統計数理（1998）
第46巻 第1号 205-225

特集「ヘルスサイエンスと統計科学」
［原著論文］

# 超多項変動を持つデータの解析

大分大学* 越 智 義 道

（受付 1997年10月2日；改訂 1998年1月29日）

## 要 旨

この論文では超多項変動をもつカテゴリカルデータの回帰分析について考える．反応カテゴリについては順序をもつ場合も含む．

まず，多項分布における反応確率に関する共変量効果の評価法として，多項ロジットや累積ロジットなど反応の順序性の有無や性質をもとに定義される関連性指標に関する回帰分析に着目し，その基本的な考え方を紹介する．次に，この多項分布を用いたモデル適合で超過変動が示唆される場合について，パラメトリックな分布の拡張としてディリクレ-多項分布を考え，その分布のもとでの分析法について考察する．また，ディリクレ-多項分布の平均-分散構造にのみ着目してこれと同様な分析を行う方法として，一般化推定方程式による分析法の提案を行う．

さらに，これら2通りの分析法を用いて現実のデータの分析を行い，分析法の適用可能性と特性について考察を行う．

キーワード：超過変動，多項分布，ロジスティック回帰，順序応答，GEE.

## 1. はじめに

本報告では多値離散反応において超過変動(over-dispersion)を考慮せねばならない場合の共変量効果の評価について考える．

催奇形性試験のような実験で得られるデータの分析では，二項分布や多項分布を用いた分析を行うと，これらの分布にもとづいて想定された構造モデルによって説明できる反応の変動よりもデータの示す変動の方が大きくなることがある．このようなデータの変動のことを超過変動（over-dispersion）あるいは基礎となる分布を明示して超二項変動（extra-binomial variation）や超多項変動（extra-multinomial variation）と呼んでいる．催奇形性試験では，同腹仔（litter mates）間の相関や母体特性による胎児の反応の変動などによる同腹仔効果（litter effect）がこの超過変動に関係していると考えられている．

本報告では，Chen et al.（1991）によって報告された催奇形性試験データ（表1）を例として考えながらこのような超過変動を含むカテゴリカルデータに関する分析について考察を行う．この試験では，まず雌雄の実験動物（マウス）を交配させ，その後雌のマウスに対してhydroxyurea（オキシ尿素，白血病治療薬）の投与を行う．投薬後一定の妊娠期間経過の後，出産直前に解剖し，母体胎内中の胎児の状態，着床の痕跡などを調べhydroxyureaの催奇形性を評価しようとするものである．投薬量については，低，中，高用量の3水準のレベルを設定している．

このデータは各母体ごとの結果を，(I)胎児の正常発育，(II)発育はしているが奇形状態のも

* 工学部 知能情報システム工学科：〒870-1192 大分市旦野原700.

表1. Chen's hydroxyurea（オキシ尿素）data（Chen et al. (1991)），(I) 正常発育，(II) 奇形，(III) 死亡-吸収．

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Low dose | | | | | | | | | | |
| (I) | 13 | 11 | 11 | 3 | 11 | 6 | 9 | 6 | 11 | 14 |
| (II) | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 0 |
| (III) | 1 | 0 | 0 | 5 | 1 | 0 | 2 | 2 | 1 | 0 |
| (I) | 7 | 6 | 9 | 5 | 11 | 8 | 12 | 6 | 10 | 8 |
| (II) | 4 | 0 | 0 | 0 | 0 | 5 | 0 | 3 | 0 | 0 |
| (III) | 2 | 1 | 0 | 4 | 0 | 1 | 2 | 1 | 0 | 1 |
| (I) | 7 | 7 | 6 | | | | | | | |
| (II) | 2 | 2 | 3 | | | | | | | |
| (III) | 1 | 0 | 1 | | | | | | | |
| Medium dose | | | | | | | | | | |
| (I) | 12 | 11 | 13 | 2 | 1 | 3 | 9 | 8 | 1 | 7 |
| (II) | 0 | 0 | 1 | 0 | 0 | 5 | 1 | 0 | 1 | 2 |
| (III) | 2 | 0 | 12 | 9 | 12 | 5 | 3 | 3 | 10 | 4 |
| (I) | 2 | 1 | 1 | 4 | 3 | 2 | 10 | 0 | 9 | 8 |
| (II) | 4 | 0 | 1 | 1 | 1 | 6 | 1 | 3 | 1 | 0 |
| (III) | 3 | 9 | 9 | 8 | 4 | 2 | 1 | 8 | 2 | 4 |
| High dose | | | | | | | | | | |
| (I) | 1 | 0 | 4 | 1 | 2 | 7 | 5 | 1 | 0 | 0 |
| (II) | 1 | 3 | 1 | 0 | 2 | 1 | 0 | 0 | 1 | 1 |
| (III) | 7 | 8 | 6 | 10 | 7 | 3 | 4 | 9 | 7 | 4 |
| (I) | 1 | 3 | 2 | 1 | 5 | 4 | 4 | 1 | 1 | 4 |
| (II) | 0 | 4 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 4 |
| (III) | 11 | 5 | 10 | 10 | 6 | 8 | 6 | 8 | 11 | 2 |
| (I) | 1 | 6 | 10 | 1 | 0 | 1 | 1 | 2 | 1 | 1 |
| (II) | 0 | 1 | 0 | 2 | 2 | 1 | 0 | 0 | 1 | 1 |
| (III) | 11 | 2 | 1 | 7 | 10 | 9 | 8 | 7 | 9 | 10 |
| (I) | 1 | 4 | 5 | 1 | 2 | 1 | 0 | 7 | 2 | 2 |
| (II) | 1 | 4 | 1 | 0 | 3 | 1 | 1 | 1 | 3 | 2 |
| (III) | 12 | 4 | 3 | 11 | 7 | 10 | 15 | 1 | 4 | 5 |
| (I) | 1 | 6 | 7 | 5 | 8 | 0 | 3 | | | |
| (II) | 3 | 2 | 1 | 2 | 0 | 1 | 0 | | | |
| (III) | 7 | 4 | 2 | 4 | 3 | 11 | 7 | | | |

の，(III)胎内死亡あるいは着床の痕跡はあるものの途中で吸収状態のもの，に分類したものである．したがって基本的にはこの試験では3カテゴリの反応が得られ，それらに関する用量効果を調査すれば良いと考えられる．

このような同腹仔にかかわる試験で通常の多項分布を用いた分析を行うと，データへの適合性が悪い場合が多く，超過変動を考慮にいれたデータの分析が必要である．

従来の催奇形性試験では出生した胎児中の奇形数に着目して2カテゴリの反応としてとらえ，二項分布あるいは超過変動を考慮にいれて二項分布を拡張した分布を用いてパラメトリックに分析を行う方法や，反応の平均と分散に着目した疑似尤度法などによる用量反応分析が広く用いられてきている（Haseman and Kupper (1979), Williams (1982), Morgan (1992)）．たし

かに，このような手法で二項分布における超過変動を考慮にいれた分析が可能になるものの，分析の基礎を出生仔における条件付き確率においているために薬物の効果を充分に解析に反映しきれていないという指摘がなされている（Haseman and Kupper (1979)）．

ところが，この研究では薬物処置前に着床したと考えられる個所全体を調査していることから，反応を3項反応としてとらえることによって，2カテゴリのみに着目した場合のような条件付きの分析によらないデータの解析が可能になる．

この論文では多項反応に関する超過変動に対応した用量反応分析について，ディリクレ-多項分布を基礎とする尤度法にもとづく分析と，一般化推定方程式にもとづく分析について考察する．これらの分析については，その分散構造の処置が重要な問題であるが，用量効果に関する基本的なモデル化自身は反応確率のモデル化として扱うことができるので，まず基本となる多項分布によるモデル化に関して考えることにする．

## 2. 多項分布にもとづく分析

ここでは多項分布を持つ反応に関する共変量効果の分析について考える．以下説明の都合上，上記の母体のような観測の基礎となるまとまりのことを観測単位 (observational unit) と呼び，各胎児（着床部位）のように反応や計数（count）の基礎となる対象のことを反応に対応する個体と呼ぶことにする．

いま $r+1$ 個の反応カテゴリを持ち多項分布に従う独立な反応変数を

$$\boldsymbol{y}_i=(y_{i0},\ y_{i1},\dots,\ y_{ir})^T \qquad (i=1,\dots,\ N),$$

とし，観測単位 $i$ の反応の総数を $n_i=\sum_{k=0}^{r} y_{ik}$ とする．また，この観測単位 $i$ の試験割り付け環境や，試薬投薬量などの共変量を

$$\boldsymbol{x}_i=(x_{i1},\dots,\ x_{iq})^T$$

とする．先の例では，$y_{ij}$ $(j=0,1,\dots,r)$ は母体 $i$ での反応，つまり胎児（着床部位）の状態の集計結果を示すことになる．さらに，$\boldsymbol{y}_i$ に対応する反応確率を

$$\boldsymbol{p}_i=(p_{i0},\ p_{i1},\dots,\ p_{ir})^T$$

とする．用量反応分析など共変量効果を調査する場合には，これを $\boldsymbol{x}_i$ の関数ととらえて，

$$\boldsymbol{p}_i=\boldsymbol{p}(\boldsymbol{x}_i)=(p_0(\boldsymbol{x}_i),\ p_1(\boldsymbol{x}_i),\dots,\ p_r(\boldsymbol{x}_i))^T$$

