交互作用は相互作用？(4)

職業により初診時癌重症度は違う？

広津　千尋｜Hirotsu　Chihiro

明星大学理工学部教授

■1968年東京大学大学院工学系研究科計数工学専攻博士課程修了。工学博士。東京工業大学助手を経て、1971年東京大学工学部計数工学科講師、助教授、教授。2000年定年退官、東京大学名誉教授。2000年４月より現職。

1. 職業による癌重症度分類

表１は国立癌センターに登院した癌患者の初診時における重症度分布を、職業別に示したものである。このデータは大分前に取られたものであり、当時は各施設で癌の疑いのある患者が発見されると、国立癌センターで精密検査を受けることが多かった。そのため、職業群によって初診時重症度分布に差があるとすると、それは職業群における癌の早期発見システムの差異によるものと考えられた。

さて、この表は10通りの３項分布が観測されていると見ることができ、興味があるのは職業による重症度プロファイルの違いである。すなわち、今回も行比較の問題であり、似通った重症度プロファイルを示す職業群への分類に興味がある。列のカテゴリは前２回と違って時間軸ではないが、軽・中・重症という自然な順序に従っており、それに沿った系統的なプロファイルの違いに興味があるという意味で、数理的な構造は同じと考えられる。ただ前回との大きな違いは、正規分布の平均

表1　職業別に分類した初登院時の癌の病状

職業＼病状	軽症		中症		重症		計
1. 専門的・技術的職業（土木技術者，教員，医師等）	148	123.3 148.5	444	473.9 452.4	86	80.8 77.1	678
2. 管 理 職	111	93.1 112.1	352	357.9 341.6	49	60.0 58.2	512
3. 事務専従者（会計事務局，タイピスト等）	645	524.6 631.4	1,911	2,015.7 1,924.4	328	343.7 328.1	2,884
4. 販売従事者	165	191.9 160.7	771	737.4 764.0	119	125.7 130.3	1,055
5. 農林，漁業，採鉱従事者	383	458.9 384.3	1,829	1,763.4 1,827.2	311	300.6 311.5	2,523
6. 運輸，通信従事者	96	79.3 95.5	293	304.7 290.9	47	52.0 49.6	436
7. 技 能 士（製鉄工，自動車修理工等）	98	88.4 106.4	330	339.7 324.3	58	57.9 55.3	486
8. 生産工程従事者，単純労働者	199	223.4 187.1	874	858.3 889.3	155	146.3 151.6	1,228
9. サービス業	59	52.4 63.1	199	201.3 192.2	30	34.3 32.8	288
10.無　職	262	330.7 276.9	1,320	1,270.7 1,316.6	236	216.6 224.5	1,818
計	2,166		8,323		1,419		11,908

括弧内数値上段は完全独立モデル(9)のあてはめ、下段は後述の自由度１交互作用モデル(10)あてはめの結果である。

ではなく、2項比率の問題としての定式化を必要とすることである。

2. 数学的定式化

いま、第i職業における第j重症度生起確率をp_{ij}で表すと、職業i, i'で重症度プロファイルが等しいという帰無仮説は

$$H_0 : \frac{p_{i1}}{p_{i'1}} = \frac{p_{i2}}{p_{i'2}} = \frac{p_{i3}}{p_{i'3}} \qquad (1)$$

と表される。一方、職業群iの方がi'より重症者が多いという対立仮説は

$$H_1 : \frac{p_{i1}}{p_{i'1}} \leqq \frac{p_{i2}}{p_{i'2}} \leqq \frac{p_{i3}}{p_{i'3}} \qquad (2)$$

で表される。(1)式と(2)式を対数変換してみると、例えば(2)式は、$\mu_{ij}=\log p_{ij}$として

$$H_1 : \mu_{i1} - \mu_{i'1} \leqq \mu_{i2} - \mu_{i'2} \leqq \mu_{i3} - \mu_{i'3}$$

