大阪大学人間科学部 〔1990年 3 月〕
『年報人間科学』第11号 1 頁—15頁

クロス表のカテゴリー統合

佐 藤 裕

クロス表のカテゴリー統合

1．本稿の目的

我々が実際に調査データなどを分析する際に、クロス表を構成する変数のカテゴリーを統合し、より小さなクロス表に加工したいという欲求を持つ場合がある。これは、主に次のような理由によるものである。

まず第一に、全ケース数が少なすぎたり、カテゴリーが多すぎたりして１つのセルの度数が少なくなりすぎる場合である。クロス表の分析でよく用いられるカイ２乗独立性検定はセルの期待度数があまり小さいと適用できないため、カテゴリーの統合によってこれを回避するのである。

第二に、カテゴリーが多すぎてクロス表が見にくく、その表が表わしていることが読みとりにくい場合である。これは特に順序尺度以上の変数をクロス表にするときに行なわれる。例えば年齢は多くのデータでは実数で与えられているが、そのままでは、あまりに大きな表になるため、例えば10歳ごとのカテゴリーにまとめて表を作る。これもまたカテゴリーの統合である。

それでは、このような場合にはどのような基準でカテゴリー統合が行なわれているだろうか。実際には統合したい変数の度数分布を見て、なるべくそれぞれのカテゴリーの度数が均等になるようにしたり、縦または横のパーセント（条件付き確率）を見て、それが似ているカテゴリーを統合したり、あるいは先の年齢のように等間隔に区切ってカテゴリーを構成したりしている。

しかし、そのような方法でのカテゴリー統合は本当にデータの特徴をよく表わすことになるだろうか。カテゴリーの統合は元のクロス表の持っていた情報を確実に減少させるが、その失われた情報の中に本当に重要な情報はなかったと言えるのだろうか。これらはカテゴリー統合の統計的妥当性の問題である。

カテゴリーを統合する際の１つの基準は、この統合によって失われる情報量の評価から与えられるだろう。統合の度合に較べて情報量の損失が少ないとき、そのクロス表が持つ情報をより凝縮したような新しいクロス表が得られたと言えるのである。

また、こういった基準を明確にし、いくつかの統合の仕方を比較し、最も適切な統合パターンを選択することができるならば、クロス表の構造を明確にすることもできるだろう。例えば、年齢と何か別の変数のクロス表の場合、年齢の最適なカテゴリー化の選択は、どの年齢で大きな変化があるか、ということを見いだすことにもなるだろう。

このような観点からのクロス表のカテゴリー統合については、すでに坂元慶行氏の業績がある(1)。坂元氏はAIC（赤池情報量基準）を用いてカテゴリー統合を行ったモデルの評価を行い、また最適な統合法を選択するコンピュータプログラムも開発している。

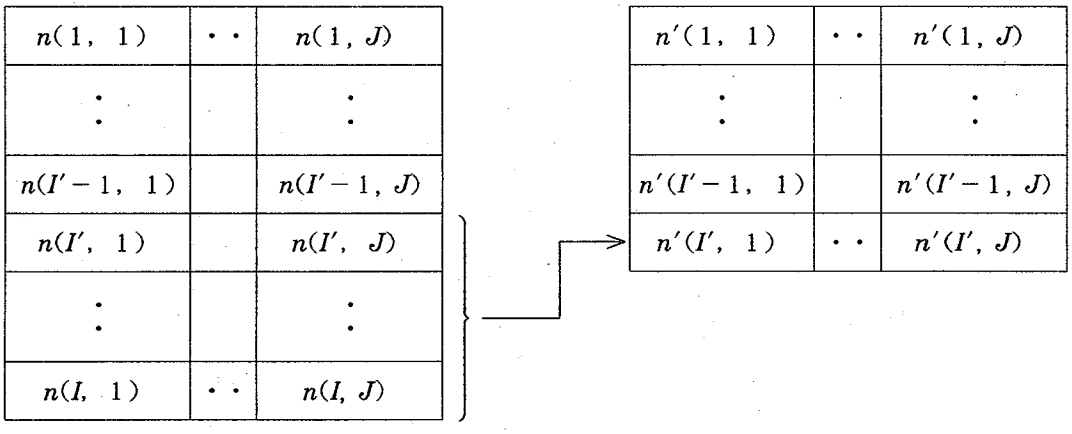
本稿は、坂元氏の方法をもとにして、実際のデータ分析を行う際により容易に利用できる方法と、坂元氏の方法を2変数のカテゴリーを同時に統合した場合へと拡張するモデルを提案するものである。