と考えることになる．したがって，観測が多項分布に従うものと考えることができると，この $N$ 個の独立な観測からなる全観測

$$\boldsymbol{y}=(\boldsymbol{y}_1^T,\dots,\ \boldsymbol{y}_N^T)^T$$

に関する対数尤度 $l$ は

$$l=\sum_{i=1}^{N} l_i=\sum_{i=1}^{N}(c_i+y_{i0}\log p_0(\boldsymbol{x}_i)+y_{i1}\log p_1(\boldsymbol{x}_i)+\cdots+y_{ir}\log p_r(\boldsymbol{x}_i))$$

あるいは

$$=\sum_{i=1}^{N}(c_i+n_i\log p_0(\boldsymbol{x}_i)+y_{i1}\eta_{i1}+\cdots+y_{ir}\eta_{ir})$$

と書ける．ただし，ここで $l_i$ は観測 $\boldsymbol{y}_i$ の対数尤度 $l$ に関する寄与，

$$c_i=\log\frac{n_i!}{y_{i0}!\,y_{i1}!\cdots y_{ir}!},$$
$$\boldsymbol{\eta}_i{}^T=(\eta_i,\ \eta_{i2},\ \ldots,\ \eta_{ir})\qquad(\eta_{ik}=\log p_{ik}/p_{i0},\ k=1,2,\ldots,r)$$

であり，この $\boldsymbol{\eta}_i$ は指数型分布族でいうところのナチュラルパラメータ (Cox and Hinkley (1974)) である．多項反応における一般化線形モデルは，まず説明変数 $\boldsymbol{x}_i$ から作られる行列 $A_i$ と対応するパラメータ $\boldsymbol{\beta}$ から線形予測子

$$\boldsymbol{\zeta}_i=A_i\boldsymbol{\beta}$$

を構成し，この $\boldsymbol{\zeta}_i$ とナチュラルパラメータ $\boldsymbol{\eta}_i$ や確率 $\boldsymbol{p}_i$ とを対応づけることによって定義できる (Jørgensen (1983), McCullagh and Nelder (1989))．ここでは，パラメータ $\boldsymbol{\eta}_i$ を線形予測子 $\boldsymbol{\zeta}_i$ の関数

$$\boldsymbol{\eta}_i=\boldsymbol{\eta}(\boldsymbol{\zeta}_i)$$

として記述することにする．

このような一般化線形モデルの枠組みにおいてよく使われるモデルとしては，反応が順序のない名義尺度反応の場合，ベースラインカテゴリと各カテゴリに関する対数オッズ比をもとにした多項ロジスティック回帰モデルがある．また反応に順序がある場合には，反応確率について累積ロジット，隣接カテゴリロジット，連続比ロジット，コンプリメンタリ log-log などの順序構造を仮定した指標を定義し，これらの指標に関する推論を線形予測子 $\boldsymbol{\zeta}$ を通して行うことが考えられる (Agresti (1984, 1990), McCullagh and Nelder (1989)，越智 他 (1991))．

例えば，関連性尺度として多項ロジスティックを採用した場合には

$$\text{(2.1)}\qquad \mathrm{logit}_k(\boldsymbol{x}_i)=\log(p_k(\boldsymbol{x}_i)/p_0(\boldsymbol{x}_i))=\eta_k(\boldsymbol{x}_i)\qquad(k=1,2,\ldots,r)$$

として

$$\boldsymbol{\eta}_i=(\eta_1(\boldsymbol{x}_i),\ \eta_2(\boldsymbol{x}_i),\ldots,\ \eta_r(\boldsymbol{x}_i))^T=\boldsymbol{\zeta}_i$$

なる対応づけを行うことになり，ナチュラルパラメータと線形予測子が直接対応づけられる．反応に順序性を仮定しないために，(2.1)式の分母の確率に対応するベースラインカテゴリについて基本的には $r+1$ 個のカテゴリの中のどのカテゴリを選んでもよく，どの場合でもパラメータ推定値は変化するものの共変量効果に関する尤度比検定やモデル適合などの解析結果は不変である．

一方，反応に順序が想定できる場合には，順序情報を分析のなかに取り込むことによって，より簡潔なデータの記述が可能になることがある．例えば，累積ロジットでは

$$F_k(\boldsymbol{x}_i)=\sum_{j=0}^{k-1}p_j(\boldsymbol{x}_i)\qquad(k=1,2,\ldots,r)$$

としてロジット

$$\text{(2.2)}\qquad \mathrm{logit}_k(\boldsymbol{x}_i)=\log\left(\frac{1-F_k(\boldsymbol{x}_i)}{F_k(\boldsymbol{x}_i)}\right)\qquad(k=1,2,\ldots,r)$$

を定義し，このロジットと線形予測子 $\boldsymbol{\zeta}_i$ を対応づけてモデルを構成することができる．場合によっては，カテゴリ区分ごとに定義されるロジットについて共変量効果を共通化することもでき，このときのモデルはプロポーショナル-オッズモデルと呼ばれることもある (Agresti (1990))．

もちろん，このような一般化線形モデルの枠組みを越えて，いわゆる用量反応曲線の形状や，実質科学の枠組みの中での構造的な解釈の立場からモデルが定義されることもある (Morgan

(1992), Jansen (1990), Zhu et al. (1994), 他). 一般化線形モデルの場合はもとより，このような状況であっても共変量効果の評価のための構造が多項分布の反応確率のモデル化としてとらえることが可能ならば，多項分布を基礎とする最尤法を用いたデータの分析が可能である．

## 3. 超多項変動

2節で述べたようなモデルを用いることにより，最尤法の枠組の中でカテゴリカルデータの分析を行うことができ，さらに適合後のモデル診断も可能である（Lesaffre and Albert (1989), O'Hara Hines et al. (1992), 越智 他 (1994)）．

ところが，所与の共変量を適切に用い，考えうる十分なモデル化を行ったにもかかわらず，モデル適合が悪い場合がある．しかも，それが少数の異常値からなるものではなく，データ全般について適合が不十分であると判断される場合がある．典型的には，モデルから想定される変動よりもデータの方が大きな変動を示す場合が多い．このような過度の変動を超過変動 (over-dispersion) と呼んだり，あるいは基礎となる分布を多項分布とするときには，基礎分布を明確にするために超多項変動 (extra-multinomial variation) と呼ぶこともある．特に1節で紹介した催奇形性試験では超過変動をもつ傾向が強い．

### 3.1 超多項変動の発生機序

超過変動が生じていると考えられる場合には，その観測単位をなす対象（例えば催奇形性試験の場合には母体）の特性が反応（胎児の状況）に影響を及ぼしていると考えることが多い．このため，その超多項変動に関する発生機序としては，

1．観測単位をなす対象における反応率を所与のものと考えると，反応自身は多項分布に従うものと考えられるが，一定の条件下でも観測単位の反応率を固定的に考えることが難しく，反応率自身を確率変数とみなした方がよい場合．例えば，催奇形性試験の場合には，試験状況が全く同一のものであっても，母体のもつ胎内の着床受精卵の成長に関する特性について本質的な個体変動を想定しなければならない場合が考えられる．
2．一つの観測単位から得られる反応について，個々の反応がそれぞれ独立でないために，多項分布仮定からのずれが生じている場合．催奇形性試験の場合，同腹仔あるいは胎内の着床受精卵の間に相関を想定する場合がこれに相当する．

の二つを考える場合が多い．つぎの2節でこれらの場合の確率的性質について整理する．これ以降，特に混乱が生じない限り，観測単位（母体）を表す添字 $i$ は省略する．

#### 3.1.1 反応率変動の場合

上記のはじめの状況は多項分布に関する混合分布を想定することによって説明可能である．まず，ある観測単位のもつ特性値として反応率 $\boldsymbol{\Pi}$ が得られていると考える．$\boldsymbol{\Pi}$ が所与であるとき，観測単位から得られた反応 $\boldsymbol{Y}|\boldsymbol{\Pi}$ が条件付き分布として多項分布 Multinom（$\boldsymbol{Y}|n, \boldsymbol{\Pi}$）に従うと考え，このときの $\boldsymbol{Y}|\boldsymbol{\Pi}$ の平均と分散共分散行列を

$$E(\boldsymbol{Y}|\boldsymbol{\Pi})=n\boldsymbol{\Pi},\ \mathrm{Var}(\boldsymbol{Y}|\boldsymbol{\Pi})=n\Delta(\boldsymbol{\Pi})=n\{\mathrm{diag}(\boldsymbol{\Pi})-\boldsymbol{\Pi}\boldsymbol{\Pi}^T\}$$

と書くことにする．

つぎに，この $\boldsymbol{\Pi}$ が平均 $\boldsymbol{p}$ 分散共分散 $\Sigma$ をもつ多変量分布 $F$ に従うと考える．このとき反応 $\boldsymbol{Y}$ の周辺分布 $P(\boldsymbol{Y})$ は

$$P(\boldsymbol{Y})=\int \mathrm{Multinom}(\boldsymbol{Y}|n, \boldsymbol{\Pi})dF(\boldsymbol{\Pi})$$

と書け，$\boldsymbol{Y}$ の平均と分散共分散は

$$\begin{aligned} E(\boldsymbol{Y}) &= E_{\Pi}(E(\boldsymbol{Y}|\boldsymbol{\Pi})) = n\boldsymbol{p}, \\ \mathrm{Var}(\boldsymbol{Y}) &= E((\boldsymbol{Y}-n\boldsymbol{p})(\boldsymbol{Y}-n\boldsymbol{p})^T) \\ &= E_{\Pi}(n\Delta(\boldsymbol{\Pi}) + n^2(\boldsymbol{\Pi}-\boldsymbol{p})(\boldsymbol{\Pi}-\boldsymbol{p})^T) \\ &= n(\Delta(\boldsymbol{p}) + (n-1)\Sigma) \end{aligned}$$

