メキシコシティの教育機会分析への「二項分布の和モデル」の適用

——周辺セルデータ分布からの内部セル確率の推計——

米村明夫／野田容助

はじめに

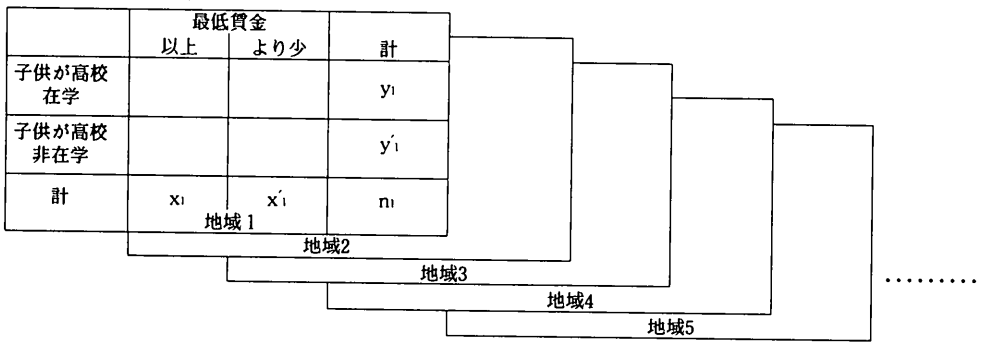
本稿は，「二項分布の和モデル」と呼ぶ確率論的因果モデルを用いて，メキシコシティに住む若年層の教育機会の格差が，家族の社会特性や住宅特性によって，どのような影響を受けるかという問題の数量的分析を行うものである。本稿は，実際に公刊されたデータをこのモデルに適用することによって，その有効性を示すと同時に，適用のしかたについても範例を示すことをねらいとしている。したがってここでは，適用結果の提示ばかりでなく，この方法の統計学的な基礎，その適用のしかたについて，詳しい議論が行われる(注1)。

このモデルに基づく分析の汎用性の高さは，具体例を挙げれば容易に理解されよう。

例えば，地域1において15～17歳人口のy_1人が高校への入学機会を持ち，y'_1人が持たないというデータがあるとする（$y_1+y'_1=n_1$とする）。また，これら15～17歳人口の家族の収入が法定最低賃金以上である家族の数をx_1，それより少ない家族の数をx'_1というデータがあるとする（$x_1+x'_1=n_1$とする）。ところが，「入学機会を持っているか否か」ということと「最低賃金以上の収入の家族か否か」ということのクロス集計データが存在しないとする。我々の目的は，このクロス集計データにあたるものを推計することにある。すなわち，2×2の分割表の周辺度数が与えられている時，その内部セルの頻度数（厳密には，内部セル確率）を推計することである。もちろん，データがこのような状態で地域1についてのみにしかないならば，そのような推計は不可能である。しかし，地域2，地域3，地域4，地域5，…，等の，多数の地域についても同様のデータがあるとする（図1参照）ならば，「二項分布の和モデル」によって，内部セル確率すなわち，最低賃金以上の収入の家族に属する者が高校に入学する確率と，最低賃金より少ない収入の家族に属する者が高校に入学する確率の推計が可能となるのである。

センサスなどの公刊されたデータでは，データが上記のような構造を持つこと(注2)，すなわち，多数の地域について各変数の（周辺）度数の情報を掲げるが，研究者が関心を持つ2つの変数間のクロス集計については情報を与えていないことがしばしば生ずる。したがって，上記

図1　2×2分割表データ



（出所）　筆者作成。

の方法が有効な場面は少なくないのである。

以下，第Ⅰ節では，モデルの設定，パラメーターの推定方法についての統計学的な議論を行い，第Ⅱ節では，メキシコの1990年センサス，メキシコシティ（連邦地区）データにそれを適用，結果を提示し，また，通常の回帰モデルによる結果との比較を行う。

Ⅰ　モデル

1．基本的な考え方

引き続き「はじめに」で述べた例を用いて，本稿で用いる方法についての基本的な考え方を述べておくことにしよう。

まず重要な前提として，教育機会は，家族の収入によって確率的に決定されるという確率論的因果関係を仮定する。すなわち，ある15～17歳の個人が最低賃金以上の収入の家族に属するならば高校に入学する確率はpであり，最低賃金より少ない収入の家族に属するならば高校に入学する確率はp'とする。p，p'は地域に依存せず，それぞれ一定とする。

すると，最低賃金以上家族の子弟数xのうち，高校へ入学する者の人数Zは確率変数であり，Zは，試行回数x回，確率pの二項分布にしたがう。同様に，最低賃金より低い家族の子弟数x'のうち，高校に入学する者の人数をZ'とすると，Z'は試行回数x'回，確率p'の二項分布に従う確率変数となる。

さらに，$Z+Z'=Y$とすると，Yは2つの二項分布の和として得られた確率変数となる。

実際に与えられたデータは，x_j，x'_jとy_j，y'_j（y_j，y'_jは確率変数Y_j，Y'_jの実現値。$j=1$，2，3，4，5，……で，各地域に対応）であり，我々が求めているのは，データと最も適合的なY，Y'の分布を与えるp，p'である。適合性の検討には，後に説明するK-L情報量と呼ばれる統計量の最小化という基準を用いる。

以下，この節の残りで，この2つの二項分布の組み合わせに基づくモデル，すなわち「二項分布の和モデル」の数式化，K-L情報量による適合性の決定方式の数式化を行うが，その前に，モデル設定の前提となっている重要な点を確認しておく必要がある。

第1に，ここでは，2×2の分割表の縦と横は対称に扱われてはいない。xとx'は確率変数

ではなく，YとY'は確率変数である。このような扱いには，「教育機会は家族の収入によって，確率論的に決定される」という社会科学的な因果関係の想定が反映されているのである(注3)。

第2に，上記のような構造のデータは，個人についてのデータを地域ごとに集計したものであり，また，推計によって得られるセル確率は，個人がそのセルに属する確率となっている。したがって，この方法による推計は集計データから個別レベルの情報を得ようとする試みであるともいえる。しかし，そのような試みは一般的には困難なものであることが古くから知られている。集計データは集計という過程がもたらす「歪み」を持っている場合が少なくなく，しかも，その「歪み」は推計結果を大きく歪めることが少なくないからである(注4)。本稿では，この「歪み」を持たない場合のみを扱っている(注5)。

2．「二項分布の和モデル」

離散型確率変数Zが試行回数x回，確率pの二項分布に従うことを，

$$Z \sim B(x, p) \quad \cdots\cdots(1)$$

と表す。さらに，離散型確率変数Z'が

$$Z' \sim B(x', p') \quad \cdots\cdots(2)$$

であり，

「Z，Z'が互いに独立である」……(3)

$$Y = Z + Z' \quad \cdots\cdots(4)$$

とする。(4)式は，確率関数Yがそれぞれ二項分布に従う2つの確率変数の和によって構成されているという，このモデルの基本仮定を示す式である(注6)。

このモデルの下では，Yも離散型確率変数となり，Yに関する離散型密度関数$f(y)$は，次のように表される。

$$f(y) = \sum_{z=zmin}^{zmax} \binom{x}{z} p^z (1-p)^{x-z} \cdot \binom{x'}{y-z} p'^{y-z} (1-p')^{x'-y+z} \quad \cdots\cdots(5)$$

ここで，$zmin = \max(0, y-x')$，また，$zmax = \min(y, x)$である。

あるいは，Yに関する累積分布関数$F(y)$の形で表すと，

$$F(y) = \sum_{\substack{z+z' \leq y \\ 0 \leq z \leq x \\ 0 \leq z' \leq x'}} \binom{x}{z} p^z (1-p)^{x-z} \cdot \binom{x'}{z'} p'^{z'} (1-p')^{x'-z'}$$

$$= \sum_{z=0}^{\min(x, y)} \left\{ \binom{x}{z} p^z (1-p)^{x-z} \cdot \sum_{z'=0}^{\min(x', y-z)} \binom{x'}{z'} p'^{z'} (1-p')^{x'-z'} \right\} \quad \cdots\cdots(6)$$

推定されるべきパラメーターp，p'を残して，(5)式，あるいは(6)式によって，モデルは完全に定められた。

パラメーターp，p'の値を決めるのが我々の課題であり，それには，概念的にいえばデータと最もよく適合するp，p'の値を以ってして定めればよいのはいうまでもない（このような値を，最適推定値と呼び，それぞれ，p^*，p'^*と表すこととする)。最適性の基準を明らかにした上で，これらの最適推定値の推定方法を述べることとしよう。