2．1変数の統合

1）モデルの設定

$I \times J$表のI'行からI行までを1つのカテゴリーとしてまとめ、新たに$I' \times J$表を作ることを考えてみよう（図1）。まとめる箇所が隣接しているのは表記の便宜のためであり、カテゴリーの順序は以後の議論に無関係であるので、離れたカテゴリーの統合も同様に扱える。

図1



カテゴリーを1つにまとめることができるための条件は、まとめるそれぞれのカテゴリーで、列の変数についての条件付き確率が等しい、すなわち、

$$p(j \mid i1)=p(j \mid i2) \quad (j=1,\ 2\ .\ .\ J,\ I' \leqq i\ 1,\ i\ 2 \leqq I) \qquad (2.\ 1)$$

であるとする。例えば行を年齢、列をある意見とするクロス表の場合、意見の回答分布が等しい年齢カテゴリーは統合できる、と考えるのである。これはごく自然な条件であると思われる。この条件付き確率はセルの度数を用いて表わすと、

$$p(j|i)=\frac{n(i,\ j)}{n(i,\ .)}$$

であるから、式（2．1）は、

$$\frac{n(i1,\ j)}{n(i1,\ .)}=\frac{n(i2,\ j)}{n(i2,\ .)} \quad (j=1,\ 2\ .\ .\ J,\ I' \leqq i1,\ i2 \leqq I)$$

と書ける。したがって、

$$\frac{n(i1,j)}{n(i2,j)}=\frac{n(i1,.)}{n(i2,.)} \quad (j=1,2\ldots J,\ I'\leqq i1,\ i2\leqq I) \qquad (2.\ 2)$$

ところが、式（2．2）の右辺はjに無関係であるので、左辺はすべての j について等しい。すなわち、

$$\frac{n(i1,j1)}{n(i2,j1)}=\frac{n(i1,j2)}{n(i2,j2)} \quad (j1,j2=1,2,\ldots J,\ I'\leqq i1,\ i2\leqq I) \qquad (2.\ 3)$$

である。また逆に、$n(i,.)$ は$n(i,j)$ をすべてのjについて加えたものだから、式（2，3）が成り立つならば式（2，2）も成り立つ。したがって式（2，3）は式（2．1）の必要充分条件である。式（2，3）は、統合してできたセルの度数を、各列に共通な比によって配分したものが、もとのセルの度数と等しい、ということを意味する。

カテゴリー統合の妥当性は、統合してできたクロス表から、式（2．3）の条件によってもとのクロス表を推定したときの適合度で評価できる。このときの推定モデルは以下のように表わすことができる。

もとのクロス表のセルの期待度数を$m(i,j)$とすると、

$$m(i,j)=\begin{cases}\dfrac{a(i,j)}{n} & (i<I')\\ \dfrac{a(I',j)\cdot b(i)}{n} & (i\geqq I')\end{cases} \qquad (2.\ 4)$$

である。

ここで、$a(i,j)$ は統合してできたクロス表の i 行 j 列のセルの確率、$b(i)$ は統合したセルをもとのセルに配分する比率である。ただし、

$$\sum_{i=1}^{I'}\sum_{j=1}^{J}a(i,j)=1,\quad \sum_{i=I'}^{I}b(i)=1 \qquad (2.\ 5)$$

である（2）。

2）最尤推定値

モデルの適合度の評価のためには、まずそのモデルのパラメータと期待度数の最尤推定値を求めなくてはならない。

パラメータ $\{a(i,j),\ b(i)\}$ において、実現度数 $\{n(i,j)\}$ が見られる確率は、

$$M(\{n(i,j)\}\,|\,\{a(i,j),b(i)\})$$
$$=\frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J}n(i,j)!}\prod_{i=1}^{I'-1}\prod_{j=1}^{J}a(I',j)^{n(i,j)}\prod_{i=I'}^{I}\prod_{j=1}^{J}a(I',j)\,b(i)^{n(i,j)}$$

