

先行変数は、媒介変数と同様 3 変数の直列的因果関係に関与しているが、順序を異にする。前者を図式的に $C \rightarrow A \rightarrow B$ と表す場合、 B に及ぼす C の影響は A により媒介されているので、モデルは A が与えられたときの条件付独立を主張する $[CA][AB]$ となる。 $C \rightarrow B \rightarrow A$ なら当然対応するモデルは $[CB][BA]$ であろう。しかし、交互作用の定義から明らかなように対数線型モデルの効果項は、変数の順序に影響されないで、これらのモデルを $[AB][AC]$, $[AB][BC]$ と表記しても差し支えない。

顕示変数は、今までとは逆に周辺分割表ではあたかも独立にみえる A, B 間の関係を要因 C の条件付で明確にするのであるから、周辺分割表 AB_+ についてはモデル $[A][B]$ があてはまることは簡単にわかる。ところが本来の変数 A, B の関係は分割表 ABC の内部で要因 C の水準間で、ねじれたように存在する。したがって、 $I \times J \times K$ の分割表におけるモデルは二次交互作用 u^{ABC} を含む $[ABC]$ がふさわしい。

テスト要因が矯正変数として働くのは周辺分割表 AB_+ にみられる 2 変数の関係が、要因 C の水準に分解したときに逆転する場合である。したがって、周辺分割表に対しては明らかにモデル $[AB]$ が成り立つ。一方、分割表 ABC の内部においては条件付見込比が周辺分割表とは (1 を挟んで) 逆向きになっているから、当然

$$u_{ij}^{AB} \neq 0, \quad u_{ijk}^{ABC} \neq 0$$

となる。よって、3 変数分割表については飽和モデル $[ABC]$ がよくあてはまるはずである。

4.4 効果の推定とモデルの選択

1) サンプルの種類

$I \times J \times K$ 三元分割表の標本は、二元分割表の場合同様ポアソン分布、多項分布、層別多項分布に従うものが代表的である。しかし、いずれの標本であっても、分割表のセル (i, j, k) の観測頻度 n_{ijk} が、確率変数 N_{ijk} の実現値とみなされる点是不変である。問題は、 N_{ijk} が従う分布により生じるパラメータ推定上の影響である。

- ① ポアソン分布に従う標本： 確率変数 N_{ijk} は、それぞれ独立に期待値

m_{ijk} のポアソン分布に従う： $N_{111} \sim \mathcal{P}(m_{111}), \dots, N_{IJK} \sim \mathcal{P}(m_{IJK})$. 当然ながら、抽出されるケースの総数 (n_{ijk}) については予め何の制約もない。この標本の尤度関数は次のとおりである：

$$L = \prod \prod \prod \frac{m_{ijk}^{n_{ijk}} e^{-m_{ijk}}}{n_{ijk}!} \quad (4.31)$$

ポアソン分布は再生性をもつので、頻度の合計 N_{+++} もまたポアソン分布に従う。ただし、このときの期待値は $m_{+++} = \sum \sum \sum m_{ijk}$ である： $N_{+++} \sim \mathcal{P}(m_{+++})$. 表 4.7 はフロリダ沖の島に棲息するトカゲの生態分布の調査結果 (Schoener, 1968) を Fienberg (1980) が再編成したものである。研究の関心はトカゲの種類による餌の相違が、木の上での棲み分けに反映されているかどうかにある。後述するように、このデータは条件付独立のモデル $[AC][BC]$ によってよく説明される。

表 4.7

C : 種類	A : 枝の高さ	B : 枝の直径 (inches)		計
		1) ≤ 4.0	2) > 4.0	
1) sagrei	1) > 4.75 feet	32	11	43
	2) ≤ 4.75	86	35	121
2) distichus	1) > 4.75 feet	61	41	102
	2) ≤ 4.75	73	70	143
合 計		252	157	409

資料：Schoener (1968)

② 多項分布に従う標本： 調査目的に応じて予め抽出するケースの総数は n_{+++} に定められる。無作為に抽出されるケースが特性 (A_i, B_j, C_k) を同時に有する確率 p_{ijk} で、そのようなケースが合わせて n_{ijk} 観測されたとする ($ijk = 111, \dots, IJK$)。ただし、 $\sum \sum \sum p_{ijk} = 1, n_{+++} = \sum \sum \sum n_{ijk}$. このとき、標本の尤度関数は

$$L = \frac{n_{+++}!}{\prod_i \prod_j \prod_k n_{ijk}!} \prod_i \prod_j \prod_k p_{ijk}^{n_{ijk}}$$