3．K-L情報量によるデータとモデルの適合性の表現

実際のデータとの適合性に基づく，p^*，p'^*の決定方式として，すぐ思いつくのが最尤推定法であろう。しかし，実際の計算をコンピュ

ーターで行う場合，xやx'を少し大きくすると（例えば，それぞれ500程度），(5)式の離散型密度関数は，極端に小さい値をとり，0として扱われ，計算が意味を持たなくなることが多い。そこで本稿では，以下で詳しく説明するように，最尤推定法とは異なったアプローチを工夫している。すなわち，まず，データとの適合性の判定に，K-L情報量の最小化という基準を採用することとする。

次に，K-L情報量の計算の際に必要なデータあるいはモデルの分布については，適当な区間設定を行い，各区間内の頻度によって，それらの分布を表現することとする。これによって，(5)式に代わって，(6)式の離散型累積分布関数を利用することが可能となり，発生する誤差を大きく減少することができる(注7)。

まず，K-L情報量は次のように定められる。いま離散型確率変数Yが，離散型密度関数$f(y)$を持つことを，

$$Y \sim f(y) \cdots\cdots (7)$$

と表す。(7)式の$f(y)$を，(5)式によって定義されたものとすれば，(7)式は，Yがここでのモデルどおりに分布していることを意味している。しかし，Yが実際には(5)式には（完全には）従っておらず，実は次式に従って分布しているとする。

$$Y \sim g(y) \cdots\cdots (8)$$

すなわち，$g(y)$がYの真の離散型密度関数を表しているものとする。

この時，K-L情報量は次式で定義される（坂元，石黒，北川 1983)(注8)。

$$KL=\sum_y g(y)\ln(g(y)/f(y)) \cdots\cdots (9)$$

ここで，

$$\sum_y g(y)=1,\ \sum_y f(y)=1 \cdots\cdots (10)$$

である。K-L情報量は，負の値をとることはなく，$f(y)$の$g(y)$に対する適合性が高いほど小さい値をとり，2つの分布が完全に一致する時，0となる，という性質を持つことが知られている(注9)。

4. K-L情報量の適用と計算方法——x, x'が固定されている場合

まず，x，x'が固定されている場合を考えよう。この時，(9)式を最小にするp，p'が，求める最適推定値p^*，p'^*であることは，K-L情報量の性質から明かである。

(9)式を計算するためには，真の分布に対応する$g(y)$の推計を行うことおよび，$f(y)$の値の計算が必要となる。yのとり得る範囲$[0, x+x']$を適当ないくつかの区間に分割することによって，これに相当する計算作業を行う。ここでは，5つの区間への分割を例として，議論を進めよう。

yのとり得る範囲の両端は，0と$x+x'$と決まっているので，間の4つの点y_1，y_2，y_3，y_4（$y_1<y_2<y_3<y_4$）を次式で定める。すなわち，ある与えられたp，p'の下で，

$$y_i=\max_{y:\text{整数}}\{F(y)\leqq i/5\} \quad i=1,2,3,4 \quad (11)$$

こうして，次の5つの区間が得られる。

$$I_1=[0, y_1],\ I_2=[y_1, y_2],\ I_3=[y_2, y_3],$$
$$I_4=[y_3, y_4],\ I_5=[y_4, x+x'] \cdots\cdots (12)$$

今，モデルによる，yがI_1において実現する確率をf_1，I_2において実現する確率をf_2，…等とすれば，f_5までが定められる。f_1からf_5は，それぞれ0.2に近い値を持ち，それらの和は1となっている。他方，実際のデータによって示される，yのI_1における実現頻度をg'_1，

I_2における実現頻度をg'_2,…等とし，g'_5までを得る。また，g'_1, g'_2,…, g'_5の和は1となっている。もし，データ数が十分であれば，g'_1, g'_2,…, g'_5を真の実現確率g_1, g_2,…, g_5とみなしてよいだろう（大数の法則）。(11)式に明らかなように，これらの計算では累積密度関数$F(y)$が用いられるので，生ずる計算誤差が小さいものとなる。

そこで，(9)式に代えて，与えられたp, p'の下でのK-L情報量が，

$$KL=\sum_i g'_i \ln(g'_i/f_i) \quad \cdots\cdots(13)$$

として計算できる(注10)。

5. K-L情報量の適用と計算方法——x, x'が固定されていない場合

しかし通常，実際に存在するデータは，x, x'が固定されていない。すなわち，x, x'が固定されてしかも多くのデータが存在するというのは実験的な状況でしかなく，社会科学で用いるデータでは，x, x'の組は様々な値を持つものとして与えられ，さらに，同一のx, x'の組に関するデータは複数はなく，複数あったとしてもデータの示す実現頻度を真の分布とみなせるほどのケース数はないのが普通である。そこで，x, x'が固定されていない場合の推計方法が必要となる。

今，様々な値を持つx, x'の組によって構成されるデータをx/x'の小さいものから大きいものの順に並べ，j番目の組をx_j, x'_j，対応する確率変数をY_jと表すことにしよう。

今，あるx_j, x'_jに固定した時の(9)(10)式によって与えられるK-L情報量を，

$$KL[g_j(y_j),\ f_j(y_j),\ y_j=0,1,2,\cdots,x_j+x'_j \mid p,\ p'] \quad \cdots\cdots(14)$$

と表すことにする。$g_j(y_j)$は，x_j, x'_jを固定した時のY_jの真の密度関数，$f_j(y_j)$は同じくY_jのモデルによる密度関数である。$y_j=0,1,2,\cdots,x_j+x'_j$は，これについて集計することを意味し，$p$, p'は，この式の値が$p=p$, $p'=p'$の場合に対応していることを示す。$y_j=0,1,2,\cdots,x_j+x'_j$の部分は，表記を省略することもある。

「二項分布の和モデル」によるモデルスペシフィケーションが完全に正しいならば，あるjの下で，(14)式を最小にするp, p'が求めるp^*, p'^*であり，さらにこの時，p^*, p'^*は任意のjの下で，(14)式を最小にするはずである。しかし，実際のデータではそのようなことは期待できない。

そこで，この時，求める最適値p^*, p'^*を，

$$\frac{\sum_j KL[g_j(y_j),\ f_j(y_j) \mid p,\ p']}{\text{全データの数}} \quad \cdots\cdots(15)$$

を，最小のものとするp, p'によって定める。これによって，p^*, p'^*が，個々のjに依存しないものとして定められることとなる。もし，「二項分布の和モデル」が比較的よく適合しているならば，こうして得られたp^*, p'^*の組と，任意のjの下で(14)式を最小にするp, p'の組の値は，等しいか，ほぼ等しいものとなろう(注11)。

以下，(15)式にあたるものを実際のデータでどのように計算するかを示す。

先に，x, x'を固定した時の方法に従って，次のようにy_jの値の取り得る区間を5つに分割する。まず，y_jのとり得る範囲の両端は，0と$x_j+x'_j$であり，他方，間の4つの点y_{j1}, y_{j2}, y_{j3}, y_{j4}($y_{j1}<y_{j2}<y_{j3}<y_{j4}$)は，次式で定める。ある与えられた$p$, p'の下で，

$$y_{ji}=\max_{y_i:\text{整数}}\{F_j(y_j)\leqq i/5\}$$
$$i=1,\ 2,\ 3,\ 4 \qquad \cdots\cdots(16)$$

すなわち，5つの区間は次のように得られる。

$$I_{j1}=[0,\ y_{j1}],\quad I_{j2}=[y_{j1},\ y_{j2}],$$
$$I_{j3}=[y_{j2},\ y_{j3}],\quad I_{j4}=[y_{j3},\ y_{j4}],$$
$$I_{j5}=[y_{j4},\ x_j+x'_j] \qquad \cdots\cdots(17)$$

いま，モデルによる，Y_jがI_{j1}において実現する確率をf_{j1}，I_{j2}において実現する確率をf_{j2}，…等とすれば，f_{j5}までが定められる。f_{j1}からf_{j5}は，それぞれ0.2に近い値を持ち，それらの和は1となっている。他方，Y_jの各区間における真の実現確率を同様に，g_{j1}，g_{j2}，…，g_{j5}と表すと，これらの和も1となっている。