したがって、$\{a(i,j),\ b(i)\}$ についての対数尤度は、パラメータに無関係な定数項を無視すると、

$$l(\{a(i,j),\ b(i)\})$$
$$=\sum_{i=1}^{I'-1}\sum_{j=1}^{J}n(i,j)\log a(i,j)+\sum_{i=I'}^{I}\sum_{j=1}^{J}n(i,j)\log a(I',j)+\sum_{i=I'}^{I}\sum_{j=1}^{J}n(i,j)\log b(i)$$

$$=\sum_{j=1}^{J}\left(\sum_{i=1}^{I'-1}n(i,j)\log a(i,j)+\sum_{i=I'}^{I}n(i,j)\log a(I',j)\right)+\sum_{i=I'}^{I}\sum_{j=1}^{J}n(i,j)\log b(i) \qquad (2.\ 6)$$

ここで、

$$n'(i,j)=\begin{cases} n(i,j) & (i<I') \\ \sum_{i=I'}^{I}n(i,j) & (i=I') \end{cases}$$

とすると、式（2．6）の右辺の第1項は、

$$\sum_{i=1}^{I'}\sum_{j=1}^{J}n'(i,j)\log a(i,j)$$

となる。

式（2．6）を最大化する $a(i,j)$ を求めるためには、式（2．5）の制約より

$$a(I',j)=1-\sum_{i=1}^{I'-1}a(i,j)$$

を式（2．6）に代入して $a(i,j)$ で偏微分し、これを0にするような $a(i,j)$ を求めればよい(3)。その結果、$a(i,j)$ の最尤推定値 $\hat{a}(i,j)$ は、

$$\hat{a}(i,j)=\frac{n'(i,j)}{n}$$
$$=\begin{cases} \dfrac{n(i,j)}{n} & (i<I') \\ \dfrac{\sum_{i=I'}^{I}n(i,j)}{n} & (i=I') \end{cases}$$

である。

また同様に、$b(i)$ の最尤推定値 $\hat{b}(i)$ は、

$$\hat{b}(i)=\frac{n(i,\ .)}{\sum_{i=I'}^{I}n(i,\ .)}$$

よって、期待度数 $m(i,j)$ の最尤推定値は

$$\hat{m}(i,j)=\begin{cases} n(i,j) & (I<I') \\ \dfrac{\sum_{i=I'}^{I}n(i,j)\cdot n(i,\ .)}{\sum_{i=I'}^{I}n(i,\ .)} & (i\geqq I') \end{cases}$$

カテゴリーを統合し、そのときの条件確率が等しいというモデルでの $m(i, j)$ の最尤推定値は、統合しないセルについては実現度数と等しく、統合したセルは統合してできたセルの度数を行の周辺度数によって比例配分したものである。

図２ａの２行目と３行目を統合するというモデルを考えると、その時の期待度数は図２ｂのようになる。例えば、２行１列のセルの期待度数は、（100＋０）×100／（100＋400）＝20である。

図２ａ

400	50	50	500
0	50	50	100
100	100	200	400
500	200	300	1000

図２ｂ

400	50	50	500
20	30	50	100
80	120	200	400
500	200	300	1000

３）モデルの検定

統合の妥当性を検定するためには、尤度比統計量 L^2 を使うことができる。L^2 は一般に以下の式で与えられる。

$$L^2=2\sum_{i=1}^{I}\sum_{j=1}^{J}n(i, j)\log\frac{n(i, j)}{m(i, j)}$$

ところが、$i<I'$ の範囲では、$n(i, j)=m(i, j)$ であり、Σの右はすべて０になるので、カテゴリー統合の妥当性を検定する $L^2{}_1$ は次の式で求められる。

$$L^2{}_1=2\sum_{i=I'}^{I}\sum_{j=1}^{J}n(i, j)\log\frac{n(i, j)\cdot\sum_{i=I'}^{I}n(i, .)}{\sum_{i=I'}^{I}n(i, j)\cdot n(i, .)}$$