となる．ここで分散共分散行列における $(n-1)\Sigma$ が超過変動分を説明する項に相当する．

#### 3.1.2 相関のある場合

つぎに相関を想定する場合について考えてみる．試験状況を同一にする観測単位（例えば母体）のなかの単一の反応に対応する個体 $l$ （例えば1胎児あるいは着床部位の1つ）における反応確率については同一の確率 $\boldsymbol{p}$ を想定し，その個体 $l$ の反応は多項分布 Multinom $(\boldsymbol{Y}_l|1, \boldsymbol{p})$ に従うものとする．ここで各個体が独立に反応すると考えれば，その母体において観測される反応 $\boldsymbol{Y} = \boldsymbol{Y}_1 + \boldsymbol{Y}_2 + \cdots + \boldsymbol{Y}_n$ は多項分布 Multinom $(\boldsymbol{Y}|n, \boldsymbol{p})$ に従い，その平均と分散共分散行列は

$$E(\boldsymbol{Y}) = n\boldsymbol{p}, \qquad \mathrm{Var}(\boldsymbol{Y}) = n\Delta(\boldsymbol{p}) = n\{\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\}$$

となる．ところが個体間の反応に相関があると考えると，反応 $\boldsymbol{Y}$ の分布は多項分布からずれてくる．いま，個体 $i$ と個体 $j$ の反応 $\boldsymbol{Y}_i, \boldsymbol{Y}_j$ に関する共分散を $\mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \Sigma_{ij}$ と考えると，このときの反応 $\boldsymbol{Y}$ の平均と分散共分散は

$$\begin{aligned} &E(\boldsymbol{Y}) = n\boldsymbol{p}, \\ &\mathrm{Var}(\boldsymbol{Y}) = E((\sum_i(\boldsymbol{Y}_i - \boldsymbol{p}))(\sum_j(\boldsymbol{Y}_j - \boldsymbol{p}))^T) = n\Delta(\boldsymbol{p}) + \sum_{i \neq j}\Sigma_{ij} \end{aligned}$$

となる．ここで反応間に等相関性 $\Sigma_{ij} = \Omega$ $(i<j)$ を仮定し，$\Sigma_0 = (\Omega + \Omega^T)/2$ とおくことによって反応率変動の場合と同様な分散共分散行列

$$\mathrm{Var}(\boldsymbol{Y}) = n(\Delta(\boldsymbol{p}) + (n-1)\Sigma_0)$$

を得ることができる．

### 3.2 発生機序の相違の意義

それぞれの超過変動の機序の違いを比較すると，相関をもとに超過変動を定義する際に用いられた $\Sigma_{ij}$ はあくまで共分散行列であり，反応率変動による定義の場合とは異なり分散共分散行列，つまり非負定値行列という性質をもつとは限らないことがあげられる．このことは反応率変動にもとづく超過変動の発生機序を考える場合には，基本的に超過変動（行列として大きい：行列差が非負定値）しか扱えないが，標本相関による発生機序を考えた場合には，過小変動（underdispersion）の場合も考慮に入れることができることを示している．

また，反応率変動にもとづく場合でも個体間の相関が導出される．このとき個体 $\boldsymbol{Y}_i, \boldsymbol{Y}_j$ 間の共分散行列は $\mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \Sigma$，つまり混合化分布の分散に相当し対称かつ同質であることが示される．

## 4. 超多項変動をもつ分布

前節で超多項変動について2種類の発生機序を紹介したが，一般的に標本相関をもとに現実的な $\boldsymbol{Y}$ の分布を構成するのは容易ではない．例えば二項分布の場合，相関を用いて超二項変動を表現する分布として Altham (1978) や Kupper and Haseman (1978) は correlated binomial

と呼ばれる分布を定義し，Paul (1987) はさらにそれを拡張した．しかし，いずれも分布パラメータに関する制約が強いため扱いづらく，また妥当な分布を与える相関の大きさも制限されてしまう．Ochi and Prentice (1984) は多変量正規分布を用いて相関構造を導入することを提案したが，数値計算上の問題をかかえている．

このため反応率変動にもとづいて $\boldsymbol{Y}$ の分布を構成する場合が多い．原則的には混合化に用いる分布に制約はなく，例えば Jansen (1990) は閾値モデルのなかで混合化分布として正規分布を用いている．ただ，多項反応の反応率変動の際に最もよく用いられる混合化分布はディリクレ分布（多変量ベータ分布）である．この分布は多項分布に対する共役分布であり，ディリクレ分布を混合化分布に用いて構成される周辺分布はディリクレ-多項分布と呼ばれている．二項反応の場合この分布はベータ-二項分布に等しい．

**ディリクレ-多項分布**

ディリクレ分布の密度関数は

$$f(\pi_0,\ \pi_1,\ldots,\ \pi_r)=\frac{\Gamma(\Theta)}{\prod_{k=0}^{r}\Gamma(\alpha_k)}\prod_{k=0}^{r}\pi_k^{\alpha_k-1}$$

$$(\Theta=\sum_{k=0}^{r}\alpha_k,\ \sum_{k=0}^{r}\pi_k=1,\ \pi_k>0,\ k=0,\ 1,\ldots,\ r)$$

と書ける．また，反応率としての確率変数 $\boldsymbol{\Pi}$ の平均と分散共分散は，$p_i=\alpha_i/\sum_{k=0}^{r}\alpha_k$, $\boldsymbol{p}^T=(p_0,\ p_1,\ldots,\ p_r)$ とすると，

$$E(\boldsymbol{\Pi})=\boldsymbol{p}, \qquad \mathrm{Var}(\boldsymbol{\Pi})=\phi\Delta(\boldsymbol{p})$$

と書ける．ただし $\phi=1/(1+\Theta)$, $\Gamma(x)$ はガンマ関数である．このとき反応 $\boldsymbol{Y}$ の周辺分布は

$$p(y_0,\ y_1,\ldots,\ y_r)=\frac{n!}{\prod_{k=0}^{r}y_k!}\frac{\Gamma(\Theta)}{\Gamma(n+\Theta)}\prod_{k=0}^{r}\frac{\Gamma(y_k+\alpha_k)}{\Gamma(\alpha_k)}$$

となり，その平均と分散共分散行列は

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

となる．さらに，同一母体からの胎児の間の共分散は $\phi\Delta(\boldsymbol{p})$ となる．このような分布をディリクレ-多項分布と呼ぶ．

このディリクレ-多項分布のパラメータに構造を導入し，反応の期待値について量反応関係を記述し，さらに必要があれば超過変動についても構造を定めておくことにより，最尤法によってデータの解析を行うことが可能である．

ただし，モデル化の場合にはパラメータの領域に注意が必要である．特に興味ある状況がパラーメータ空間の端にある場合には通常の最尤法における漸近理論が破綻をきたすので気をつけなければならない．超過変動に関するモデル化でディリクレ-多項分布を採用した場合，通常は混合化分布のパラメータについて $\alpha_0>0$, $\alpha_1>0,\ldots,\ \alpha_r>0$ なる条件を想定することが多いが，こうすると多項分布はこの分布族の端に位置することになり問題となる．とくに反応率変動を機序とするモデル化の場合，混合化分布の分散共分散によって超過変動を記述することになるので注意が必要である．

ちなみにベータ-二項分布の場合には Prentice (1986) が，一般のディリクレ-多項分布の場合について Zhu et al. (1994) がパラメータの範囲を負の範囲まで拡張しても分布として定義可能であることを指摘し，その限界の範囲を示している．このため，この拡張された分布族の中では通常の最尤法による多項分布の妥当性の調査，つまり超過変動の存在の有無にかかわる検定が可能である．また，パラメータ値が負になりえることから過小変動を示すことも可能にな

る．

さらにディリクレ-多項分布はベータ-二項分布による分解が可能である．Chen and Li (1992) はこの分解を利用したデータの解析例も示している．

## 5. 一般化推定方程式

前節までの議論により最尤法によるデータの解析が可能である．ところが最尤法のもとでデータの解析を行う限り指定された分布の妥当性は常に問題となる．適合の妥当性とデータ解析の推論の限界が共に適用された分布の特性に依存するという問題は避けられない．

そこで，一般化線形モデルあるいは疑似尤度法 (Nelder and Wedderburn (1972), Wedderburn (1974), McCullagh (1983), McCullagh and Nelder (1989)) などでは，基本となる分布の平均と分散の構造に着目した分析が可能となり，ある程度柔軟な分析が可能となる．二項反応の場合については Williams (1982)，Moore (1986, 1987) らが疑似尤度法にもとづいて超過変動が存在する場合の分析法を提案している．また，Ryan (1992) は超多項変動が存在する場合について，多項分布の平均-分散構造にスケールパラメータを導入した形の拡張を行い，疑似尤度法にもとづいた分析法について考えている．ここでは，一般化線形モデルを拡張した一般化推定方程式の枠組み (Liang and Zeger (1986), Zeger and Liang (1986), Prentice and Zhao (1991)) によるデータ解析法について考える．