そこで，(15)式は，次式で近似される。

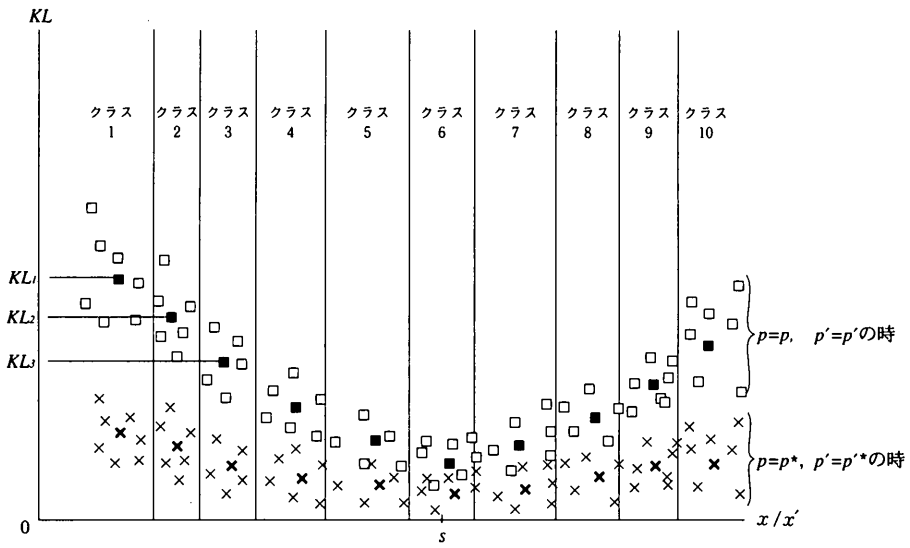
$$\frac{\sum_j KL[g_{ji},\ f_{ji},\ i=1,2,\cdots,5|\ p,\ p']}{\text{全データの数}} \qquad \cdots\cdots(18)$$

もし，任意のあるデータのx_j, x'_jの組とそれぞれ同一の値を持つx，x'のデータの組が常に十分な数だけあるならば，(18)式の計算は，先に見たx，x'が固定した場合と同様に行うことができる。しかし，通常，任意のデータのx_j, x'_jの組と同一の値を持つデータの数は，それ自身のただ1つか，多くとも数個程度であろうから，そのx_j, x'_jに対応した実現頻度を計算することができない。

そこで，全データをx/x'の値によってクラス分けし，各クラスごとにまとめた扱いをすることによって，実現頻度を得ることにする。以下，図を利用しながらこのことを説明する。図は，わかりやすさのため，「二項分布の和モデル」が比較的よく真の分布にあてはまっている場合を示すことにする。

図2は，横軸にx/x'の値を，縦軸を$KL[g_{ji},\ f_{ji},\ i=1,2,\cdots,5|\ p,\ p']$の値（(18)式の各$j$番目の成分）とし，$p=p^*$，$p'=p'^*$の

図2　各データに対するK-L情報量



（出所）　筆者作成。

時，および$p=p\,(>p^*)$，$p'=p'\,(<p'^*)$の時について，それぞれプロットしたものであり，1つの点が1つのデータを表している。$p=p^*$，$p'=p'^*$の時(×印の点)，K-L情報量は，x/x'の値によってあまり変化せず，×印の点はKL=一定の直線の付近に存在している。また，$p=p$，$p'=p'$の時(□印の点)，sをある実数値とすると，□印の点は，$x/x'=s$の付近で最小値を持つU型の曲線の付近に存在している。また，いずれの場合も，もしx/x'が同一の値を持つ点が複数存在するならば，それらの点は互いに近くに存在している。これらは，モデルが二項分布を基礎としていること，モデルのあてはまりが比較的よいとしたことから導かれる。

ここで，例えば，□印の点に注目することにしよう。すなわち，あるp，p'の組を固定する。これらをx/x'の値によってk個のクラスに分け，各クラスのK-L情報量の平均値を$\underline{KL_1}$，$\underline{KL_2}$，…，$\underline{KL_c}$，…，$\underline{KL_k}$，等とする。すなわち，

$$\underline{KL_c}=\frac{\sum_{\substack{c\text{内の}\\ \text{全ての}j}} KL[g_{ji},\ f_{ji},\ i=1,2,\cdots,5\,|\,p,\ p']}{c\text{内のデータの数}}$$

$$c=1,\ 2,\cdots,\ c,\cdots,\ k \quad \cdots\cdots(19)$$

図中では，$k=10$とし，それらを■印の点で示してある。

$$w_c=\frac{c\text{内のデータ数}}{\text{すべてのデータ数}} \quad \cdots\cdots(20)$$

とする時，(18)式は次式に等しい。

$$\sum_c w_c\cdot\underline{KL_c} \quad \cdots\cdots(21)$$

実際のデータからは，これら□印の点を得ることができないので，■印の点すなわち，$\underline{KL_1}$，$\underline{KL_2}$，…，$\underline{KL_c}$，…，$\underline{KL_k}$も得ることができない。しかし，次の近似を想定することにする(注12)。

$$\underline{KL_c}\fallingdotseq KL[\underline{g_{ci}},\ \underline{f_{ci}},\ i=1,2,\cdots,5\,|\,p,\ p'] \quad \cdots\cdots(22)$$

ここで，$\underline{g_{ci}}$および，$\underline{f_{ci}}$は次式によって定められる。

$$\underline{g_{ci}}=\frac{\sum_{\substack{c\text{内の}\\ \text{全ての}j}} g_{ji}}{c\text{内のデータ数}} \quad \cdots\cdots(23)$$

$$\underline{f_{ci}}=\frac{\sum_{\substack{c\text{内の}\\ \text{全ての}j}} f_{ji}}{c\text{内のデータ数}} \quad \cdots\cdots(24)$$

(24)式の値は，(16)式から容易に計算される(注13)。(23)式の値には，データを用いて次のように接近する。

まず，離散型確率変数G'_{ji}を次のように定義する。

$$g'_{ji}=\begin{cases}1 & (y_j\in I_{ji}\text{の時})\\ 0 & (y_j\notin I_{ji}\text{の時})\end{cases} \quad \cdots\cdots(25)$$

この式に従い，各データのy_jの値によって，g'_{ji}の値を得ることができる。また，この離散型密度関数$f(g'_{ji})$は，明らかに，

$$f(g'_{ji})=\begin{cases}g_{ji} & (g'_{ji}=1\text{の時})\\ 1-g_{ji} & (g'_{ji}=0\text{の時})\end{cases} \quad \cdots\cdots(26)$$

そこで，c内のデータ数が十分大きいものとすると，大数の法則を応用することにより(注14)，

$$\frac{\sum_{\substack{c\text{内の}\\ \text{全ての}j}} g'_{ji}}{c\text{内のデータ数}}\fallingdotseq\underline{g_{ci}} \quad \cdots\cdots(27)$$

を得ることができる(注15)。

6. p^*, p'^*の関係の推定

確率変数Z，Z'が互いに独立であるということが成立している時，p，p'の最適推定値p^*，

p'^*の間に，次の関係を想定することができる。

$$p^*\sum x_j + p'^*\sum x'_j = \sum y_j \quad \cdots\cdots(28)$$

この式は，次のように導かれる。仮に，図1の全地域集計表において，Zの実現値データの全地域集計値$\sum z_j$が知られているとする。明らかに，最尤推定によって，

$$p^*\sum x_j = \sum z_j \quad \cdots\cdots(29)$$

同時に，Z'の実現値データの全地域集計値$\sum z'_j$が知られているとすると，

$$p'^*\sum x'_j = \sum z'_j \quad \cdots\cdots(30)$$

(29)(30)式の両辺の和をそれぞれとれば，(28)式が得られる。この式によって，推定すべきパラメーターを事実上1つとすることができた(注16)。

II　データと計算結果

1．被説明変数および説明変数，適用のためのデータ変換

本稿で用いるセンサスデータ(メキシコの1990年センサス，メキシコ連邦地区，基本地域 Area Geográfica Básica 別）は図3のような構造を持っている(注17)。変数1，変数1′，変数2，変数2′，…等の各変数は，地域ごとに，ある属性を有する（該当の）個人あるいは住居がいくつあったか，有していない（非該当の）個人あるいは住居がいくつあったかを示している。

より詳しく説明すれば，変数1～変数2′は教育機会の有無に関する変数であり，変数1は，「(15～17歳で)高校入学以上の学歴」の該当者数を表し，変数1′はその非該当者数を表わし，変数2は「(6～14歳で)いずれかの教育機関に在学中」の該当者数を表わし，変数2′はその非該当者数を表わしている。他方，変数3～変数36′は個人や住居の社会特性の有無を表す変数であり，例えば，変数3は「連邦地区生まれ」該当者数，変数3′はその非該当者数，変数4は「1985年に連邦地区に居住」該当者数，変数4′はその非該当者数，…，あるいは，変数23は「(個人住宅中)屋根がコンクリート・瓦」該当住宅数，変数23′はその非該当住宅数，変数24