で表される。ところで、セル (i, j, k) の期待頻度は $m_{ijk} = n_{+++} p_{ijk}$ で求められ、また標本の設定条件から $m_{+++} = n_{+++}$ がいえるので、この尤度関数は次のように書き直せる。

$$L = \frac{n_{+++}!}{\prod_i \prod_j \prod_k n_{ijk}!} \prod_i \prod_j \prod_k \left(\frac{m_{ijk}}{m_{+++}} \right)^{n_{ijk}}$$

$$= \prod_i \prod_j \prod_k \left[\frac{m_{ijk}^{n_{ijk}} e^{-m_{ijk}}}{n_{ijk}!} \right] / \frac{m_{+++}^{n_{+++}} e^{-m_{+++}}}{n_{+++}!} \quad (4.32)$$

(4.32) は多項分布に従う標本を, $m_{+++} = n_{+++}$ なる条件付のポアソン分布からの標本とみなせることを示している.

表 4.8

C: クリニック	A: 出生前の 看護期間	B: 幼児の生存		計
		1) 死亡	2) 生存	
1)	1) 1カ月未満	3	176	179 } 476
	2) 1カ月以上	4	293	
2)	1) 1カ月未満	17	197	214 } 239
	2) 1カ月以上	2	23	

資料: Bishop (1969)

表 4.9

(a) 分割表 ABC

C: 植樹の時期	A: 切断の長さ	B: 生育結果		計
		1) 枯	2) 生	
1) 直後	1) 長	84	156	240
	2) 短	133	107	240
2) 春	1) 長	156	84	240
	2) 短	209	31	240
		582	378	960

(b) 周辺分割表 AB₊

C: 時期	A: 切断の長さ	B: 生育結果		計
		1) 枯	2) 生	
—	1) 長	240	240	480
—	2) 短	342	138	480
		582	378	960

(c) 周辺分割表 +BC

C: 植樹の時期	A: 切断の長さ	B: 生育結果		計
		1) 枯	2) 生	
1) 直後	—	217	263	480
2) 春	—	365	115	480
		582	378	960

資料: Bartlett (1935)

③ 層別多項分布の標本： 標本全体はいくつかの層からなっており、層ごとに選ぶケース数を前もって決めて抽出する。その事例 (Bishop, 1969) を表 4.8 に示す。母親や医師にとって出生前の看護期間(変数 A)が新生児の生存の可能性(変数 B)に影響を及ぼすか否かは重大な関心事である。表 4.8 のデータは二つのクリニック(変数 C)から抽出するケース数($n_{++1}=476$, $n_{++2}=239$)を決めて集められた。層別に標本を抽出する方法は、このほかに 2 変数の組合せを考慮に入れて行われることもある。実験的色彩の強い調査に好まれる方法で、表 4.9 に紹介するデータはよく引用される例である。これは、すももの生育状況を、切断した根の長さ(変数 A)と植樹の時期(変数 C)との関係で調べようとしたもので、実験要因の組合せごとの件数を等しくしてある：

$$n_{1+1}=n_{2+1}=n_{1+2}=n_{2+2}=240$$

最初に、表 4.8 の例にそって 1 変数についてのみ層別抽出を行った標本の尤度関数を考える。ここでは変数 C の各カテゴリーの大きさが n_{++k} ($k=1, \dots, K$) に固定されている。第 k 層に属する母集団から無作為に選んだケースが特性 (A_i, B_j) を兼ね備えている条件付確率を $p_{ij:k}=P(A_i B_j | C_k)$ で表そう(ただし、 $\sum_i \sum_j p_{ij:k}=1$)。すると、第 k 層において観測頻度 n_{11k}, \dots, n_{IJK} が得られる確率は、

$$f(n_{11k}, \dots, n_{IJK}) = \frac{n_{++k}!}{\prod_i \prod_j n_{ijk}!} \prod_i \prod_j p_{ij:k}^{n_{ijk}}$$