のように表され、丁度第2回に扱った対立仮説H_1と同じ形式になる。かくして確率モデルに正規分布と多項分布の違いはあるものの、以下の基本的なアイディアは第2回と同様である。

さて、第2回においては、対立仮説H_1に対応して差分行列D_b'を導入し、その一般逆行列$(D_b' D_b)^{-1} D_b'$を観測値ベクトル$\boldsymbol{y}$に作用させて累積和統計量を導いている。実はその理論は正規分布、及び2項分布モデルを含めてより一般的に拡張することができ、エフィシェントスコアを仮定された順序に沿って累積すればよいことが導かれる。それを2項分布モデルに適用すると、セル度数をy_{ij} ($i=1,\cdots,a$; $j=1,\cdots,b$)で表して、累積和統計量

$$Y_{il} = \sum_{j=1}^{l} y_{ij} \quad , l = 1,\cdots, b-1$$

を基礎とすればよいことが導かれる（詳しくは参考文献[2,3]参照）。結局第2回の(4)式に対応する行間の2乗距離として累積χ^2統計量

$$S^*(i;i') = y_{\cdot\cdot}\left(\frac{1}{y_{i\cdot}} + \frac{1}{y_{i'\cdot}}\right)^{-1} \sum_{l=1}^{b-1}\left\{\left(\frac{1}{Y_{\cdot l}} + \frac{1}{y_{\cdot\cdot} - Y_{\cdot l}}\right) \times \left(\frac{Y_{il}}{y_{i\cdot}} - \frac{Y_{i'l}}{y_{i'\cdot}}\right)^2\right\} \qquad (3)$$

が導かれる。群間の2乗距離に対応する統計量は、一般性を失うことなく2群を$I_1=(1,\cdots,u)$, I_2 $(u+1,\cdots,u+v)$として、

$$S^*(I_1;I_2) = y_{\cdot\cdot}\left(1\Big/\sum_{i=1}^{u} y_{i\cdot} + 1\Big/\sum_{i=u+1}^{u+v} y_{i'\cdot}\right)^{-1} \sum_{l=1}^{b-1}\left\{\left(\frac{1}{Y_{\cdot l}} + \frac{1}{y_{\cdot\cdot} - Y_{\cdot l}}\right) \times \left(\sum_{i=!}^{u} Y_{il}\Big/\sum_{i=1}^{u} y_{i\cdot} - \sum_{i=u+!}^{u+v} Y_{il}\Big/\sum_{i=u+1}^{u+v} y_{i\cdot}\right)^2\right\} \qquad (4)$$

で与えられる。これまで同様、(4)式はさらに多群間の2乗距離に拡張され、それらすべてが漸近的にWishart行列 $W(C^{*\prime}C^*, a-1)$の最大根の分布で押さえられる。ただし、$C^{*\prime}C^*$は、

$$\lambda_j = \frac{y_{\cdot 1} + \cdots + y_{\cdot j}}{y_{\cdot j+1} + \cdots + y_{\cdot k}}, \quad j = 1,\cdots, k-1$$

を用いて

$$C^{*\prime}C^* = \{\rho_{ij}\}, \quad \rho_{ij} = \rho_{ji} = \sqrt{\lambda_i / \lambda_j} \qquad (1 \leqq i \leqq j \leqq k-1) \qquad (5)$$

のように構成される。その最大根の分布は$C^{*\prime}C^*$の最大固有値を$\gamma_{(1)}$とするとき、

$$\gamma_{(1)} \chi^2_{(1)} (a-1)$$

の分布でよく近似されることも分かっている。しかし、$k=3, 4$の場合には別に数表が作られており[1]、それを利用することができる。なお、正規分布モデルの場合と異なり、この最大根の分布は撹乱母数σ^2を含んでいない。それはこの場合、(3)式や(4)式で定義される累積χ^2統計量が、丁度Pearsonの適合度χ^2と同じように、周辺和で構成される漸近分散で基準化されたχ^2統計量になっているからである。