図3　センサスデータの基本構造

	(被説明変数)				(説明変数)						
	変数1	変数1′	変数2	変数2′	変数3	変数3′	変数4	変数4′	……	変数36	変数36′
	(15～17歳人口中)高校入学者		(6～14歳人口中)就学中の者		(全人口中)連邦地区出生者		(5歳以上人口中)1985年連邦地区居住者		……	(個人住宅中)居住者非所有の物	
	該当	非該当	該当	非該当	該当	非該当	該当	非該当	……	該当	非該当
地域　1	y_1	y'_1			u_1	u'_1					
地域　2	y_2	y'_2			u_2	u'_2					
地域　3	y_3	y'_3			u_3	u'_3			……		
地域　4	y_4	y'_4			u_4	u'_4					
・	・	・			・	・					
・	・	・			・	・					
地域 1945	y_{1945}	y'_{1945}			u_{1945}	u'_{1945}					

(出所)　筆者作成。

は「(個人住宅中)壁がレンガ」該当住宅数，変数24′はその非該当住宅数，…，等を表している。

第Ⅰ章の議論での確率変数（被説明変数）Yが変数1にあたり，Yの実現値が各地域の該当者数y_1，y_2，y_3，…，にあたることは明らかであろう。ここでは，Yとして変数1の場合だけでなく，変数2の場合についても，推計を行うことにする。すなわち，被説明変数として用いられる変数の数は，2通りである。

また，第Ⅰ章の議論での説明変数x，x'に対応しているのは，変数20，変数20′である。ここでは，x，x'にあたるものとして，それ以外の変数セットの場合（すなわち，変数3と変数3′のセットから変数36と変数36′のセットまでのすべて，計34セット）についても，推計を行うことにする。

ただし，ここで注意しなければならないのは，このセンサスデータの変数のセットをそのまま，第Ⅰ章で扱ったx，x'そのものとして扱うことはできない，ということである（それゆえここでは，センサスデータの変数のセットを，u，u'と表わすこととし，この変数の各地域での値を，図2では，u_1，u'_1，u_2，u'_2，u_3，u'_3，…，等と表わしてある)。なぜなら，$Y+Y'$が，変数1＋変数1′の時，これは15～17歳人口に対応し，変数2＋変数2′の時，これは6～14歳人口に対応している。以下同様に，変数3＋変数3′，…はそれぞれ，一定の年齢の人口，労働力人口，あるいは個人住宅数であったりしている。したがって，$y_j+y'_j=n_j$（$j=1,2,3,\cdots$）とすると，ほとんどの場合，$u_j+u'_j\neq n_j$（$j=1,2,3,\cdots$）であり，このままでは第Ⅰ章で議論したような2×2の分割表の形をとることができないからである。

そこで，u_j，u'_jから2×2の分割表に適応するx_j,x'_jを得るために次のような変換を行う。

$$\frac{x_j/u_j}{x'_j/u'_j}=m \quad j=1,\ 2,\ 3,\cdots \qquad \cdots\cdots\cdots\cdots(31)$$

ここで，

$$x_j+x'_j=n_j(=y_j+y'_j) \quad j=1,\ 2,\ 3,\cdots \qquad \cdots\cdots\cdots\cdots(32)$$

である。mはパラメーターであり，本稿では，$m=1/2$，1，2の3通りの場合を計算している。また，x_j，x'_jは整数でなくてはならないから，(31)式をx_jについて解き，その小数点以下を四捨五入したものをx_jの値とし，$x'_j=n_j-x_j$とする。この変換式（(31)式）の意味，解釈は以下の計算結果の提示の際に行う。

以上，推計で用いられる説明変数のセットの数は，34×3通りとなる。

2. データのクラス分け，実際のp^*，p'^*の計算方法

こうして，センサスの原データに基づいて，モデルの適用に必要なデータが準備されたことになるが，推計を行うモデルの個数を改めて確認すると，2（＝被説明変数の数）×(34×3)（＝説明変数のセットの数）の計204通りである。また，この1通りごとに，1945のデータ（＝地域数）が存在する，言い換えると，この1通りごとに，地域ごとの2×2分割表が1945個得られる（すなわち，$j=1,\ 2,\ 3,\cdots,\ 1945$）こととなる。

前節で述べたモデルに適用するために，1つの説明変数のセットごとに，データを，x_j/x'_jが小さいものから順に並べて，小さいものから，194ずつのデータを含む9個のクラスと199のデータを含む1個のクラス，計10個のクラスに分

表1　15～17歳の者の家族あるいは住宅特性による高校入学以上である確率：「二項分布の和モデル」に基づく推計結果

Σn_j＝15～17歳人口＝583,099人　Σy_j＝うち高校入学以上の者＝162,145人　$\Sigma y_j/\Sigma n_j$＝0.28

		m＝0.5を仮定した時				m＝1を仮定した時				m＝2を仮定した時			
		確率 p に対応する家族／住宅	15～17歳の者が高校入学以上である確率		推計された確率に対応するK-L情報量	確率 p に対応する家族／住宅	15～17歳の者が高校入学以上である確率		推計された確率に対応するK-L情報量	確率 p に対応する家族／住宅	15～17歳の者が高校入学以上である確率		推計された確率に対応するK-L情報量
			家族全員／住宅がその特性を有する時	家族全員／住宅がその特性を有しない時			家族全員／住宅がその特性を有する時	家族全員／住宅がその特性を有しない時			家族全員／住宅がその特性を有する時	家族全員／住宅がその特性を有しない時	
変数番号	説明変数(家族メンバー／住宅の特性)	$\Sigma x_j/\Sigma n_j$	p	p'	KL	$\Sigma x_j/\Sigma n_j$	p	p'	KL	$\Sigma x_j/\Sigma n_j$	p	p'	KL
3	(全人口中)連邦地区出生者	.60	.30	.25	.36	.75	.30	.20	.36	.85	.29	.21	.36
4	(5歳以上人口中)1985年連邦地区居住者	.91	.29	.17	.34	.95	.29	.09	.35	.98	.29	.00	.35
5	(15歳以上人口中)小学校入学以上者	.89	.31	.00	.51	.94	.29	.01	.56	.97	.29	.00	.61
6	(15歳以上人口中)小学校卒業以上者	.71	.39	.00	.30	.82	.34	.00	.46	.90	.31	.00	.57
7	(15歳以上人口中)中学校入学以上者	.48	.54	.04	.20	.64	.44	.00	.22	.77	.36	.00	.40
8	(15歳以上人口中)中学校卒業以上者	.38	.60	.08	.23	.54	.50	.01	.19	.69	.40	.00	.29
9	(15歳以上人口中)高校入学以上者	.25	.70	.14	.27	.38	.54	.12	.20	.53	.48	.05	.14
10	(18歳以上人口中)高校入学以上者	.26	.60	.17	.27	.39	.52	.13	.20	.54	.47	.06	.15
11	(18歳以上人口中)大学入学以上者	.10	.80	.22	.38	.17	.68	.20	.34	.27	.52	.19	.29
12	(12歳以上人口中)生徒・学生	.12	.56	.24	.38	.21	.45	.23	.38	.34	.37	.23	.38
13	(労働力中)就業者	.95	.29	.00	.36	.97	.29	.00	.37	.99	.28	.05	.37
14	(労働力中)第3次産業就業者	.49	.49	.07	.27	.65	.43	.00	.28	.79	.35	.00	.36
15	(労働力中)被雇用者	.61	.40	.08	.35	.75	.36	.02	.35	.86	.32	.01	.37
16	(労働力中)非自営業者	.73	.34	.11	.34	.84	.33	.00	.35	.91	.30	.00	.36
17	(労働力中)週32時間以下就業者	.09	.79	.23	.37	.17	.58	.22	.37	.29	.47	.20	.36
18	(労働力中)週40時間以下就業者	.33	.70	.07	.28	.50	.56	.00	.26	.66	.42	.00	.35
19	(労働力中)週48時間以下就業者	.59	.42	.07	.36	.74	.37	.00	.38	.85	.33	.00	.42
20	(労働力中)月収1法定最低賃金より大の者	.67	.41	.00	.35	.80	.35	.00	.43	.89	.31	.00	.50
21	(労働力中)月収2法定最低賃金より大の者	.23	.69	.15	.26	.36	.57	.11	.23	.52	.50	.03	.22
22	(労働力中)月収5法定最低賃金より大の者	.06	.71	.25	.42	.11	.70	.22	.41	.20	.50	.22	.39
23	(個人住宅中)屋根がコンクリート・瓦の物	.67	.37	.10	.19	.78	.35	.01	.21	.87	.32	.00	.32
24	(個人住宅中)壁がレンガの物	.92	.30	.05	.31	.96	.29	.00	.34	.98	.28	.02	.36
25	(個人住宅中)床が非セメントの物	.25	.46	.22	.21	.35	.41	.21	.17	.47	.38	.19	.18
26	(個人住宅中)2室以上の物	.88	.32	.00	.36	.93	.30	.01	.44	.96	.29	.00	.51
27	(個人住宅中)6室以上の物	.11	.64	.23	.35	.19	.50	.23	.33	.30	.43	.21	.32
28	(個人住宅中)2寝室以上の物	.50	.42	.14	.27	.65	.39	.07	.26	.78	.36	.00	.28
29	(個人住宅中)専用キチンを持つ物	.69	.41	.00	.26	.81	.34	.00	.36	.89	.31	.00	.45
30	(個人住宅中)キチンを持つ物	.80	.34	.01	.34	.89	.31	.00	.37	.94	.29	.04	.38
31	(個人住宅中)調理用ガスを使用する物	.95	.29	.06	.34	.97	.29	.00	.36	.99	.28	.06	.36
32	(個人住宅中)公共下水道施設を有する物	.78	.31	.16	.34	.83	.30	.18	.36	.87	.30	.14	.36
33	(個人住宅中)下水道施設を有する物	.88	.30	.08	.30	.93	.30	.02	.32	.95	.29	.00	.34
34	(個人住宅中)電気施設を有する物	.98	.28	.01	.38	.99	.28	.00	.39	.99	.28	.02	.40
35	(個人住宅中)住居内に水道施設を有する物	.57	.38	.15	.18	.68	.35	.12	.19	.78	.33	.09	.24
36	(個人住宅中)それが非自己所有である物	.21	.38	.25	.34	.33	.36	.24	.34	.47	.31	.25	.34