先のディリクレ-多項分布の平均と分散の構造を用い，平均

$$E(\boldsymbol{Y}_i)=\boldsymbol{\mu}_i=n_i\boldsymbol{p}_i \qquad (\boldsymbol{p}_i=\boldsymbol{p}(\boldsymbol{x}_i))$$

については多項分布の場合と同様な共変量効果に関するパラメータの導入法を考える．また，分散共分散構造については

$$V(\boldsymbol{Y}_i)=n_i\Delta(\boldsymbol{p}_i)(1+(n_i-1)\phi)$$

を考え，この行列を一般化推定方程式の構成の際に用いる作業相関行列から作られた分散共分散行列に対応する行列と考えることにする．

超過変動を示すパラメータ $\phi$ についても必要に応じて共変量効果に関するモデル化を考える．以下の Chen のデータの分析では，$\boldsymbol{\gamma}$ を導入して，

$$\operatorname{logit}\phi_i = \log\frac{\phi_i}{1-\phi_i}=\boldsymbol{x}_i^T\boldsymbol{\gamma}$$

なるモデル化を考えた．もちろん，最も単純な場合には，共変量効果を考えずに，すべてのデータに共通な超過変動パラメータ $\phi$ が対応するモデルを考えることができる．

上記のモデルのパラメータ $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ の推定には次の 2 つの推定方程式を用いる．まず，平均構造にかかわるパラメータ $\boldsymbol{\beta}$ については

$$\sum_i D_{1i}^T V_{1i}^{-1}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_i^*)=0 \tag{5.1}$$

を考える．ただし $V_{1i}$ は上で想定した反応変数ベクトル $\boldsymbol{Y}_i^*$ の分散 $V(\boldsymbol{Y}_i^*)$ であり，$D_{1i}$ は $\boldsymbol{\mu}_i^*$ の $\boldsymbol{\beta}$ に関する導関数

$$D_{1i}=\frac{\partial\boldsymbol{\mu}_i^*}{\partial\boldsymbol{\beta}}=\left(\frac{\partial\boldsymbol{\zeta}_i}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial\boldsymbol{\mu}_i^*}{\partial\boldsymbol{\zeta}_i}\right)$$

である．また超過変動に関わるパラメータ $\boldsymbol{\gamma}$ については，Zhu et al. (1994) が提案している推定方程式と同様な

$$\sum_i D_{2i}^T V_{2i}^{-1}(Q_i-\sigma_i)=0 \tag{5.2}$$

を考える．ただし，ここで $Q_i$ は反応変数 $\boldsymbol{y}_i$ に関するピアソン $\chi^2$ 型の統計量

$$Q_i=(\boldsymbol{y}_i^*-n_i\boldsymbol{p}_i^*)^T\Delta^*(\boldsymbol{p}_i)^{-1}(\boldsymbol{y}_i^*-n_i\boldsymbol{p}_i^*),$$

$\boldsymbol{y}_i^*=(y_{i1},\ldots,y_{ir})^T$, $\boldsymbol{p}_i^*=(p_{i1},\ldots,p_{ir})^T$, $\boldsymbol{\mu}_i^*=n_i\boldsymbol{p}_i^*$, $\Delta^*(\boldsymbol{p}_i)$ は $\Delta(\boldsymbol{p}_i)$ から第1行，第1列を除いた行列とし，

$$\begin{aligned}\sigma_i=E(Q_i)&=2n_i\{1+(n_i-1)\phi_i\}\\ \text{or}\quad &=2n_i\{1+(n_i-1)\phi(\boldsymbol{x}_i)\}\qquad(\phi(\boldsymbol{x}_i)=\phi(\boldsymbol{x}_i^T\boldsymbol{\gamma}))\end{aligned}$$

$$D_{2i}=\partial\sigma_i/\partial\boldsymbol{\gamma},\qquad V_{2i}=\sigma_i^2$$

とする．上記の推定方程式の解法については，まず $\boldsymbol{\gamma}$ 所与のもとで $\boldsymbol{\beta}$ について(5.1)式を解く．つぎに，ここで得られた推定値 $\hat{\boldsymbol{\beta}}$ を(5.2)式の $\boldsymbol{\beta}$ とし $\boldsymbol{\gamma}$ について(5.2)式を解く．さらに，そうして得られた推定値 $\hat{\boldsymbol{\gamma}}$ を(5.1)式の $\boldsymbol{\gamma}$ として，また $\boldsymbol{\beta}$ について(5.1)式を解く．この操作を $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ が収束するまで繰り返すことによって解を得ることができる．また各方程式の数値解法にはNewton-Raphson 法を用いることができる．得られた解については漸近正規性が成立し，例えば $\hat{\boldsymbol{\beta}}$ については

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\to N(\boldsymbol{0},\ V_\beta)$$

$$V_\beta=\lim_{n\to\infty}\left(\frac{1}{n}\sum D_{1i}^T V_{1i}^{-1}D_{1i}\right)^{-1}\left\{\frac{1}{n}\sum D_{1i}^T V_{1i}^{-1}\mathrm{Var}(\boldsymbol{Y}_i^*)V_{1i}^{-1}D_{1i}\right\}\left(\frac{1}{n}\sum D_{1i}^T V_{1i}^{-1}D_{1i}\right)^{-1}$$

が成り立ち，$\hat{\boldsymbol{\gamma}}$ についても同様の式が成立する（Liang and Zeger (1986), Zhu et al. (1994)）．

ここで，$\mathrm{Var}(\boldsymbol{Y}_i^*)$ は反応変数 $\boldsymbol{Y}_i^*$ の真の分散共分散であり，$V_{1i}$ は反応変数 $\boldsymbol{Y}_i^*$ がディリクレ-多項分布に従うと想定した場合の分散共分散であることに注意が必要である．したがって，実際にデータに関してディリクレ-多項分布の分散共分散構造を想定してよい場合には $V_\beta$ が

$$V_\beta=\lim_{n\to\infty}\left(\frac{1}{n}\sum D_{1i}^T V_{1i}^{-1}D_{1i}\right)^{-1}$$

となり，最尤推定量の分散共分散行列を計算する際に用いる式と同様の計算式の推定量が対応することが分かる．パラメータの分散共分散 $(1/n)V_\beta$ の推定値については，得られたパラメータを用いて上記の式を評価することになるが，ディリクレ-多項分布を想定でき簡略化した分散共分散行列を用いる場合には，その推定量（ナイーブ推定量）

$$(\sum\hat{D}_{1i}^T\hat{V}_{1i}^{-1}\hat{D}_{1i})^{-1}$$

を，真の分布がディリクレ-多項分布の想定からずれた場合を考えなければならない場合には，$\mathrm{Var}(\boldsymbol{Y}_i^*)$ の部分について標本残差 $\boldsymbol{S}_i=\boldsymbol{y}_i^*-\hat{\boldsymbol{\mu}}_i^*$ を用いて

$$(\sum\hat{D}_{1i}^T\hat{V}_{1i}^{-1}\hat{D}_{1i})^{-1}(\sum\hat{D}_{1i}^T\hat{V}_{1i}^{-1}\boldsymbol{S}_i\boldsymbol{S}_i^T\hat{V}_{1i}^{-1}\hat{D}_{1i})(\sum\hat{D}_{1i}^T\hat{V}_{1i}^{-1}\hat{D}_{1i})^{-1}$$

の様に計算するサンドイッチ推定量を用いることにする．

## 6. 解析例：Chen データの分析

ここでは，前節までに述べたディリクレ-多項分布を用いた最尤法にもとづく分析や，一般化推定方程式を用いたデータ解析の例として1節で述べた Chen のデータの分析について紹介す

る．

表1で示したように，ここでは試験動物を3群に分けて，低用量，中用量，高用量のオキシ尿素を投与し，反応としては胎内の着床部位について，(I)正常発育，(II)奇形，(III)死亡-吸収，の状態を検査している．

そこで，共変量として用量効果のトレンドを調べるための回帰型の変数

$$x_{trend}=\begin{cases}0 & \text{Low Dose}\\ 1 & \text{Medium Dose}\\ 2 & \text{High Dose}\end{cases}$$

と，用量群を示すダミー変数

$$x_{mid\text{-}dose}=\begin{cases}1 & \text{Medium Dose}\\ 0 & \text{else}\end{cases}\qquad x_{high\text{-}dose}=\begin{cases}1 & \text{High Dose}\\ 0 & \text{else}\end{cases}$$

を用いることにした．

さらに，反応に関わる関連性の尺度として，反応を順序のない名義尺度反応とみなすときには，正常発育カテゴリをベースラインカテゴリとする多項ロジット(2.1)式を，また順序のある反応とみなす場合には，累積ロジット(2.2)式を採用しこれらのロジットに関して

$$\text{logit}_1(\boldsymbol{x})=\beta_{0(1)}+\beta_{1(1)}x_1+\cdots+\beta_{q(1)}x_q$$
$$\text{logit}_2(\boldsymbol{x})=\beta_{0(2)}+\beta_{1(2)}x_1+\cdots+\beta_{q(2)}x_q$$

となるように線形予測子を対応づけて考えることにした．

また，超過変動に関わるパラメータ $\phi$ については

$$-\log\Theta=\log\frac{\phi}{1-\phi}=\gamma_0+\gamma_1x_1+\cdots+\gamma_qx_q$$

なるモデル化をここでは考える．ただし $x_1,\cdots,x_q$ は適当に選択された共変量を示す．例えば，ここで用量群を示すダミー変数を用いる場合には，$x_1=x_{mid\text{-}dose}$, $x_2=x_{high\text{-}dose}$ として2つの共変量を用いることになり，回帰型の共変量を用いて分析する場合には，$x_1=x_{trend}$ として分析を行うことになる．