3. 適用例

表1のデータで$S^*(i\ ;\ i')$ (3)を計算したの

が表２である。ただし、距離の小さいもの同士が近くに、大きいもの同士はなるべく離れるように行を並べ替えてあることに注意する。

一見してI_1＝(10, 5, 4, 8)、I_2＝(7, 9, 2, 1, 6, 3)の２群に分離しているので、(4)式を計算すると

$$S^*(I_1;I_2)=90.96 \qquad (6)$$

が得られる。この値は表１においてI_1, I_2に含まれる行をそれぞれプールして得られる２×３表に対して計算される累積χ^2統計量(3)の値に等しい。

参考文献[1]の表の補間から得られる$W(C^{*'}C^*, 9)$の最大根の分布のα点は

$$21.85(\alpha=0.05),\quad 27.28(\alpha=0.01)$$

なので、表２中の幾つかの値及び(6)式は極めて高度に有意である。特に(6)式はプールする前の累積χ^2統計量（[3]参照）の値99.64の約91.3%を説明しており、２群I_1, I_2への群分けは十分意味があると思ってよい。

4. 列の群分け

一方、列の群分けにも興味があるが、列には水準間に自然な順序のある点で、行の群分けとは異なっている。すなわち、あらゆる組合せの群分けを考える替わりに、列の順序に沿ってどこで分割すると全体の変動を最もよく表すかという視点で考える方が適切である。この例の場合、それは、軽症とそれ以外、及び重症とそれ以外という分割を考えることに相当する。そこで、表１において２列・３列をプールしてできる10×２分割表と、１列・２列をプールしてできる10×２分割表でそれぞれPearsonの適合度χ^2を計算すると、

$$\chi^2(1;2,3)=91.25,\quad \chi^2(1,2;3)=8.39 \qquad (7)$$

が得られる。３節の終わりに述べた行をプールする前の累積χ^2統計量99.64とはこの和のことである。この大きい方91.25の全体に対する寄与率も91.6%あり、２列と３列をプールすることによる情報損失は小さい。つまり、表１に潜む交互作用は次の２×２表（表３）のパターンで代表される。表３に対するPearson適合度χ^2は87.71となり、これは全累積χ^2値の88%を説明している。このように本節の方法は累積χ^2統計量の最大成分を用いているので最大累積χ^2(max acc. χ^2) 法と呼ばれる。

なお、(7)式の２つのχ^2統計量の同時分布はパラメータをρ_{12}(5)とする２変数χ^2分布であり、χ^2(1;2,3)＝91.25はその最大値の分布を用いて評価することができる。今の場合そのp値は0.00となり、極めて高度に有意である。

表2　行の多重比較のための補助表（行を並べ替えてあることに注意）

行番号	10	5	4	8	7	9	2	1	6	3
10	0	0.85	2.52	1.67	8.93	7.72	18.6	18.3	15.3	50.1**
5		0	0.88	0.65	6.86	5.79	15.2	15.9	12.5	47.8**
4			0	1.10	4.71	3.73	9.41	11.4	8.51	23.5*
8				0	3.83	3.95	10.5	9.29	8.35	23.2*
7					0	0.41	1.71	0.68	0.82	1.48
9						0	0.30	1.24	0.30	0.85
2							0	2.7	0.35	1.48
1								0	0.92	1.01
6									0	0.16
3										0

表3　均質な行及び列をそれぞれプールした２×２表

行＼列	$m=1(1)$	$m=2(2,3)$	
$l=1$ (1,2,3,6,7,9)	1,157	4,127	5,284
$l=2$ (4,5,8,10)	1,009	5,615	6,624
計	2,166	9,742	11,908