(出所)　1990年メキシコ人口・住居センサス (Instituto Nacional de Estadística, Geografía e Información 1992. *CODICE 90, Resultados definitivos, XI censo general de población y vivienda, 1990.* Aguascalientes. CD-ROM版) のメキシコ連邦地区，基本地域別データより計算。

表2　6～14歳の者の家族あるいは住宅特性による就学中である確率：「二項分布の和モデル」に基づく推計結果

$\sum n_j$＝6～14歳人口＝1,500,546人　$\sum y_j$＝うち就学中の者＝1,426,647人　$\sum y_j / \sum n_j = 0.95$

変数番号	説明変数(家族メンバー／住宅の特性)	m＝0.5を仮定した時				m＝1を仮定した時				m＝2を仮定した時			
		確率 p に対応する家族／住宅	6～14歳の者が就学中である確率		推計された確率に対応するK-L情報量	確率 p に対応する家族／住宅	6～14歳の者が就学中である確率		推計された確率に対応するK-L情報量	確率 p に対応する家族／住宅	6～14歳の者が就学中である確率		推計された確率に対応するK-L情報量
			家族全員／住宅がその特性を有する時	家族全員／住宅がその特性を有しない時			家族全員／住宅がその特性を有する時	家族全員／住宅がその特性を有しない時			家族全員／住宅がその特性を有する時	家族全員／住宅がその特性を有しない時	
		$\sum x_j / \sum n_j$	p	p'	KL	$\sum x_j / \sum n_j$	p	p'	KL	$\sum x_j / \sum n_j$	p	p'	KL
3	(全人口中)連邦地区出生者	.60	.96	.94	.29	.75	.97	.91	.29	.85	.96	.91	.29
4	(5歳以上人口中)1985年連邦地区居住者	.91	.96	.83	.26	.95	.96	.74	.27	.98	.96	.64	.27
5	(15歳以上人口中)小学校入学以上者	.89	.98	.72	.19	.94	.98	.56	.20	.97	.97	.32	.22
6	(15歳以上人口中)小学校卒業以上者	.70	.98	.88	.20	.82	.98	.82	.18	.90	.98	.71	.18
7	(15歳以上人口中)中学校入学以上者	.47	1.00	.91	.26	.63	.99	.88	.23	.76	.99	.83	.20
8	(15歳以上人口中)中学校卒業以上者	.37	1.00	.92	.28	.53	.99	.90	.25	.68	.98	.88	.22
9	(15歳以上人口中)高校入学以上者	.24	1.00	.94	.30	.37	.99	.93	.27	.52	.99	.91	.24
10	(18歳以上人口中)高校入学以上者	.24	1.00	.93	.31	.38	.99	.93	.28	.53	.98	.92	.25
11	(18歳以上人口中)大学入学以上者	.09	1.00	.95	.37	.16	.98	.94	.35	.26	.99	.94	.32
12	(12歳以上人口中)生徒・学生	.11	.95	.95	.30	.20	1.00	.94	.30	.34	.99	.93	.29
13	(労働力中)就業者	.95	.95	.94	.27	.97	.95	.93	.27	.99	.95	.90	.27
14	(労働力中)第3次産業就業者	.49	.99	.91	.23	.65	.98	.89	.22	.78	.99	.81	.21
15	(労働力中)被雇用者	.60	1.00	.88	.25	.75	1.00	.81	.25	.86	.99	.69	.27
16	(労働力中)非自営業者	.73	.98	.88	.25	.84	.97	.83	.25	.91	.97	.78	.26
17	(労働力中)週32時間以下就業者	.09	1.00	.95	.27	.17	.99	.94	.27	.28	.98	.94	.26
18	(労働力中)週40時間以下就業者	.33	1.00	.93	.26	.49	1.00	.90	.19	.65	1.00	.86	.15
19	(労働力中)週48時間以下就業者	.59	1.00	.88	.25	.74	1.00	.81	.23	.85	1.00	.69	.23
20	(労働力中)月収1法定最低賃金より大の者	.67	1.00	.85	.20	.80	.99	.78	.20	.89	.99	.66	.22
21	(労働力中)月収2法定最低賃金より大の者	.22	1.00	.94	.28	.36	1.00	.92	.26	.51	.98	.92	.24
22	(労働力中)月収5法定最低賃金より大の者	.06	1.00	.95	.34	.11	.98	.95	.34	.19	.98	.94	.33
23	(個人住宅中)屋根がコンクリート・瓦の物	.65	.97	.92	.20	.77	.97	.89	.19	.86	.97	.84	.19
24	(個人住宅中)壁がレンガの物	.92	.96	.86	.22	.96	.96	.78	.23	.98	.96	.64	.23
25	(個人住宅中)床が非セメントの物	.23	.98	.94	.27	.33	.98	.94	.25	.45	.97	.93	.23
26	(個人住宅中)2室以上の物	.87	.97	.83	.18	.93	.97	.73	.18	.96	.97	.55	.18
27	(個人住宅中)6室以上の物	.10	.99	.95	.29	.17	.99	.94	.29	.28	.97	.94	.27
28	(個人住宅中)2寝室以上の物	.48	.98	.93	.22	.63	.97	.91	.22	.77	.97	.88	.22
29	(個人住宅中)専用キチンを持つ物	.68	.99	.87	.18	.80	.98	.82	.16	.89	.98	.71	.16
30	(個人住宅中)キチンを持つ物	.80	.98	.85	.22	.88	.97	.80	.22	.94	.97	.68	.23
31	(個人住宅中)調理用ガスを使用する物	.95	.96	.77	.24	.97	.96	.60	.25	.99	.96	.45	.26
32	(個人住宅中)公共下水道施設を有する物	.76	.96	.92	.23	.82	.96	.92	.24	.86	.96	.91	.25
33	(個人住宅中)下水道施設を有する物	.87	.96	.89	.22	.92	.96	.87	.22	.95	.96	.82	.23
34	(個人住宅中)電気施設を有する物	.97	.96	.70	.23	.99	.96	.48	.23	.99	.96	.16	.23
35	(個人住宅中)住居内に水道施設を有する物	.54	.97	.93	.20	.66	.97	.92	.18	.76	.97	.90	.17
36	(個人住宅中)それが非自己所有である物	.21	.98	.94	.24	.33	.97	.94	.24	.47	.96	.94	.24

(出所)　表1に同じ。

類する（すなわち，$c=1, 2, 3, \cdots, 10$）。また，第Ⅰ節で述べた$\underline{f_{ci}}$を定め，$\underline{g_{ci}}$を求めるための区間分割は，そこで述べたように，5区間への分割とする（すなわち，$i=1, 2, 3, 4, 5$）。