表2は反応に関して名義尺度を想定した場合の飽和モデル（つまり，この場合には，各用量群で個別の反応確率ベクトルを想定することに相当するモデル）の適合を示したものである．このうち，多項分布を用いた分析では，ピアソン $\chi^2$ の値が467.48(d.f.174)と非常に大きくモデル適合が悪いことが分かる．ちなみにこのモデルにおける $G^2$ の値も480.57（d.f.174）と大きな値を示すことが分かる．

このように大きな値を示す場合には，いくつかのはずれ値がデータ適合に大きな影響を与えている可能性が考えられる．ところが以下の影響診断に関するインデックスプロットに示されるように，パラメータの挙動に大きな影響を与える特定の観測単位はあまり見受けられない．

図1のCookの距離はパラメータ全体への各観測の影響を示し，図2は第2ロジット(正常発育と死亡-吸収に関するロジット）に関する高用量群と低用量群との比較のためのパラメータ(表2のHigh-Dose(2)に相当）の影響度インデックスプロットである．図の縦軸のスケールに注意すると，これらのばらつきが実質的に大きなものでないことを読み取ることができる．この傾向は他のパラメータについても同様であった．一方，ピアソン残差（図3）は各観測単位，つまり各母体，の反応に対するモデルの適合の程度を示しているが，全体的に大きな値となっていることが分かる．

以上のことから，このデータに関しては超過変動を考慮にいれた分析法を適用することが必要なことが分かる．表2の2，3列目に先に述べた多項分布の構造に加えて，3用量群に共通

表2. 名義尺度の場合の飽和モデル適合.（$\chi^2$ 値欄の括弧内は超過変動未調整の値，パラメータ推定値の隣りの括弧は推定値の標準偏差，GEE の場合にはサンドイッチ推定量を用いた．また，Intercept(1) は $\text{logit}_1$ における切片項に対応するパラメータ，Mid-Dose(1) は $\text{logit}_1$ における $x_{mid\text{-}dose}$ に対応するパラメータ，High-Dose(1) は $\text{logit}_1$ における $x_{high\text{-}dose}$ に対応するパラメータ，Intercept(2) は $\text{logit}_2$ における切片項に対応するパラメータ，Mid-Dose(2) は $\text{logit}_2$ における $x_{mid\text{-}dose}$ に対応するパラメータ，High-Dose(2) は $\text{logit}_2$ における $x_{high\text{-}dose}$ に対応するパラメータを示すものとする.）

| | Multinomial (MLE)(model1) | Dirichlet-Multinomial (MLE) | Dirichlet-Multinomial (GEE) |
|---|---|---|---|
| Log-likelihood | -827.96 | -764.67 | |
| $\chi^2$ | 467.48 | 176.89(461.41) | 181.07(462.50) |
| D.F. | 174 | 173 | 173 |
| Over-dispersion:$\gamma$ | | -1.637(0.181) | -1.670(0.191) |
| ($\phi$) | | 0.163(0.025) | 0.159(0.026) |
| Intercept(1) | -2.148(0.220) | -2.285(0.357) | -2.145(0.332) |
| Intercept(2) | -2.025(0.209) | -1.801(0.278) | -1.973(0.295) |
| Mid-Dose(1) | 0.905(0.308) | 1.223(0.476) | 0.955(0.506) |
| Mid-Dose(2) | 2.151(0.251) | 1.966(0.365) | 2.107(0.422) |
| High-Dose(1) | 1.345(0.273) | 1.692(0.418) | 1.330(0.389) |
| High-Dose(2) | 2.984(0.234) | 2.732(0.327) | 2.903(0.352) |

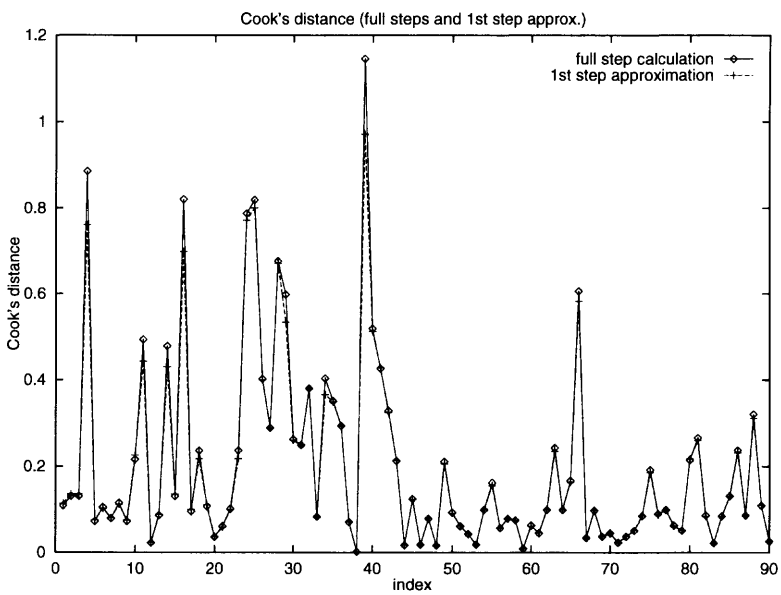


図1. 多項ロジットモデル (model 1) における Cook の距離に関するインデックスプロット（実線は Cook の距離，波線はその 1-step 近似（越智 他 (1994))).

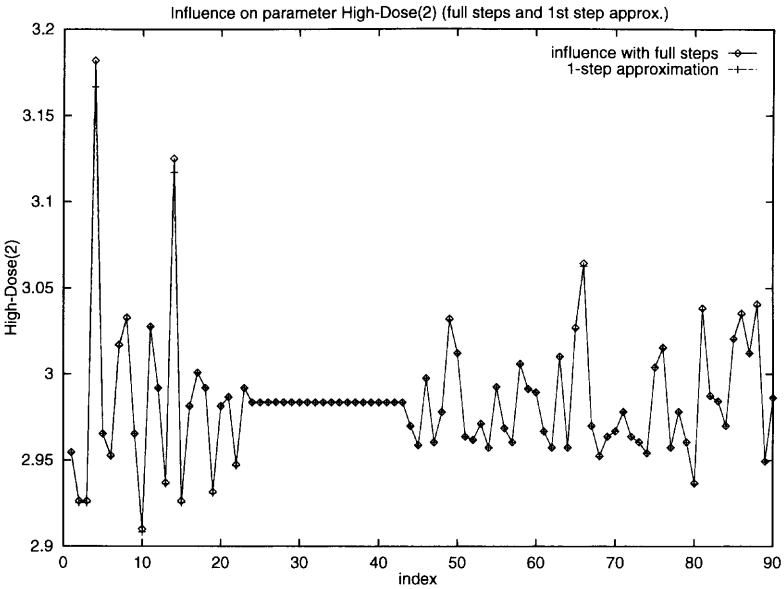


図2. 多項ロジットモデル (model 1) におけるパラメータ High-Dose(2)への影響インデックスプロット (High-Dose(2)は高用量の第2ロジットに関する効果を示すパラメータ，実線はそのインデックスプロット，波線はその1-step近似 (越智 他 (1994))).

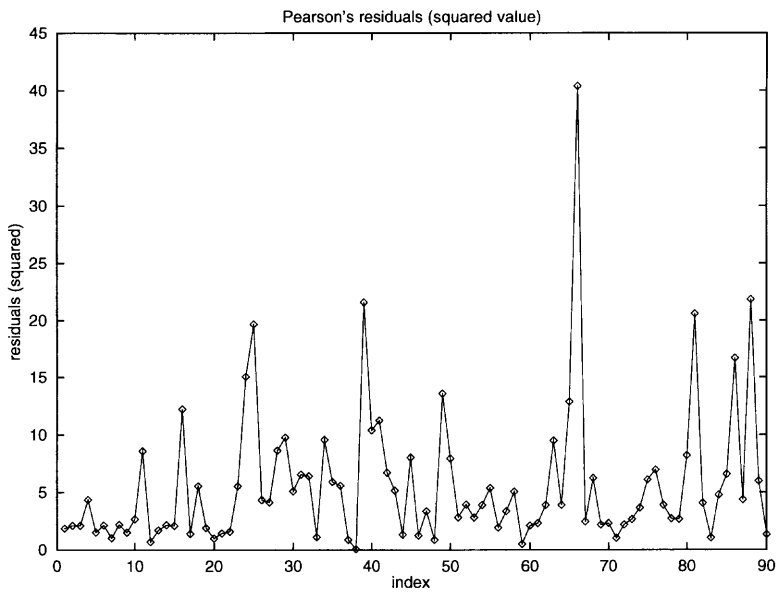


図3. 多項ロジットモデル (model 1) におけるピアソン残差の平方値 (ピアソン適合度 $\chi^2$ への寄与).

