量を用いることとした。

②上記で導出を行った分析アプローチについて，そのカテゴリ間相関の変化が共変量効果の統計的推測に関してどのような効果をもたらすかについてシミュレーション実験によって検討を加えた。

表４：共変量効果に関するシミュレーション結果

Case 1　β = 0.10	Multinom MLE	Dirichlet-Multinom MLE	GEE Unstruc. Disp.	GEE Diag. Disp.	GEE Scale Fac. Disp.
Mean of estimates	0.101	0.100	0.101	0.101	0.101
S.E. of estimates	0.070	0.068	0.070	0.070	0.070
Mean of S.E. estim.	0.041	0.067	0.068	0.068	0.068
Rejection Prob.*	0.62	0.32	0.32	0.32	0.31
Case 2　β = 0.10					
Mean of estimates	0.101	0.100	0.101	0.101	0.101
S.E. of estimates	0.074	0.070	0.073	0.073	0.070
Mean of S.E. estim.	0.041	0.072	0.071	0.072	0.068
Rejection Prob.*	0.61	0.27	0.30	0.29	0.32

表４は反応カテゴリを 3，共変量を 1 変数とし 0(1)4 とし，観測単位数を 100(各群 20)，観測個体での観測数を 8-12,反応確率の構造として累積ロジスティック回帰モデルを想定したときの結果の一部である。

ここでケース１はちょうど超過変動構造$\boldsymbol{\Psi}$についてディリクレ多項型の構造を想定した場合，ケース２は対角化構造を想定した場合を示している。このとき，ケース１では無構造，対角化構造であってもその結果はスケール倍のそれ，つまりディリクレ多項の結果を同等に示し，それからずれた状態では，ディリクレ多項のモデルでは超過変動をとらえきれていない状況を示している。ディリクレ多項分布の最尤法の結果は完全にモデルの適合ミスとなるため，推定誤差のそれが検出力に反映できていないことも示唆されている。この設定の他の多くの場合で同様の結果が得られた。

③以上のシミュレーションによる分析アプローチの挙動の結果から十分にデータの構造的な変化をとらえられることが確認でき

たことから，このアプローチによる実データの分析について検討した。

用いたデータはHydroxyureaデータ(Chenら，1991)である。これは，催奇形性試験データであり，雌マウスに交配後試験薬を投与し，出産直前に胎内の状況を調査し死産，奇形，正常の状態を記録したものである。

表５:Hydroxyurea データの分析　超過変動の調整
（括弧内は標準誤差を示す。）

	Multinom (MLE)	Dirichlet-Multinom (MLE)	Scale Factor (GEE)	Diagonal (GEE)	Unstruct (GEE)
Intercept(1)	-1.416	-1.315	-1.384	-1.327	-1.327
	(0.160)	(0.232)	(0.234)	(0.246)	(0.246)
Intercept(2)	-1.962	-1.920	-1.934	-1.872	-1.871
	(0.166)	(0.237)	(0.236)	(0.246)	(0.243)
Mid-Dose	1.804	1.765	1.780	1.724	1.722
	(0.203)	(0.310)	(0.356)	(0.363)	(0.364)
High-Dose	2.546	2.418	2.490	2.432	2.431
	(0.186)	(0.277)	(0.289)	(0.297)	(0.297)

表５は平均構造について累積ロジスティック回帰モデルを想定して分析を行った結果の要約を示しているが，パラメータの標準誤差の推定から,多項分布では十分にその変動がとらえられていないこと，また，ディリクレ多項分布(MLE)では改善はされているものの,今回の提案分析手法による，Ψに関するスカラー倍(Scale Factor)，対角化(Diagonal),無構造化(Unstructured)の構造化の結果から，ディリクレ多項分布の構造でも十分でないこと，さらに，構造的には対角化を想定することで十分な表現が得られることが確認された。

(3)本研究では超過変動を含む分析アプローチとして最尤法の枠組みと推定方程式によるアプローチについて検討を加えた。その結果，推定方程式による分析法によって，カテゴリ間相関の平均構造に関わる分析に関して，より柔軟な分析の方法を提案できた。このことにより，データの分析の際により詳細な構造化の検討を行うことが可能になった。

これまでChenらのデータの分析についてはディリクレ多項分布以上の構造的な分析に関しては詳細な議論が困難であったが，今回の分析によって，その超過変動のディリクレ多項分布からのずれについてずれの方向性に関する議論ができるようになった意義は大きい。

今後，(1)で調査したカテゴリ間相関にかかわるモデルについてその構造数理の調査を進め，詳細な数理モデルの可能性について検討を進める必要がある。ただし，これらのモデルの数理表現は確率関数について閉じた表現を与えない可能性が高いので，直接的な尤度法に基づく解析は制限されたものになる可能性があり，別のアプローチを検討する必要もある。その一つがEMアルゴリズムである。本研究でも限られた範囲でその適用について検討を加えたが，現時点では，収束性の安定化にいたらなかったために比較検討の観点から十分な調査を行うことができなかった。これは,基本的に潜在分布として連続多変量分布を想定し，その離散化としてデータをとらえ，データから潜在分布の計量値を推定してパラメータを推定する方式である。計量値の適切な推定値が得られれば，パラメータ推定は大きく簡素化できる可能性もある。また，計算機集約的な方法になるがMCMC法等により，シミュレーションを基礎としてモデル解析を進めることも視野に入れるべきである。観測個体数が限定的である場合については，標本空間全体に関する全数列挙あるいはその効率的な算出法に基づく正確推測の枠組みの中で評価される可能性を含んでいる。

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）

〔学会発表〕（計　４　件）

①　長本州彦(発表者), 柴田佳直, 越智義道, ロジスティック回帰分析での正確推測法における離散性への対応，シンポジウム「医学データ解析の数理的基盤」，2009.2.5.(大分市)

②　長本州彦(発表者)，越智義道，ロジスティック回帰モデルにおける条件付正確検定での２次元統計量の利用について，大分統計談話会第38回大会，　2008.10.16.(大分市)

③　T. Obata(発表者) and H. Ishii, “Evaluation of Nearness of Alternatives from Ranked Preference Data”, Proceedings of the 13th Asia Pacific Management Conference, 2007.11.19. (Melbourne: Australia).

④　越智義道，数理概念の具現化ツールR：離散データ解析への応用，大分統計談話会第36回大会，2007.10.19.（大分市）.

〔図書〕（計　１　件）

①　小西貞則，越智義道，大森裕浩，計算統計学の方法，朝倉書店，223(71-141)，2008

６．研究組織

(1)研究代表者

越智　義道（OCHI YOSHIMICHI）
大分大学・工学部・教授
研究者番号：60185618

(2)研究分担者

小畑　経史（OBATA TSUNESHI）
大分大学・工学部・助教
研究者番号：00244153