5. ブロック交互作用モデル

今回のデータでは行、列がそれぞれ2群に分かれた。ここで一般に行が$I_1,\cdots,I_L$のL群、列が$j_1,\cdots,j_M$のM群に分かれたとして、群内は無交互作用であるとするモデルを考えよう。すなわち、i, i'が同じ群Iに属するか、j, j'が同じ群Jに属するとき、交互作用パラメータは0、すなわち、

$$\log p_{ij} - \log p_{i'j} - \log p_{ij'} + \log p_{i'j'} = 0$$

とする。そして、i, i'及びj, j'が行、列のそれぞれ異なる群に属するときのみ、交互作用パラメータが0でないとすると、ブロック交互作用モデル

$$p_{ij} = p_{i\cdot}p_{\cdot j}\lambda_{lm}, \quad i \in I_l,\ j \in J_m, \tag{8}$$

が導かれる。モデル(8)は完全独立モデル

$$p_{ij} = p_{i\cdot}p_{\cdot j} \tag{9}$$

に対し、交互作用パラメータλ_{lm}, $l=1,\cdots,L$; $m=1,\cdots,M$を余分に用いているが、その実質的な数（自由度）は$(L-1)(M-1)$である。$L=M=1$が完全独立モデル、$L=a$, $M=b$が全交互作用モデルに相当する。

6. 表1のデータへのあてはめと考察

3節、4節の群分けの結果から、今回のデータには自由度1の交互作用モデル

$$p_{ij} = p_{i\cdot}p_{\cdot j}\lambda_{lm}$$
$$\begin{pmatrix} l=1 \text{ for } \ i=1,2,3,6,7,9\,; \ \ l=2 \text{ for } i=4,5,8,10 \\ m=1 \text{ for } \ j=1\,; \ \ m=2 \text{ for } j=2,3 \end{pmatrix} \tag{10}$$

が示唆されたことになる。このモデルをあてはめたときの、セル度数の推定量は次のような考え方で求められる。

例えば、元表の(1, 1)セルに対しては、表3においてその属する$l=1$, $m=1$に対する度数が交互作用パターンの推定を与える。これをそこに属する行和及び列和で比例配分して

$$\hat{y}_{11} = 1157 \times \frac{678}{5284} \times \frac{2166}{2166} = 148.5$$

がセル度数の推定量になる。例えば、(10, 3)セルに対しては同様の考え方から

$$\hat{y}_{10,3} = 5615 \times \frac{1818}{6624} \times \frac{1419}{9742} = 224.5$$

が得られる。このようにして得られた推定値を表1括弧内の下段に与える。これを完全独立モデル(9)に対する推定値$y_{i\cdot}y_{\cdot j}/y_{\cdot\cdot}$（括弧内上段の値）と比べると改良の度合は明らかである。ちなみに、モデル(10)あてはめ後の適合度χ^2は自由度$(10-1)(3-1)-1=17$に対し、8.04となり、極めて良い適合を示す。結局このデータから、職業群(10, 5, 4, 8)が(7, 9, 2, 1, 6, 3)に比べて、初診時重症者の割合が多いことが言える。重症者の多い群には10.無職、5.農林,漁業,採鉱従事者などが含まれる。現在は癌の早期発見システムがかなり広汎に行き渡り、当時のように歴然とした差は最早ないと思われる。例えば現在は、特定の企業に属していなくても、市町村の健康診断のシステムが行き渡っているのは衆知の事柄である。いずれにせよ、このような調査データ、及び解析は疫学上大変有用なことと思われる。

＊参考文献

[1] 会田雅人・広津千尋(1983)：順序制約下での多項分布比較の一方法と数表：応用統計学会誌 12, pp.101-110.

[2] Hirotsu, C.(1982)：Use of cumulative efficient scores for testing ordered alternatives in discrete models：Biometrika 69, pp.567-577.

[3] 広津千尋(1992)：実験データの解析－分散分析を超えて－：共立出版, 東京.