この式は、統合する部分を１つのクロス表と見なし、独立性の検定を行なった場合の式と同じである。

自由度を求めるにはセルの数から（パラメータ数＋１）を引けばよい。$a(i, j)$ は式（２．５）の制約より、$I'J-1$ 個、$b(i)$ も同じく $I-I'$ 個であるから、自由度 df_1 は、

$$\begin{aligned}df_1&=IJ-(I'J-1+I-I'+1)\\&=(I-I')(J-1)\end{aligned}$$

$I-I'$ は結合する部分の行数－１と等しいので、自由度もまた結合する部分を１つのクロス表と見なして独立性の検定を行なった場合と等しい。

以上のことから、カテゴリー統合の妥当性を検定するには、統合する部分を１つのクロス表と見なして独立性の検定を行なえばよいということがわかる。

また、統合してできたクロス表の独立モデルの尤度比統計量をL^2_2とすると、

$$L^2_2=2\sum_{j=1}^{J}\left(\sum_{i=1}^{I'-1}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}+\sum_{i=I'}^{I}n(i,j)\cdot\log\frac{n\cdot\sum_{i=I'}^{I}n(i,j)}{\sum_{i=I'}^{I}n(i,\ .)n(.\ ,j)}\right)$$

これに、先ほどのL^2_1を加えると、

$$L^2_1+L^2_2=2\sum_{j=1}^{J}\left\{\sum_{i=1}^{I'-1}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}+\sum_{i=I'}^{I}n(i,j)\log\left(\frac{n\cdot\sum_{i=I'}^{I}n(i,j)}{\sum_{i=I'}^{I}n(i,\ .)n(.\ ,j)}\cdot\frac{n(i,j)\cdot\sum_{i=I'}^{I}n(i,\ .)}{\sum_{i=I'}^{I}n(i,j)\cdot n(i,\ .)}\right)\right\}$$

$$=2\sum_{j=1}^{J}\left(\sum_{i=1}^{I'-1}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}+\sum_{i=I'}^{I}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}\right)$$

$$=2\sum_{i=1}^{I}\sum_{j=1}^{J}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}$$

となり、これはもとのクロス表の独立モデルの尤度比統計量と等しい。すなわち、もとのクロス表の独立性の尤度比統計量をL^2_3とすると、

$$L^2_3=L^2_1+L^2_2 \qquad (2.\ 7)$$

という式が成り立っているのである。また、自由度についても、

$$\begin{aligned}df_1+df_2&=(I-I')(J-1)+(I'-1)(J-1)\\&=(I-1)(J-1)\\&=df_3\end{aligned} \qquad (2.\ 8)$$

と、同様の関係が成り立つ。

すなわち、カテゴリー統合の妥当性を検定するには、もとのクロス表のL^2と統合してできたクロス表のL^2の差を自由度の差で検定すればよい、ということである。

これまでの議論は1ヶ所のみの統合のみを扱っていたが、2ヶ所以上を統合する場合もこれをそのまま拡張することができる。統合の妥当性のL^2は統合したそれぞれの箇所を1つのクロス表と見なした独立性のL^2をすべて足し合わしたものになり、また、式（2．7）、（2．8）も1ヶ所のみの統合と同様成り立っている。

3．2変数の統合

(1) モデル設定と最尤推定値

次に、行と列の両方を統合する場合について考えてみよう。$I\times J$表のI'行目からI行目とJ'列目からJ列目を統合するとすると、統合してできたクロス表からもとのクロス表を推定するモデルは、1変数の場合と同様にもとのクロス表のセルの期待度数を$m(i,\ j)$とすると、

$$m(i,j)=\begin{cases}\dfrac{a(i,j)}{n} & (i<I',\ j<J')\\ \dfrac{a(I',j)\cdot b(i)}{n} & (i\geqq I',\ j<J')\\ \dfrac{a(i,J')\cdot c(j)}{n} & (i<I',\ j\geqq J')\\ \dfrac{a(I',J')\cdot b(i)\cdot c(j)}{n} & (i\geqq I',\ j\geqq J')\end{cases}$$