こうして，p，p'を与えると，第Ⅰ節の(22)式の近似を通じて，(21)式にあたるものの計算が可能となった。p^*，p'^*を解析的に求めることは困難であるので，コンピューターを用いて，第Ⅰ節の(28)式の関係を前提としながら，いわゆる格子探索（グリッド・サーチ）法によりながら，様々な値のp，p'の下で(21)式を計算し，それを最小化する最適のものを探していくという方法を用いる(注18)。

3．計算結果，その評価のための回帰分析結果との比較

表1，表2は，以上の手続きを経て得た，p^*，p'^*の推計結果である。

ここでは，「二項分布の和モデル」の有効性を示すことを主眼として，分析結果の社会科学的な議論は省略するが，ただ，(31)式で用いたパラメーターmをどのように解釈するかということについては，説明しておく必要があろう。もともとセンサスデータでは，変数3から変数22′までは人数を単位としており，変数23から変数36′までは住居数を単位としていたことを考慮して，次のような解釈が可能であろう。

まず，変数23から変数36′までの住居環境の変数についてみよう。住居数＝家族数とみなせば，u_jとu'_jは，それぞれある特性を持つ住居に住む家族数とその特性を持たない住居に住む家族数を表わす。また，前者に属す子供の数をx_j，後者に属す子供の数をx'_jとすれば，(x_j/u_j)は前者の家族における平均的な子供数，(x'_j/u'_j)は後者の家族における平均的な子供数である。そこで，mは，(31)式によって，これらの2種類の家族の平均的な子供数の比となっている。

変数3から変数22′までの個人の社会的特性の変数については，これを家族特性の変数と読み替えるための次のような手続きが必要である。

まず，家族メンバー全員がその社会的特性を持つ時，その子供が教育機会を持つ確率をpとし，家族メンバー全員がその社会的特性を持たない時，その子供が教育機会を持つ確率をp'とする。家族メンバーa人がその社会的特性を持ち，a'人が持たない時には，このような構成を持つ家族のうち$a/(a+a')$の家族が子供の教育機会の確率pを持ち，$a'/(a+a')$の家族が子供の教育機会の確率p'を持つとする。このようにしてすべての家族を，子供の教育機会に関して，確率pの家族と確率p'の家族に分ける。地域j内における確率pの家族数をx_j，確率p'の家族数をx'_jとする。

この時，(31)式を変形して得られる$x_j/x'_j=m(u_j/u'_j)$から明らかなように，mはu_j/u'_jからx_j/x'_jを得るための「修正係数」となっている。このような修正の必要性は，次の3つの要因から発生する。(1)教育機会有無の問題の対象となっている子供を有する家族メンバーと有しない家族メンバーの社会特性分布が異なる可能性，(2)確率pの家族メンバー数平均と確率p'家族メンバー数平均が異なる可能性，(3)確率pの家族と確率p'の家族では，教育機会有無の問題の対象となっている子供に関して，その平均的な数が異なっている可能性。仮に，(1)と(2)が無視できる程度のものであるとすれば，mは，住宅環境の変数と同様に，これら2種類の家族の平均的な子供数の比とみなすことがで

きる。

表1と表2に示された「二項分布の和モデル」による計算結果を評価するにあたって，地域ごとの集計値を使った回帰分析の結果を参照することにしよう。

回帰モデルは，次のように表される。

$$Y = px + p'x' + E \quad \cdots\cdots(33)$$

ここで，Eは誤差項を表わす確率変数であり，次の正規分布に従うと仮定する。

$$E \sim N(0,\ xp(1-p) + x'p'(1-p')) \quad \cdots\cdots(34)$$

この仮定によって，(33)式は，本質的には，「二項分布の和モデル」を正規分布近似したものとみなすことができる。

これの両辺を $n(=x+x')$ で割る。すると次式が得られる。

$$Y_r = cx_r + p' + E_r \quad \cdots\cdots(35)$$

ここで，

$$Y_r = Y/n,\quad x_r = x/n,\quad c = p - p',$$
$$E_r = E/n \quad \cdots\cdots(36)$$

である。

E_r は次の正規分布に従う。

$$E_r \sim N\left(0,\ \frac{xp(1-p)}{n^2} + \frac{x'p'(1-p')}{n^2}\right) \quad \cdots\cdots(37)$$

(35)式の推計を通常の回帰分析によって行うことは不適切である。E_rの分散は(37)式に示されるように，一定値ではなく，しかもY_rと関わり無くx，x'，p，p'，nによって与えられているからである。

しかし，ここで，(37)式や(34)式の仮定から離れ，次式が成立しているとする。

$$E_r = N(0,\ \theta^2/n) \quad \cdots\cdots(38)$$

θ^2 は，定数である。これは，既存のコンピューター統計分析パッケージを用いた推計を行うための便宜的な近似法と解釈してもよいし，他の攪乱要因が作用して実際このモデルが適切となっている，と仮定してもよい。すると，重み付けした回帰分析が適用可能となる。

また，パラメーター$m=1$とする時，(31)式，(32)式による変換を行わずに，

$$x_r = u/(u+u') \quad \cdots\cdots(39)$$

とすることができる。

表3は，$m=1$の場合について，このような手続きによる推計，すなわち，各地域ごとに，それぞれの属性を持つ者の比率を算出し，この比率を用いて回帰分析（従属比率変数作成に使われた各地域の人口で重み付けした）を行った結果である。

これら2つの方法による「教育機会享受の確率」の推定値を比較しよう。回帰分析（表3）において，それらがマイナスの値や1を超える値をとっているのに対し，当然のことながら，「二項分布の和モデル」（表1や表2）においてはそのようなことは生じていない。

また，表3において，変数4に対応する属性は変数1と変数2の場合で，一貫しない効果を示している（その属性があるほうが，教育機会の存在に関して，一方ではより低い確率を示しているのに，他方では，より高い確率を示している）。これに対し，表1と表2は，一貫した結果が見られる。

上記の点を除くと，$m=1$を仮定した時の「二項分布の和モデル」による結果と回帰分析結果とは近い数値となっていることが多い。

表3　各地域の従属変数の人口によって重みを付けた単純回帰分析の結果

変数番号	独立変数	従属変数			
		15～17歳の者が高校入学以上である確率		6～14歳の者が就学中である確率	
		家族全員／住宅がその特性を有する時 p	家族全員／住宅がその特性を有しない時 p'	家族全員／住宅がその特性を有する時 p	家族全員／住宅がその特性を有しない時 p'
3	(全人口中)連邦地区出生者	0.29[a]	0.25	0.97	0.89
4	(5歳以上人口中)1985年連邦地区居住者	0.28[a]	0.32	0.96	0.80
5	(15歳以上人口中)小学校入学以上者	0.40	−1.83	0.98	0.50
6	(15歳以上人口中)小学校卒業以上者	0.44	−0.46	0.99	0.79
7	(15歳以上人口中)中学校入学以上者	0.48	−0.08	0.99	0.88
8	(15歳以上人口中)中学校卒業以上者	0.51	0.00[a]	1.00	0.90
9	(15歳以上人口中)高校入学以上者	0.54	0.12	1.00	0.92
10	(18歳以上人口中)高校入学以上者	0.52	0.13	1.00	0.92
11	(18歳以上人口中)大学入学以上者	0.65	0.20	1.02	0.94
12	(12歳以上人口中)生徒・学生	0.83	0.14	1.15	0.90
13	(労働力中)就業者	0.33	−1.74	0.95	0.80
14	(労働力中)第3次産業就業者	0.46	−0.06	0.98	0.89
15	(労働力中)被雇用者	0.42	−0.14	1.00	0.81
16	(労働力中)非自営業者	0.38	−0.26	0.98	0.81
17	(労働力中)週32時間以下就業者	0.87	0.16	0.99	0.94
18	(労働力中)週40時間以下就業者	0.62	−0.06	1.02	0.88
19	(労働力中)週48時間以下就業者	0.52	−0.41	1.02	0.75
20	(労働力中)月収1法定最低賃金より大の者	0.48	−0.54	1.00	0.76
21	(労働力中)月収2法定最低賃金より大の者	0.58	0.10	1.01	0.92
22	(労働力中)月収5法定最低賃金より大の者	0.68	0.23	1.02	0.94
23	(個人住宅中)屋根がコンクリート・瓦の物	0.36	−0.01[a]	0.97	0.89
24	(個人住宅中)壁がレンガの物	0.30	−0.16	0.96	0.83
25	(個人住宅中)床が非セメントの物	0.42	0.20	0.98	0.94
26	(個人住宅中)2室以上の物	0.34	−0.57	0.97	0.75
27	(個人住宅中)6室以上の物	0.46	0.24	0.98	0.94
28	(個人住宅中)2寝室以上の物	0.39	0.07	0.97	0.91
29	(個人住宅中)専用キチンを持つ物	0.40	−0.25	0.98	0.82
30	(個人住宅中)キチンを持つ物	0.34	−0.25	0.97	0.82
31	(個人住宅中)調理用ガスを使用する物	0.30	−0.37	0.96	0.72
32	(個人住宅中)公共下水道施設を有する物	0.30	0.18	0.96	0.93
33	(個人住宅中)下水道施設を有する物	0.30	0.02	0.96	0.88
34	(個人住宅中)電気施設を有する物	0.29	−0.63	0.95	0.68
35	(個人住宅中)住居内に水道施設を有する物	0.35	0.13	0.97	0.92
36	(個人住宅中)それが非自己所有である物	0.30[b]	0.27	0.96	0.95