の超過変動パラメータを導入した場合のディリクレ-多項分布に関する最尤推定値とディリクレ-多項分布の平均-分散構造を利用した一般化推定方程式による推定値を示している．このうち，一般化推定方程式のパラメータ推定値に関する標準誤差はサンドイッチ推定量に基づいて

計算された値である．

最尤法における比較の立場からは対数尤度は多項分布の場合の－827.96（カーネル部）からディリクレ-多項の場合－764.67 と大幅に減少している．また，増加したパラメータは 1 個であるので自由度の減少は 1 である．

ディリクレ-多項分布における適合度，ピアソンの $\chi^2$ を

$$\chi^2=\sum_{i=0}^{N}(\boldsymbol{y}_i^*-n_i\hat{\boldsymbol{p}}_i^*)^T[n_i\Delta^*(\hat{\boldsymbol{p}}_i)\{1+(n_i-1)\hat{\phi}_i\}]^{-1}(\boldsymbol{y}_i^*-n_i\hat{\boldsymbol{p}}_i^*)$$

とすると，この値は多項分布の場合の適合度 467.48 (d.f. 174) に対して，ディリクレ-多項分布の場合 176.89 (d.f. 173)，一般化推定方程式によるモデル適合では 181.07 (d.f. 173) と適合度の面で著しく改善できていることが分かる．表 2 中の $\chi^2$ 値の欄の括弧内の数字は超過変動パラメータの調整を行わない場合の $\chi^2$ 値を示している．基本的に多項分布での適合自身が悪かったために，ディリクレ-多項分布あるいはその分散-共分散構造を利用したモデルではその値も多少減少しているが，超過変動パラメータの調整によるほど劇的には変化していないことも分かる．この分析では，ディリクレ-多項分布における最尤法と一般化推定方程式の結果はかなり相似した結果となっている．ちなみにこの場合の平均構造にかかわるパラメータの分散共分散行列としてのナイーブ推定量とサンドイッチ推定量は表 3，表 4 の通りであり，両者が大きくは相違していない様子がわかる．したがってこのデータに関しては，ディリクレ-多項分布を想定

表 3．平均パラメータ推定量に関するナイーブ分散共分散推定量（多項ロジット）．

| Intercept(1) | Intercept(2) | Mid-Dose(1) | Mid-Dose(2) | High-Dose(1) | High-Dose(2) |
|---|---|---|---|---|---|
| 0.107 | -0.107 | -0.107 | 0.013 | -0.013 | -0.013 |
| | 0.160 | 0.107 | -0.013 | 0.041 | 0.013 |
| | | 0.135 | -0.013 | 0.013 | 0.033 |
| | | | 0.125 | -0.125 | -0.125 |
| | | | | 0.246 | 0.125 |
| | | | | | 0.190 |

表 4．平均パラメータ推定量に関するサンドイッチ分散共分散推定量（多項ロジット）．

| Intercept(1) | Intercept(2) | Mid-Dose(1) | Mid-Dose(2) | High-Dose(1) | High-Dose(2) |
|---|---|---|---|---|---|
| 0.087 | -0.087 | -0.087 | 0.015 | -0.015 | -0.015 |
| | 0.178 | 0.087 | -0.015 | 0.074 | 0.015 |
| | | 0.124 | -0.015 | 0.015 | 0.038 |
| | | | 0.110 | -0.110 | -0.110 |
| | | | | 0.256 | 0.110 |
| | | | | | 0.151 |

表 5．超過変動パラメータ $\phi$ の推定値．

| | Low | Medium | High |
|---|---|---|---|
| 最尤推定値 | 0.132 | 0.230 | 0.139 |
| (S. E.) | (0.0521) | (0.0545) | (0.0312) |
| GEE | 0.116 | 0.251 | 0.137 |
| (S. E.) | (0.0512) | (0.0534) | (0.0331) |

表6. ディリクレ-多項分布にもとづく回帰パラメータの最尤推定値.（$\chi^2$値はピアソンの$\chi^2$値を示す．また，Intercept(1)は$\text{logit}_1$における切片項に対応するパラメータ，Trend(1)は$\text{logit}_1$における$x_{trend}$に対応するパラメータ，Intercept(2)は$\text{logit}_2$における切片項に対応するパラメータ，Trend(2)は$\text{logit}_2$における$x_{trend}$に対応するパラメータを示すものとする．）

| パラメータ | 最尤推定値 | S.E. |
|---|---|---|
| Over-dispersion:$\gamma$ | -1.594 | 0.1785 |
| ($\phi$) | 0.169 | 0.0251 |
| Intercept(1) | -2.169 | 0.314 |
| Intercept(2) | -1.543 | 0.243 |
| Trend(1) | 0.837 | 0.202 |
| Trend(2) | 1.287 | 0.157 |
| Log-likelihood=-766.939, $\chi^2_{(175)}=176.94$ | | |

表7. ディリクレ-多項分布にもとづく回帰パラメータのGEE推定値.（ただし，$\text{S.E.}_s$はサンドイッチ推定量による標準偏差，$\text{S.E.}_n$はナイーブ推定量による標準偏差，$\chi^2$値はピアソンの$\chi^2$値を示す．また，Intercept(1)は$\text{logit}_1$における切片項に対応するパラメータ，Trend(1)は$\text{logit}_1$における$x_{trend}$に対応するパラメータ，Intercept(2)は$\text{logit}_2$における切片項に対応するパラメータ，Trend(2)は$\text{logit}_2$における$x_{trend}$に対応するパラメータを示すものとする．）

| パラメータ | 推定値 | $\text{S.E.}_s$ | $\text{S.E.}_n$ |
|---|---|---|---|
| Over-dispersion:$\gamma$ | -1.622 | 0.192 | 0.204 |
| ($\phi$) | 0.165 | 0.0264 | 0.0281 |
| Intercept(1) | -2.083 | 0.303 | 0.321 |
| Intercept(2) | -1.608 | 0.244 | 0.251 |
| Trend(1) | 0.677 | 0.191 | 0.218 |
| Trend(2) | 1.325 | 0.162 | 0.161 |
| $\chi^2_{(175)}=180.58$ | | | |

した分析でも適切な分析が可能であるといえる．

上記表2では超過変動パラメータ$\phi$について3群共通のパラメータを想定したが，3群個別のパラメータをもつようにして分析した場合，それぞれの群における超過変動パラメータは表5のようになる．ただし，実際にはこれら3群の間の差は有意ではない：尤度法（尤度比検定）$\chi^2(2)=2.638$ ($p=0.267$)，GEE（コントラストを用いたワールドタイプの検定）$\chi^2(2)=4.564$ ($p=0.102$)．したがって，個々の群での超過変動パラメータについては表2の場合のように共通パラメータを用いたモデルでよいと考えられる．この場合のパラメータ$\phi$の推定値は尤度法による場合0.163 (S.E. 0.025)，GEEの場合0.159 (S.E. 0.026) となる．ちなみに，この超過変動パラメータ$\phi$は4節で述べたように着床部位間の相関を表わすパラメータとしてもとらえることができる．

表2の各ロジットにおける共変量効果を調べてみると用量が増加するにつれてその効果も増大していることが分かる．したがって，共変量に$x_{trend}$を用いてモデル適合を行うと表6，表7

のようになり，いずれの反応カテゴリに対するロジットについても強い関連があることが分かり，投与量が増加すれば奇形や死亡が起こり易くなることが確かめられた．超過変動パラメータ $\phi$ については，0.169（S.E. 0.0251, MLE），あるいは 0.165（S.E. 0.0264, GEE）なる推定値が得られた．

ディリクレ-多項分布としてのモデル適合の上では表 2 の結果とこの線形モデルとの違いは有意ではない（尤度比検定 $\chi^2(2)=4.529$ ($p=0.104$)）．また GEE としての比較においても表 2 の結果をもとにした線形性に関する検定（$H_0$：2Mid-Dose(1)－High-Dose(1)＝0, 2Mid-Dose(2)－High-Dose(2)＝0）の結果は有意ではなかった（$\chi^2(2)=3.578$, ($p=0.167$)）．ただし，これ以上のモデルの簡略化は難しく，例えば各ロジットにおける用量効果パラメータが等しいと言う仮説は，尤度比検定でも GEE のコントラストを用いたワールドタイプ検定でも棄却された（それぞれ $\chi^2$ 値は 4.390, 10.630（d.f. 1）であった）．

このデータは 3 カテゴリを持ち，カテゴリ間に順序性を仮定できる．したがって，順序を考慮にいれた分析を行うことによってより簡略化を行ったモデルの適合が行え，データの持つ情報のより簡潔な表現が可能になることが期待できる．そこで，ここでは累積ロジットを用いたデータの分析を行った．表 8 は表 2 と同様に各用量群に個別の反応確率ベクトルを想定して 飽和モデルの適合を行った結果である．この場合，平均構造にかかわる部分については基本的に表 2 と表 8 は同等なモデルの適合を行っていることになる．そのため表 8 の対数尤度や適合度，また超過変動パラメータは表 2 のそれと同一である．また，超過変動に関する議論は先の名義尺度の場合の議論と全く同一となり，共通パラメータで十分であると判断できる．