ここで、$a\ (i,\ j)$ は統合してできたクロス表の i 行 j 列のセルの確率、$b(i)$、$c(j)$ は統合したセルをもとのセルに配分する比率である。ただし、

$$\sum_{i=1}^{I'}\sum_{j=1}^{J'}a(i,j)=1,\quad \sum_{i=I'}^{I}b(i)=1,\quad \sum_{j=J'}^{J}c(j)=1 \qquad (3.\ 1)$$

$\{a\ (i,\ j)$、$b(i)$、$c(j)\}$ についての対数尤度は、パラメータに無関係な定数項を無視すると、

$$\begin{aligned}&l(\{a(i,j),\ b(i),\ c(j)\})\\ &=\sum_{i=1}^{I'-1}\sum_{j=1}^{J'-1}n(i,j)\log a(i,j)\\ &+\sum_{i=I'}^{I}\sum_{j=1}^{J'-1}n(i,j)\log a(I',j)+\sum_{i=I'}^{I}\sum_{j=1}^{J'-1}n(i,j)\log b(i)\\ &+\sum_{i=1}^{I'-1}\sum_{j=j'}^{J}n(i,j)\log a(i,J')+\sum_{i=1}^{I'-1}\sum_{j=j'}^{J}n(i,j)\log c(j)\\ &+\sum_{i=I'}^{I}\sum_{j=j'}^{J}n(i,j)\log a(I',J')+\sum_{i=I'}^{I}\sum_{j=j'}^{J}n(i,j)\log b(i)\\ &+\sum_{i=I'}^{I}\sum_{j=J'}^{J}n(i,j)\log c(j)\end{aligned} \qquad (3.\ 2)$$

ここで、

$$n'(i,j)=\begin{cases}n(i,j) & (i<I',\ j<J')\\ \displaystyle\sum_{i=I'}^{I}n(i,j) & (i=I',\ j<J')\\ \displaystyle\sum_{j=J'}^{J}n(i,j) & (i<I',\ j=J')\\ \displaystyle\sum_{i=1}^{I}\sum_{j=J'}^{J}n(i,j) & (i=I',\ j=J')\end{cases}$$

とし、式（3．2）から $a(i,\ j)$ を含む項のみを抜き出すと、

$$\sum_{i=1}^{I'}\sum_{j=1}^{J'}n'(i,j)\log a(i,j)$$

となるので、1変数の場合と同様に、（3．1）の制約式を代入し、それぞれのパラメータで偏微分して $a(i,\ j)$ の最尤推定値を求めると、

$$\hat{a}(i,j)=\frac{n'(i,j)}{n}$$

また、$b(i)$ を含む項は、

$$\sum_{i=I'}^{I}\sum_{j=1}^{J}n(i,j)\log b(i)$$

と整理できるので、（3．1）の制約式から $b(i)$ の最尤推定値は

$$\hat{b}(i)=\frac{n(i,\ .)}{\sum_{i=I'}^{I}n(i,\ .)}$$

$c(j)$ の最尤推定値も同様に、

$$\hat{c}(j)=\frac{n(.\ ,j)}{\sum_{j=J'}^{J}n(.\ ,j)}$$

となる。

したがって、$m(i,\ j)$ の最尤推定値は、

$$\hat{m}(i,j)=\begin{cases} n(i,j) & (I<I',\ j<J') \\ \dfrac{\sum_{i=I'}^{I}n(i,j)\cdot n(i,\ .)}{\sum_{i=I'}^{I}n(i,\ .)} & (i\geqq I',\ j<J') \\ \dfrac{\sum_{j=J'}^{J}n(i,j)\cdot n(.\ ,j)}{\sum_{j=J'}^{J}n(.\ ,j)} & (i<I',\ j\geqq J') \\ \dfrac{\sum_{i=I'}^{I}\sum_{j=J'}^{J}n(i,j)\cdot n(i,\ .)\cdot n(.\ ,j)}{\sum_{i=I'}^{I}n(i,\ .)\cdot\sum_{j=J'}^{J}n(.\ ,j)} & (i\geqq I',\ j\geqq J') \end{cases}$$