(注)　通常の t 検定で，a：10%水準で有意でない。b：5%水準で有意。他はすべて，1%水準で有意。
ただし，p の欄については，$c\,(=p-p')$ の有意性を示す。

(出所)　表1に同じ。

おわりに

理論的には，「二項分布の和モデル」と回帰モデルの差は，前者が二項分布を基礎にした確率モデルであるのに対し，後者は正規分布を基礎にした確率モデルである点である。それらのモデルの基本構造は変わらないということもでき，先の結果はそのことを裏書きしていよう。

しかし，本稿で扱ってきたような p，p' といった確率自体を推計するには，従来の回帰分析より，「二項分布の和モデル」は好ましい結果を与えるものと考えられる。このことも，先の結果から明らかである。

本稿の方法が，実際の計算においてよい結果をもたらすには，ある程度大きいサンプル数を必要とする。また，既成のコンピューター用統計計算パッケージをそのまま適用した計算はできないので，自分でかなり複雑なプログラムを組む必要がある。そして，パーソナルコンピューターを使用する場合は，計算時間もかなりなものになる(注19)。

しかし，この方法におけるモデルの単純明快さ（したがって解釈が簡単にできること）は魅力的なものであり，また，データが2カテゴリー変数やそれに変換可能な形で与えられることの多い社会科学的分析において，応用可能な場面も少なくないと考えられる。

また，理論的には容易に，本稿の方法の拡張が思いつかれる。すなわち，第1に，第Ⅰ節において(注3)で述べた，「多項分布モデル」への拡張がある。これは，因果関係を前提としない，統計的分布の推計モデルとなっている。第2は，「多数の二項分布の和モデル」への拡張である。これは，「被説明変数」は2カテゴリー変数であるが，「説明変数」が3カテゴリー以上の変数の場合（すなわち，2×3，2×4，2×5，…，等の分割表）に適用されるものである。いずれも，方法的な構造は，本稿に示されたものと同様であるので，後者についてだけ，最も簡単な，3カテゴリーの「説明変数」の場合に対応する「3つの二項分布の和モデル」の素描を行っておく。

まず，離散型確率変数 Z，Z'，Z'' をそれぞれ次の二項分布として定める。

$$\begin{aligned} &Z\sim B(x,\ p)\\ &Z'\sim B(x',\ p')\\ &Z''\sim B(x'',\ p'') \end{aligned} \quad \cdots\cdots(40)$$

さらに，

「Z, Z', Z'' はそれぞれ互いに独立である」…(41)

$$Y=Z+Z'+Z'' \quad \cdots\cdots(42)$$

とする。すると，第Ⅰ節の(5)式に対応する次式が得られる。

$$f(y)=\sum_{\substack{y=z+z'+z''\\ \text{を満たす可能な}\\ z,\ z',\ z''\text{の組}\\ \text{み合わせ}}}\binom{x}{z}p^{z}(1-p)^{x-z}\cdot\binom{x'}{z'}p'^{z'}(1-p')^{x'-z'}\cdot\binom{x''}{z''}p''^{z''}(1-p'')^{x''-z''} \quad \cdots\cdots(43)$$

さらに，(28)式に対応するものとして，

$$p\Sigma x+p'\Sigma x'+p''\Sigma x''=\Sigma y \quad \cdots\cdots(44)$$

という条件を与える。したがって，p，p'，p'' の3つのパラメーターのうち，いずれかの2つを推定する必要があり，実際の推計を格子探索

法によって行っていくとかなり大変なものとなる。

ところで，いうまでもなく，統計学的あるいは数学的観点から本稿の理論的不十分点を多く指摘することができよう。最適解決定におけるK-L情報量の使用の意味や(注20)，注で扱ってきたような問題（近似計算の問題，データをx/x'の値によってクラス分けすることの意味，推計の「安定性」の問題，K-L情報量をp，p'の関数とした時の解析的性質の検討）をより厳密に議論することは，残された課題といえよう。

同時に，応用という立場から見た時，重要な問題点は，用いられるデータが，集計による歪みを持たないことを前提としてきたことであろう。しかし，本稿のモデルの基本構造やその実際の計算結果は，回帰分析によるものと基本的に大きく異なるものではなく，集計データに歪みがある時，本稿のアプローチも，集計データを利用した回帰分析の持つ問題（いわゆるエコロジカルファラシー）を逃れていないことを示唆しているといってよい。

「二項分布の和モデル」を，実用的な意味で真に価値あるものたらしめるには，集計データに歪みが存在する時にも，適切な推計を行えるように「改良」を加える必要がある。筆者らは，このモデルを基礎とした改良が可能であると考えており，(注5)で述べたように，他稿にてその点を展開する予定である。

[付記] 本稿作成には，コンピューターの使用が重要な役割を果たしたが，その際，El Colegio de México の Coordinación de Servicios de Computo の次の方々にお世話になった。記して感謝する。Ricardo Ramírez Corona, Gerardo Julian Naranjo Villares, Laura Esther Cienfuegos Ambriz, Oscar Sánchez de La Barquera Jorge Rolando Rodríguez Ariano, Mario Alberto Martínez Yánez.

（注1） ここで示される分析方法は社会科学的分析において汎用性の高いものと考えられるが，管見するところ適用例は存在しない。

（注2） $y_1+y'_1 \neq x_1+x'_1$の場合もしばしばあろう。この時，適当なパラメーターを与えることによって，等式が成立するよう，データを変形する。第II節で，メキシコシティの事例によってその方法を示す。

（注3） すなわち，xは，ある収入特性に属するということのベルヌーイ試行の回数を示しており，したがって「原因」に対応しており，Zはその時の教育機会享受の回数を表わしており，したがって「結果」に対応している。x'とZ'の関係も同様である（それぞれZ，Z'を従属変数，x，x'を独立変数とする回帰モデルに対比させて考えるとわかりやすい）。故に，$Y(=Z+Z')$は，「結果」に対応している。

このような理由によって，データ値y，y'は，それぞれ確率変数Y，Y'の実現値として扱われ，他方，データ値x，x'は，（それらが確率変数の実現値であるとしても）それらとY，Y'の関係を扱う限りにおいては，確率変数でない通常の変数の値として扱ってよいこととなる（ただしそのためには，厳密には，データ値x，x'は測定誤差や計算誤差等を全く含んではならない）。

もし，このような因果関係の想定が許されないならば，表の縦横を対称に扱う方法が自然であろう。詳論は避けるが，この時，「多項分布モデル」を適用することができる。

（注4） 例えば，生態学的相関と個体相関（集計データに基づく分析と集計する以前の個別データに基づく分析）の相違の問題は有名である。

（注5） したがって，本来，実際の集計データがこの前提を満たしているかどうかについてもチェックする必要があるが，本稿ではそうした検討に踏み込むことは避ける。本稿の方法は，「歪み」の有無の検討，さらに「歪み」がある場合の推計も，一定の条件の下で可能とするものであるが，こうした方向での展開については，別稿を予定している。

（注6） ここでは，回帰モデルと異なり，誤差項を持たないことに注意。すなわち，このモデルの下では，

Yの持つ不確定性は，ただZ, Z'が二項分布にしたがうという不確定性のみに由来しているのである。このことは，このモデルが厳密に適合するためには，xおよびx'ばかりでなく，確率変数Y(およびY')の実現値データy(およびy')も測定誤差や計算誤差を含んでいてはならないことを意味する。

(注7) この方式でも，なお，二項分布が計算の一部に用いられ，その値は時には極めて小さく，0として扱われる。しかし，そのもたらす誤差は十分小さい。

(注8) 坂元慶行・石黒真木夫・北川源四郎 1983.『情報量統計学』共立出版.