このモデル適合から，名義尺度の場合とは異なり，累積ロジスティックモデルでは各カテゴ

表 8. 累積ロジスティックモデルの飽和モデル適合．（$\chi^2$ 値欄の括弧内は超過変動未調整の値，パラメータ推定値の隣りの括弧は推定値の標準偏差，GEE の場合にはサンドイッチ推定量を用いた．また，Intercept(1) は $\text{logit}_1$ における切片項に対応するパラメータ，Mid-Dose(1) は $\text{logit}_1$ における $x_{mid\text{-}dose}$ に対応するパラメータ，High-Dose(1) は $\text{logit}_1$ における $x_{high\text{-}dose}$ に対応するパラメータ，Intercept(2) は $\text{logit}_2$ における切片項に対応するパラメータ，Mid-Dose(2) は $\text{logit}_2$ における $x_{mid\text{-}dose}$ に対応するパラメータ，High-Dose(2) は $\text{logit}_2$ における $x_{high\text{-}dose}$ に対応するパラメータを示すものとする．）

| | Multinomial (MLE) | Dirichlet-Multinomial (MLE) | Dirichlet-Multinomial (GEE) |
|---|---|---|---|
| Log-likelihood | -827.96 | -764.67 | |
| $\chi^2$ | 467.48 | 176.89(461.41) | 181.07(462.50) |
| D.F. | 174 | 173 | 173 |
| Over-dispersion:$\gamma$ | | -1.637(0.181) | -1.670(0.191) |
| ($\phi$) | | 0.1629(0.0247) | 0.159(0.026) |
| Intercept(1) | -1.391(0.1596) | -1.321(0.235) | -1.362(0.237) |
| Intercept(2) | -2.136(0.2074) | -1.898(0.275) | -2.084(0.292) |
| Mid-Dose(1) | 1.744(0.2075) | 1.743(0.322) | 1.732(0.373) |
| Mid-Dose(2) | 2.008(0.2451) | 1.766(0.352) | 1.952(0.395) |
| High-Dose(1) | 2.508(0.1900) | 2.448(0.286) | 2.453(0.299) |
| High-Dose(2) | 2.724(0.2272) | 2.388(0.315) | 2.647(0.335) |

リ区分ごとのロジットに対して個別の用量効果を考える必要が無いことがうかがえる．実際，カテゴリ間での用量効果の共通化に関する仮説検定（$H_0$：Mid-Dose(1)＝Mid-Dose(2)，High-Dose(1)＝High-Dose(2)）については，尤度比検定では $\chi^2(2)=0.244$（$p=0.885$），GEE におけるコントラストを用いた検定では $\chi^2(2)=0.671$（$p=0.715$）であった．

そこで用量効果についてカテゴリ間で共通化した分析の結果を表9に示す．この分析では，ディリクレ-多項分布にもとづく分析でもそれを拡張した GEE における分析でも似た結果が得られている．このときの平均パラメータに関するナイーブ分散共分散推定量とサンドイッチ分散共分散推定量は表10，表11の通りであり，これら2つの推定量が非常に近いとは言いがたいが，非常に大きく異なっているとも言えないように見える．両者の客観的な近さの評価については今後の検討課題である．

また，用量効果に関する線形モデルの適合にはこの場合注意が必要である．ダミー変数の代りに線形効果を示す共変量 $x_{trend}$ を用いた分析では，カテゴリ区分別の用量効果の違いを考慮する必要はないものの，そのモデルにおける適合はダミー変数を用いたモデルに比べて有意に劣ることが示される．例えば，カテゴリ区分ごとに定義されるロジットについて共通の線形効果を考える場合，ディリクレ-多項分布にもとづくモデルの適合に関する最大尤度の対数は－767.29（カーネル部のみ）であり，ダミー変数を用いたモデル（表9の model 2）との差違に

表9．累積ロジスティックモデルによる分析．（$\chi^2$ 値欄の括弧内は超過変動未調整の値，パラメータ推定値の隣りの括弧は推定値の標準偏差，GEE の場合にはサンドイッチ推定量を用いた．また，Intercept(1) は $\text{logit}_1$ における切片項に対応するパラメータ，Intercept(2) は $\text{logit}_2$ における切片項に対応するパラメータ，Mid-Dose は $\text{logit}_1$，$\text{logit}_2$ 共通に適用される $x_{mid\text{-}dose}$ のパラメータ，High-Dose は $\text{logit}_1$，$\text{logit}_2$ 共通に適用される $x_{high\text{-}dose}$ のパラメータを示すものとする．）

| | Multinomial (MLE) | Dirichlet-Multinomial (MLE)(model2) | Dirichlet-Multinomial (GEE) |
|---|---|---|---|
| Log-likelihood | -829.31 | -764.80 | |
| $\chi^2$ | 473.94 | 175.33(457.54) | 180.64(469.08) |
| D.F. | 176 | 175 | 175 |
| Over-dispersion $\gamma$ | | -1.637(0.181) | -1.644(0.184) |
| $\phi$ | | 0.163(0.025) | 0.162(0.025) |
| Intercept(1) | -1.416(0.160) | -1.315(0.232) | -1.384(0.234) |
| Intercept(2) | -1.962(0.166) | -1.920(0.237) | -1.934(0.236) |
| Mid-Dose | 1.804(0.203) | 1.765(0.310) | 1.780(0.356) |
| High-Dose | 2.546(0.186) | 2.418(0.277) | 2.491(0.289) |

表10．平均パラメータ推定量に関するナイーブ分散共分散推定量（累積ロジット）．

| Intercept(1) | Intercept(2) | Mid-Dose | High-Dose |
|---|---|---|---|
| 0.065 | 0.064 | -0.065 | -0.064 |
| | 0.070 | -0.067 | -0.068 |
| | | 0.108 | 0.067 |
| | | | 0.088 |

表 11. 平均パラメータ推定量に関するサンドイッチ分散共分散推定量（累積ロジット）.

| Intercept(1) | Intercept(2) | Mid-Dose | High-Dose |
|---|---|---|---|
| 0.055 | 0.053 | -0.054 | -0.053 |
| | 0.056 | -0.056 | -0.056 |
| | | 0.127 | 0.056 |
| | | | 0.084 |

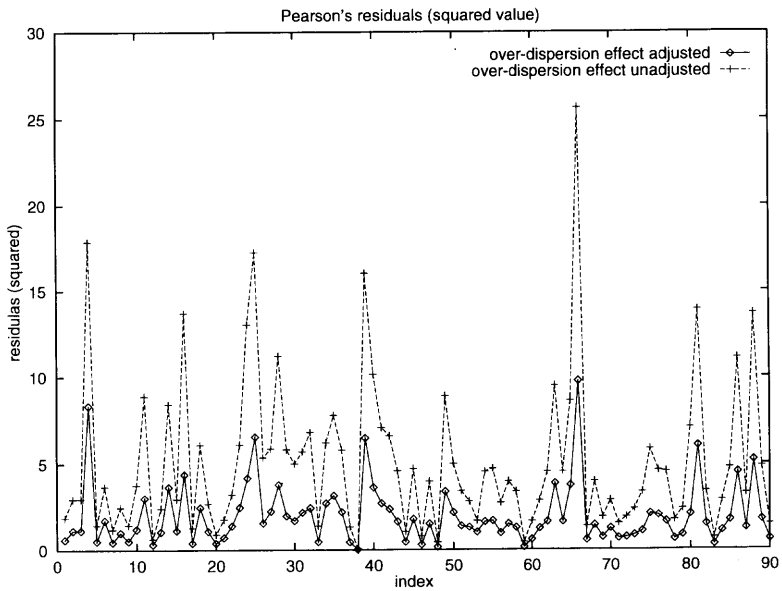


図 4. 累積ロジットモデル（model 2）におけるピアソン残差の平方値（ピアソン適合度 $\chi^2$ への寄与）. ただし，実線は超過変動調整後の，波線は超過変動調整前の各観測の寄与.

関する尤度比検定統計量の値は $\chi^2(1)=4.994$ ($p=0.025$) であった. また，表 8 の GEE の分析結果では，得られたパラメータの線形性に関する検定（$H_0$ : 2Mid-Dose(1)−High-Dose(1)=0, 2Mid-Dose(2)−High-Dose(2)=0）について $\chi^2(2)=5.159$ ($p=0.023$) となり線形性が棄却される. ところが，表 9 の GEE の分析での線形性に関する検定（$H_0$ : 2 Mid-Dose−High-Dose=0）では $\chi^2(1)=3.105$ ($p=0.078$) となり棄却できない. これは GEE での分析の方が，最尤法の結果より多少パラメータ推定値の標準偏差を大きく推定しているために生じた結果であると考えられる.

以上のデータ解析の結果得られた超過変動調整後のモデルにおける適合度はいずれも大きくモデルからのずれを示すものではなかった. 例えば，ディリクレ-多項分布の場合の累積ロジットモデル（表 9 の model 2）のピアソン適合度 $\chi^2$ は 175.33（d.f. 175）であるが，各観測のこの適合度に関する寄与は図 4 の通りである. 図 4 の実線は超過変動調整後の,波線は超過変動調整前の各観測の寄与を示しているが，調整後相当に適合の改善ができていることが分かる. また，推定されたパラメータに関する各観測の影響として Cook の距離に関する影響を調べると図 5 のようになる. 基本的にあまり大きな影響を与える観測はないと考えられるが，25 番目，66 番目のデータには気をつけておかなければならないかもしれない. ちなみに，これらのデー

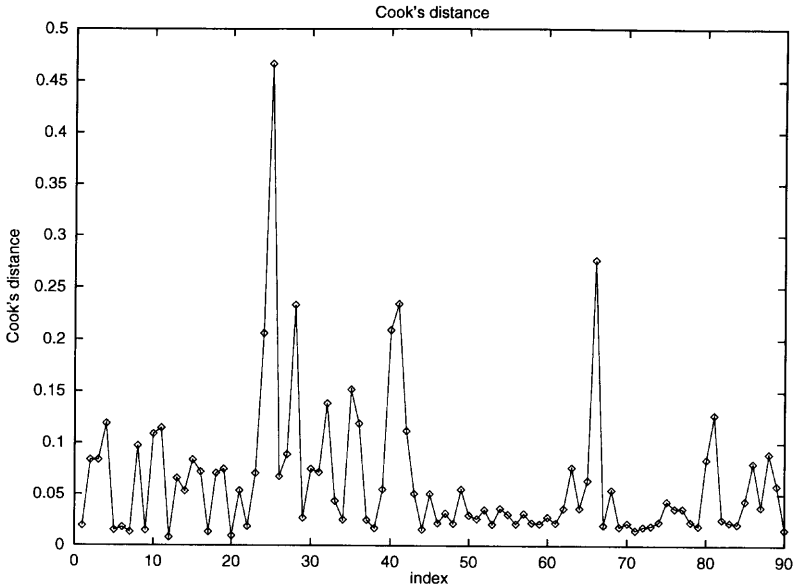


図5. 累積ロジットモデル (model 2) における Cook の距離.