図2 aのクロス表の2行目と3行目、2列目と3列目をそれぞれ統合したときの期待度数は図3のようになる。例えば2行3列のセルの場合であれば、(50+50+100+200) ×100×200／100+400) ／ (200+300)=32である。

図3

400	40	60	500
20	32	48	100
80	128	192	400
500	200	300	1000

(2) モデルの検定

統合の妥当性を検定する L^2_1 は、やはり行も列も統合しない範囲は無視できるので、

$$L^2_1=2\sum_{i=I'}^{I}\sum_{j=1}^{J'-1}n(i,j)\log\frac{n(i,j)\cdot\sum_{i=I'}^{I}n(i,\ .)}{\sum_{i=I'}^{I}n(i,j)\cdot n(i,\ .)}$$

$$+2\sum_{i=1}^{I'-1}\sum_{j=J'}^{J}n(i,j)\log\frac{n(i,j)\cdot\sum_{j=J'}^{J}n(.\ ,j)}{\sum_{j=J'}^{J}n(i,j)\cdot n(.\ ,j)}$$

$$+2\sum_{i=I'}^{I}\sum_{j=J'}^{J}n(i,j)\log\frac{n(i,j)\cdot\sum_{i=I'}^{I}n(i,\ .)\cdot\sum_{j=J'}^{J}n(.\ ,j)}{\sum_{i=I'}^{I}\sum_{j=J'}^{J}n(i,j)\cdot n(i,\ .)\cdot n(.\ ,j)}$$

これに、統合してできたクロス表の独立性 L^2

$$L^2_2=2\sum_{j=1}^{J'-1}\sum_{i=1}^{I'-1}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}$$

$$+2\sum_{i=I'}^{I}\sum_{j=1}^{J'-1}n(i,j)\log\frac{n\cdot\sum_{i=I'}^{I}n(i,j)}{\sum_{i=I'}^{I}n(i,\ .)\cdot n(.\ ,j)}$$

$$+2\sum_{i=1}^{I'-1}\sum_{j=J'}^{J}n(i,j)\log\frac{n\cdot\sum_{j=J'}^{J}n(i,j)}{\sum_{j=J'}^{J}n(.\ ,j)\cdot n(i,\ .)}$$

$$+2\sum_{i=I'}^{I}\sum_{j=J'}^{J}n(i,j)\log\frac{n\cdot\sum_{i=I'}^{I}\sum_{j=J'}^{J}n(i,j)}{\sum_{i=I'}^{I}n(i,\ .)\cdot\sum_{j=J'}^{J}n(.\ ,j)}$$

を加えると、

$$L^2_1+L^2_2=2\sum_{i=1}^{I}\sum_{j=1}^{J}n(i,j)\log\frac{n\cdot n(i,j)}{n(i,\ .)n(.\ ,j)}$$

となり、やはりもとのクロス表の独立性 L^2 と等しい。

自由度は、

$$df_1=IJ-(I'J'-1+I-I'+J-J'+1)$$
$$=IJ-I'J'-I+I'-J+J'$$
$$df_2=(I'-1)(J'-1)$$
$$=I'J'-I'-J'+1$$

$$df_3=(I-1)(J-1)$$
$$=IJ-I-J+1$$

であるので、やはり

$df_1 + df_2 = df_3$

が成り立っている。

2変数の統合の場合は、このように、もとのクロス表と統合してできたクロス表の独立性L^2を比較することによって間接的に妥当性を検定する方法が便利である。

ここまでで紹介した方法は、特別な計算プログラムを必要とせず、SPSSやSASなどの汎用統計パッケージを利用してだれでもすぐに計算することができる。特に、統合する前後のクロス表の独立性L^2を比較する方法は、1変数を統合する場合も両方の変数を統合する場合も利用でき、しかも汎用統計パッケージの出力を比較してカイ2乗分布表を調べるだけなので、非常に手軽で有効な方法である。