(注9) K-L情報量を利用する方法でも，次のように，大数の法則を通じて，最尤推定に用いられるものと同じ項（尤度関数）を持つ式に近似変形できる。しかし，すでに述べた理由から，この近似は行わない。

$$KL=\sum_{y} g(y)\ln\frac{g(y)}{f(y)}=E\left[\ln\frac{g(Y)}{f(Y)}\right]$$
$$=E[\ln g(Y)]-E[\ln f(Y)]$$
$$\fallingdotseq E[\ln g(Y)]-\left(\sum_{j=1}^{n}\ln f(y_j)\right)/n \quad \cdots\cdots(a)$$

ここで，$E[\ \ ]$は期待値，$y_j(j=1,2,3,\cdots,n)$は，データの与える値を意味する。

(注10) 以上では，モデルによる実現確率が0.2に近い値を持つ5つの部分に区間分割を行ったが，この分割方法は随意である。分割をあまり細かく行うと，各分割部分あたりのサンプル数が少なくなって，実現頻度を真の確率分布とみなすことは不適当になる。他方，分割が大雑把過ぎれば，モデルの分布と真の分布の差異を識別できないこととなる。

また，実際計算において，$f_1=f_2=f_3=f_4=f_5=0.2$とみなして問題がない場合が多いであろう。この近似によって，コンピュータープログラム作成上のこみいった手間を省くことができる。

ところで以上では，区間の設定は，事実上，f_1, $f_2,\cdots,f_5$の値を設定することによってなされ，その後，得られた区間に対応した$g'_1, g'_2,\cdots,g'_5$の実現頻度を得た。逆に，$g'_1, g'_2,\cdots,g'_5$の実現頻度の値を設定することによって区間を定め，それに対応したf_1, $f_2,\cdots,f_5$を計算することは，理論的には考えられる。しかし，実際的ではない。すなわち，yの実現値データが全区間で可能なすべての値をとらずに，とびとびの値を示している時，適切な分割点を見出すのは困難である。

(注11) このことは，個票データが各jごとに集計された時に，歪みが発生しなかった，と言い換えることができる。

(注12) このような近似がもたらす問題点について，簡単に考察しておこう。最適解に等しくないp, p'を固定した時，x_j/x'_jが十分大きい範囲と，十分小さい範囲では，モデルは常に「過剰」推定（「過大」推定，または「過小」推定）をもたらすであろう。ただし，「過大」推定とは，$E[f^*_j(y_j)]<E[f_j(y_j)]$，「過小」推定とは，$E[f^*_j(y_j)]>E[f_j(y_j)]$の時であり，$f^*_j(y_j)$は$p=p^*$, $p'=p'^*$である$f_j(y_j)$を意味する。分けられたクラスが「過大」推定のみの区間によって成立しているか，「過小」推定のみの区間によって成立していれば，(24)式による平均を利用した近似も必ず対応する「過剰」推定をもたらす。そして，その「過剰」の程度は，それを構成する各$f_j(y_j)$の「過剰」の程度の「平均的なもの」となっている。この意味で，(22)式の近似も成立しており，全く問題ない。しかし，「過大」推定の区間と「過小」推定の区間の両者を含むクラスでは，(24)式は「過大」と「過小」を打ち消し合う分布をもたらし，「過剰」の程度は，各$f_j(y_j)$の「過剰」の程度の平均ではなく，極端な場合は，各$f_j(y_j)$の「過剰」の程度のいずれよりも低い程度となる可能性がある。この意味で，(22)式の近似は必ずしも成立しない。

しかし，あるクラスdに「過大」と「過小」の区間が含まれていたとしても，したがって，この意味で(22)式の近似が成立しないとしても，$p\neq p^*$, $p'\neq p'^*$の時，常に，

$$KL[\underline{g_{di}}, \underline{f_{di}}\,|\,p, p']$$
$$>KL[\underline{g_{di}}, \underline{f_{di}}\,|\,p^*, p'^*] \quad \cdots\cdots(b)$$

が成立しているなら，推計には全く問題が生じない。さらに，(b)式が成立しなくとも，常に，

$$\sum_c w_c\cdot KL[\underline{g_{ci}}, \underline{f_{ci}}\,|\,p, p']$$
$$>\sum_c w_c\cdot KL[\underline{g_{ci}}, \underline{f_{ci}}\,|\,p^*, p'^*] \quad \cdots\cdots(c)$$

が成立しているなら，推計に問題は生じない。しかし，この時，クラスdは，不適切なp, p'をリジェクトするために貢献していないので，むしろこのクラスのデータは推計に用いない方がよいこととなる。しかし，「過剰」推計をもたらす（あるいはもたらさない）x/x'の範囲は，p, p'およびp^*, p'^*の値に依存するから，効率的な推計のために，データの「理想的な」クラス分け方法や，推計に貢献しない一部のデータの排除方

法をあらかじめ固定的に示すことはできない。

少しでも確実に，推計に貢献するデータの比率を増やす方法としては，$p \neq p^*$，$p' \neq p'^*$の時，「過大」と「過小」推計をもたらす区間の混在するクラスの中のデータ数をできる限り減らすことが考えられよう。このためには，クラス分けをできる限り細かく行い，1クラス中のデータ数を減らすことが好ましい。しかし，他方，以下で述べるように，1クラス中にある程度以上のデータがなければ，頻度計算が行えず，推計が不可能となる。

しかし，特殊な条件を設定しない限り，(c)式は成立するであろうから，本注で問題にしたような意味の限りでは，クラス分けをそれほど細かくすることに気をつかう必要はないであろう。

(注13) 実際には，任意の j，i に対し，$f_{ji}=0.2$と計算しても，生ずる誤差は無視できる程度のものである。

(注14) 証明はここでは省略するが，通常の大数の法則の場合の証明を拡張すればよい。

(注15) x/x' が様々な値を持つ形でデータが与えられていることは，例えば同じ数のデータが，x/x' が一つの値に固定されている場合に比べ，推計をより「安定的」なものにする。x/x' が一つの値に固定されている場合，K-L情報量は，$(p, p')=(p^*, p'^*)$で最小値をとり，そこから，(p, p')を㉘式の直線を満たして動かす時，それ以外の方向に動かすより，ゆっくりとK-L情報量は増加する((注12)を参照)。この意味で，最適解 (p^*, p'^*) は，㉘式の直線方向に関しては，より「安定度の低い」ものとなっている。これに対し，x/x' が様々な値を持つ場合，これらすべての値のデータに対して，同時にK-L情報量をゆっくりと増加させるような1つの直線は存在しない。この意味で，最適解 (p^*, p'^*) は，より「安定的」な解となっている。

他方，「二項分布の和モデル」の重要な特徴は，x，x' を固定した状態でのデータがたくさんあれば（正確な推計が様々な値の x，x' を持つ場合に比較すれば困難であるにせよ），p^*，p'^*の推計が可能になることを挙げることができる。第II節で見る回帰分析による方法では，データがいくらあっても，x，x' が固定されていては推計は不可能なのである。

ところで，ここでのクラス分けという作業は，改めて，集計という操作を行っていることになる。最初の集計データに歪みがなければ（本稿ではそれを前提とした），それをさらに集計した，このクラス分けしたデータにも歪みがないと考えてよい。

(注16) ㉙，㉚式は，Z，Z' が互いに独立であることを前提に成立していること，したがってまた，㉘式もそうした前提のもとに成立していることに注意。

(注17) Instituto Nacional de Estadística, Geografía e Información 1992. *CODICE 90, Resultados definitivos, XI censo general de población y vivienda, 1990.* Aguascalientes. CD-ROM版.

(注18) 厳密にいえば，格子探索法によるアプローチが正しいためには，K-L情報量を表す式の数学的性質が検討されていなくてはならない。本稿では，それを欠いたまま，K-L情報量がこのアプローチによって誤りへと導かれるような不規則性を有していないことを前提としている。

(注19) 計算作業はパーソナルコンピュータを用い，SPSSプログラム (version 6) の上で，かなり複雑なプログラムを作成した。プログラムが複雑になる一つの理由は，モデルの個数が204通りに上るため（それぞれが異なったデータセットに対応している），マクロを用いる構造にしていることがある。格子探索法による推定で，1つのモデルに対応する1つの p^* を推定するのに1～2時間かかっている。

(注20) 本稿では近似計算に関わる便宜的な理由を述べたに止まるが，より積極的な意味を論ずることができる。

（米村・アジア経済研究所研究コーディネーター／野田・同開発研究部主任研究員）