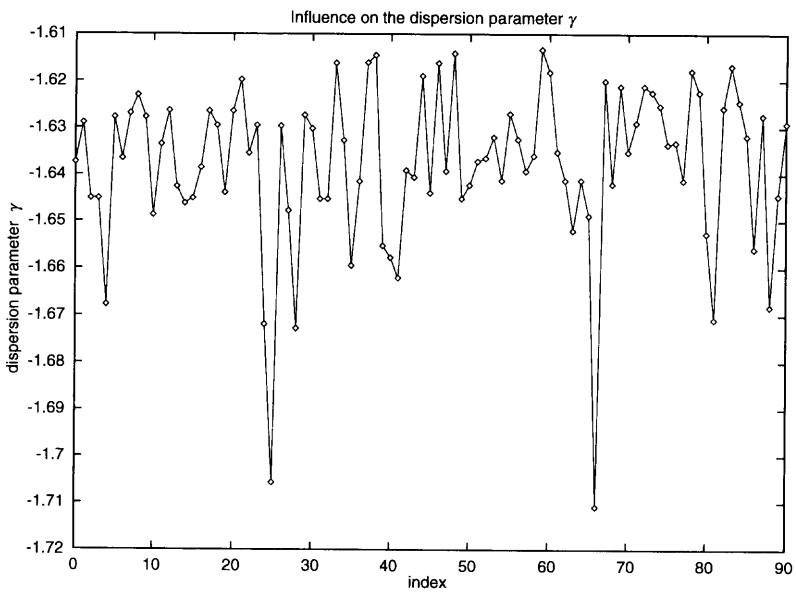


図6. 累積ロジットモデル (model 2) における超過変動パラメータに関する影響.

タは超過変動に比較的大きな影響を与えるデータであり，図6からも分かるように超過変動を増大させる効果をもっている．実際，表1でもこれらのデータは群としての全体的な傾向とは異なった反応を持つデータであり注意が必要なことが確認できる．ただ，これらのデータを除くことによるデータ解析全般への影響は超過変動自身の評価を除き軽微なものであった．

## 7. まとめ

以上超過変動をもつカテゴリカルデータの分析について，ディリクレ-多項分布をもとにしたモデル適合の観点から分析法とそのデータ解析の一例を述べてきた．

超過変動の処理に関して重要なポイントはパラメータの分散推定の問題である．つまり，超過変動をもつデータに対して，それを無視してデータの分析を行うと分散の過小評価につながり，ひいては共変量効果の有意性を過大評価してしまう点にある．今回の分析例においても，多項分布にもとづく分析のパラメータの分散推定値よりディリクレ-多項分布を用いたデータ解析のそれの方が大きく推定されている．ここでのデータに関しては，対象とする共変量の効果が強いものであったため，超過変動の有無に関わらずその効果の有意性に変化は無かったが，微妙な効果の予想されるデータの場合には注意が必要である．

また，平均に関するパラメータ推定値自身に関しては，その分散推定値ほど大きな変化はなかった．ただ，総じて最尤推定量の方がGEE推定量よりも共変量効果を低く見積もり，超過変動を考慮に入れない多項分布にもとづく推定量が最もラディカルに共変量効果を推定しているようであった．ベータ-二項分布の場合，超過変動推定値を適切に処理しなければ最尤法ではパラメータ推定に関してバイアスが生じることが知られている (Kupper et al. (1986), Yamamoto and Yanagimoto (1988), Liang and Hanfelt (1994))．このデータ解析の場合には，超過変動に関して中用量群でその差は有意ではなかったものの大きな推定値を生じていたので，ここで適用したモデルでもこのバイアスの問題が生じている懸念がある．実際3群別の超過変動を考えた分析では，最尤推定量における用量効果について，共通な超過変動を想定した値よりわずかながら大きく見積もる傾向が見られた．これはGEEでの分析でも同様な傾向であったが，その程度は最尤法の場合に比べて小さなものであった．これらのバイアスや推定効率の問題については今後より詳細な研究が必要である．

また，今回は超過変動に関するモデル化として，基本的にディリクレ-多項分布を用いた分析について考えたが，これ以外の平均-分散構造をもとにしたデータ分析について考えることも重要な問題である．一般化推定方程式では，想定した作業分散共分散構造が真の分散構造でなくても柔軟な分析が可能であることから，今回想定したディリクレ-多項分布の平均-分散構造による分析法である程度広範な分析が行えると考えられるが，その適用可能性に関する検討も今後の課題である．

## 謝辞

丁寧な査読を賜り，貴重なご意見とご指摘を頂きました査読者に感謝いたします．
本報告は一部統計数理研究所共同研究9-共研A-63の補助を受けた．

## 参考文献

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, Wiley, New York.

Agresti, A..(1990). *Categorical Data Analysis*, Wiley, New York.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution, *Applied Statistics*, **27**, 162-167.

Chen, J. J. and Li, L.-A. (1992). An evaluation of beta-binomial and Dirichlet-trinomial models for the analysis of reproductive and developmental effects, *Biometrical J.*, **34**, 231-241.

Chen, J. J., Kodell, R. L., Howe, R. B. and Gaylor, D. W. (1991). Analysis of trinomial responses from reproductive and developmental toxicity experiments, *Biometrics*, **47**, 1049-1058.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Halls, London.

Haseman, J. K. and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments, *Biometrics*, **35**, 281-293.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present, *Applied Statistics*, **39**, 75-84.

Jørgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models, *Biometrika*, **70**, 19-28.

Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics*, **34**, 69-76.

Kupper, L. L., Portier, C., Hogan, M. D. and Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology, *Biometrics*, **42**, 85-98.

Lesaffre, E. and Albert, A. (1989). Multiple-group logistic regression diagnostics, *Applied Statistics*, **38**, 425-440.

Liang, K.-Y. and Hanfelt, J. (1994). On the use of the quasi-likelihood method in teratological experiments, *Biometrics*, **50**, 872-880.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.

McCullagh, P. (1983). Quasi-likelihood functions, *Ann. Statist.*, **11**, 59-67.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Halls, London.

Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions, *Biometrika*, **73**, 583-588.

Moore, D. F. (1987). Modelling the extraneous variance in the presence of extra-binomial variation, *Ann. Statist.*, **36**, 8-14.

Morgan, B. J. T. (1992). *Analysis of Quantal Response Data*, Chapman and Halls, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *J. Roy. Statist. Soc. Ser. A*, **135**, 370-384.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model, *Biometrika*, **71**, 531-543.

越智義道，杉山陽一，牧野智明(1991)．多項分布にもとづく一般化線形モデルと関連性評価，大分大学工学部研究報告，**24**，49-56.

越智義道，岸野清広，上杉博紀，小畑経史 (1994)．多項反応における回帰分析と診断法，計算機統計学，**7**，111-125.

O'Hara Hines, R. J., Lawless, J. F. and Carter, E. M. (1992). Diagnostics for a cumulative multinomial generalized linear model, with applications to grouped toxicological mortality data, *J. Amer. Statist. Assoc.*, **87**, 1059-1069.

Paul, S. R. (1987). On the beta-correlated binomial (BCB) distribution : a three parameter generalization of the binomial distribution, *Comm. Statist. Theory Methods*, **16**, 1473-1478.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *J. Amer. Statist. Assoc.*, **81**, 321-327.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics*, **47**, 825-839.

Ryan, L. (1992). Quantitative risk assessment for developmental toxicity, *Biometrics*, **48**, 163-174.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method, *Biometrika*, **61**, 439-447.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics*, **31**, 144-148.

Yamamoto, E. and Yanagimoto, T. (1988). Litter effects to dose-response curve estimation, *J. Japan Statist. Soc.*, **18**, 97-106.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, **42**, 121-130.

Zhu, Y., Krewski, D. and Ross, W. H. (1994). Dose-response models for correlated multinomial data from developmental toxicity studies, *Applied Statistics*, **43**, 583-598.

# Analysis of Categorical Data with Extra-multinomial Variation

Yoshimichi Ochi

(Department of Computer Science and Intelligent Systems, Faculty of Engineering, Oita University)

In this paper, we consider regression analysis of categorical data with extra-multinomial distribution. The response categories may have some ordinal nature.

An extension of the multinomial distribution, dirichlet-multinomial distribution is considered as an approach to model the logistic regression with and/or without ordinal nature. An alternative approach via GEE is also investigated, using the mean-variance structure of the dirichlet-multinomial distribution, and the technical details of fitting these two methods are described.

Furthermore, some comparisons are made on the applicability of the two methods for an actual data set.

Key words: Over-dispersion, multinomial distribution, logistic regression, ordered response, GEE.