4．統合法の選択

カテゴリーの統合を行なう場合に、いくつかの方法が考えられる場合がある。例えば年齢をカテゴリー化する場合に、10歳ごとに区切るのか、あるいはなるべくそれぞれのカテゴリーに含まれる人数が等しくなるようにするのか、といった場合である。このようないくつかの統合法から1つを選択するといった場合にも、今まで述べた方法が適用できる。

複数の統合法を比較するための基準としては、AIC（赤池情報量基準）が便利である。AICは次の式で求めることができる(4)。

$AIC = L^2 - 2df$

L^2もdfも今まで述べた方法で計算できるので、AICの計算も簡単である。こうしてすべての統合法について計算し、AICが最も小さい統合法が、もとのクロス表の持つ情報を最も凝縮して表わした最適なクロス表である。

クロス表の統合法は、すべての可能性を含めると、比較的小さなクロス表でも膨大な数になる。例えば5カテゴリーを統合する組合せは51通りだが、これが6カテゴリーになると202通りへと急増する。さらに行と列の両方を統合する組合せを考えると単純に計算してこれの2乗、5×5だと2601通りにのぼる。このような膨大な計算を行なうには当然コンピュータの利用が考えられる。コンピュータを用いればこの程度の計算なら楽々とこなしてくれるので機械的に最適なクロス表を選択するには有効である。

しかし、もっと大きなクロス表の場合は、組合せがあまりにも多くなりすぎ、またクロス表を構成する変数の性質や分析の目的によってはすべての組合せを計算する必要がない場合もある。例えば変数が順序尺度を構成していると考える場合は隣合うカテゴリー以外の統合は無意味であり、これを除くと、考えられる統合の組合せは大幅に減少する(5)。

従来の独立性の検定は、行または列のすべてのカテゴリーを1つにまとめた場合の妥当性の検討と等しい。様々な統合法はもとのクロス表と、すべてを1カテゴリーにまとめた場合との

中間に位置づけられる。コンピュータの利用によって最適なクロス表を選択することは、2つの変数が独立であるかそうでないか、という両極端の中間にいくつものレベルを設定し、どのレベルが最も適当かということを求めることであるとも言える。

5．例題

最後に、前節で述べたような最適なクロス表の選択法を、実際のデータに適用して考えてみよう。表4は85年SSM調査報告書からのデータである(6)。これは40歳以上の既婚女性のキャリアタイプ別の年齢を表わしている。この表をそのまま見ると結婚退職や中年退職は年齢がやや低く、就業継続は年齢が高いように見えるが、それほどはっきりしたものではない。そこでまずキャリアタイプによる年齢構成の違いをはっきりさせるために、キャリアタイプのみの統合を行ってみると、AICが最も小さい統合法は表5のようになった。これを見ると、就業継続と不就業が類似した年齢構成を持ち、中年退職、再就職、結婚退職の3つも年齢構成が類似している。前者は比較的高い年齢が多く、後者は若い年齢層に多い。逆に、年齢によるキャリアタイプの違いに着目して年齢カテゴリーの統合を行なったのが表6である。この場合、年齢は順序尺度を構成すると考えて隣合うカテゴリーのみを統合した。表5と同様に、就業継続と不就業は高い年齢層ほど比率が多くなっており、中年退職は高い年齢層ほど比率が低くなっているが、再就職は年齢による差がほとんどなく、結婚退職も高い年齢層ほど比率が低くなってはいるものの、それほど大きな差ではない。最後に年齢、キャリアタイプ両方を統合したのが表7である。このときの最適な統合法は、それぞれの変数を別々に統合した場合と必ずしも一致しないことが注目される。この表から読み取れることは、就業継続が最も高い年齢層に集中し、ついで不就業も高い年齢層に偏っている。最も若い年齢層に集中しているのは中年退職であり、再就職と結婚退職はやや若い年齢層に偏っている。

表4と表7を見比べると、表4では細かい数値に紛れて読み取りにくかった傾向が、表4では非常にはっきりと現われてきている。このように、最適な統合法を求めることは複雑な構造を持つクロス表を解読する場合にも有効な方法であると言えよう。

表4

実数／比率（%）		年齢 40～44歳	45～49歳	50～54歳	55～59歳	60～64歳	65～69歳	合　計
キャリア	就業継続	5	9	19	9	11	21	74
タイプ		(6.8)	(12.2)	(25.7)	(12.2)	(14.9)	(28.4)	
	中年退職	10	17	11	10	2	3	53
		(18.9.)	(32.1)	(20.8)	(18.9)	(3.8)	(5.7)	
	再就職	14	27	21	13	12	11	98
		(14.3)	(27.6)	(21.4)	(13.3)	(12.2)	(11.2)	
	結婚退職	78	83	70	66	34	36	367
		(21.3)	(22.6)	(19.1)	(18.0)	(9.3)	(9.8)	
	不就業	16	17	22	21	16	18	110
		(14.5)	(15.5)	(20.0)	(19.1)	(14.5)	(16.4)	
	合　計	123	153	143	119	75	89	702
		(17.5)	(21.8)	(20.4)	(17.0)	(10.7)	(12.7)	

表5

実数／比率（%）		年齢 40～44歳	45～49歳	50～54歳	55～59歳	60～64歳	65～69歳	合　計
キャリア	就業継続	21	26	41	30	27	39	184
タイプ	不就業	(11.4)	(14.1)	(22.3)	(16.3)	(14.7)	(21.2)	
	中年退職	102	127	102	89	48	50	518
	再就職	(19.7)	(24.5)	(19.7)	(17.2)	(9.3)	(9.7)	
	結婚退職							
	合　計	123	153	143	119	75	89	702
		(17.5)	(21.8)	(20.4)	(17.0)	(10.7)	(12.7)	

AIC=−12,329

表6

実数／比率（%）		キャリアタイプ 就職継続	中年退職	再就職	結婚退職	不就業	合　計
年齢	40～49歳	14	27	41	161	33	276
		(5.1)	(9.8)	(14.9)	(58.3)	(12.0)	
	50～59歳	28	21	34	136	43	262
		(10.7)	(8.0)	(13.0)	(51.9)	(16.4)	
	60～69歳	32	5	23	70	34	164
		(19.5)	(3.0)	(14.0)	(42.7)	(20.7)	
	合　計	74	53	98	367	110	702
		(10.5)	(7.5)	(14.0)	(52.3)	(15.7)	

AIC=−13,954

表7

実数／比率（%）		年齢 40～49歳	50～59歳	60～69歳	合　計
キャリヤタイプ	就業継続	14	28	32	74
		(18.9)	(37.8)	(43.2)	
	中年退職	27	21	5	53
		(50.9)	(39.6)	(9.4)	
	再就職結婚退職	202	170	93	465
		(43.4)	(36.6)	(20.0)	
	不就業	33	43	34	110
		(30.0)	(39.1)	(30.9)	
	合　計	276	262	164	702
		(39.3)	(37.3)	(23.4)	

AIC=−17,041

注

（1）坂元慶行、『カテゴリカルデータのモデル分析』、共立出版、1985

（2）坂元慶行氏は同じ条件から、条件付き確率をパラメータとしたモデルをつくり、以下と同様の結論を得ている。本稿のモデルと坂元氏のモデルは同等であるが、後で述べる2変数の統合への拡張を容易にするためにこのモデルを採用した。

坂元慶行、同書、p31

（3）この部分の計算については、以下を参照のこと。

坂元慶行、同書、p 8

（4）正確には、最もパラメータ数の多いモデル（飽和モデル）のAICとの差である。

坂元慶行、同書、p44

（5）上田尚一氏は、最も適当な2つのカテゴリー統合を繰り返して、徐々にカテゴリー数を減らしていく階層的手法を提唱している。分析の目的によってはこれを用いることもできるだろう。一般には階層的手法での解と総当たりでの最適解は一致しない。以下の例題の計算は総当たり法で計算した。

上田尚一、『データ解析の方法』、朝倉書店、1982、p99

（6）岡本英雄、「女性の職業経歴」、『1985年社会階層と社会移動全国調査報告書　第4巻』1985年社会階層と社会移動全国調査委員会、1989、pp61－74