である。こうして互いに独立な K 層の多項分布を重ねた分割表全体にわたって頻度 $\{n_{ij1}\}, \dots, \{n_{ijK}\}$ を観測する確率は、

$$f(n_{111}, \dots, n_{IJK}) = \prod_k \left[\frac{n_{++k}!}{\prod_i \prod_j n_{ijk}!} \prod_i \prod_j p_{ij:k}^{n_{ijk}} \right]$$

となる。次に、 $p_{ij:k}=m_{ijk}/m_{++k}$ を上式に代入して、標本の尤度関数を求めておこう。

$$\begin{aligned} L &= \prod_k \left[\frac{n_{++k}!}{\prod_i \prod_j n_{ijk}!} \prod_i \prod_j \left(\frac{m_{ijk}}{m_{++k}} \right)^{n_{ijk}} \right] \\ &= \frac{n_{+++}!}{\prod_i \prod_j \prod_k n_{ijk}!} \prod_i \prod_j \prod_k \left(\frac{m_{ijk}}{m_{+++}} \right)^{n_{ijk}} / \frac{n_{+++}!}{\prod_k n_{++k}!} \prod_k \left(\frac{m_{++k}}{m_{+++}} \right)^{n_{++k}} \\ &= \prod_i \prod_j \prod_k \left[\frac{m_{ijk}^{n_{ijk}} e^{-m_{ijk}}}{n_{ijk}!} \right] / \prod_k \frac{m_{++k}^{n_{++k}} e^{-m_{++k}}}{n_{++k}!} \end{aligned} \quad (4.33)$$

今度は、二つの要因の組合せについて大きさを決める標本の尤度関数について、表 4.9 の例に沿って調べてみる。この例では周辺分割表 A_+C の頻度が n_{i+k} に固定されている。切断の長さ、植樹の時期の組合せ (A_i, C_k) のなかで植物が枯れてしまう条件付確率は $p_{1:ik}$ 、逆にうまく生育してくれる確率は $p_{2:ik}$ で表される(ただし $\sum_j p_{j:ik}=1$)。このとき、セル $(i, 1, k)$, $(i, 2, k)$ において n_{i1k} , n_{i2k} のケースを観測する確率は次の多項分布に従う:

$$f(n_{i1k}, n_{i2k}) = \frac{n_{i+k}!}{\prod_j n_{ijk}!} \prod_j p_{j:ik}^{n_{ijk}} \quad (j=1, 2)$$

よって、分割表 ABC にわたって頻度 $\{n_{i1j}\}, \dots, \{n_{iJK}\}$ が観測される確率は

$$f(\{n_{ijk}\}) = \prod_i \prod_k \left[\frac{n_{i+k}!}{\prod_j n_{ijk}!} \prod_j p_{j:ik}^{n_{ijk}} \right] \quad (i=1, \dots, I; j=1, \dots, J; k=1, \dots, K)$$

で与えられる。標本全体の尤度関数を期待頻度を用いて表すために、上式に

$$p_{j:ik} = m_{ijk}/m_{i+k} = m_{ijk}/n_{i+k}$$

を代入してみると、(4.33) に類似したものが導かれる。

$$L = \prod_i \prod_j \prod_k \left[\frac{m_{ijk}^{n_{ijk}} e^{-m_{ijk}}}{n_{ijk}!} \right] / \prod_i \prod_k \frac{m_{i+k}^{n_{i+k}} e^{-m_{i+k}}}{n_{i+k}!} \quad (4.34)$$

3種類のサンプリング方法について導かれた尤度関数 (4.31)~(4.34) を比較すると、すべて、

$$L = \prod_i \prod_j \prod_k \left[\frac{m_{ijk}^{n_{ijk}} e^{-m_{ijk}}}{n_{ijk}!} \right] / A \quad (4.35)$$

なる共通様式にまとめられることに気づく。いうまでもなく、(4.31) では $A=1$ である。その他の場合、 A はサンプリングの設計に伴い固定された周辺度数の関数となっている。母集団が単純な多項分布の場合では $m_{+++}=n_{+++}$ 、層別多項分布の場合では、(4.33) については $m_{++k}=n_{++k}$ 、また (4.34) では $m_{i+k}=n_{i+k}$ であるから、結局 A の部分は対数線型モデルによる尤度の変化には無関係な定数である。したがって、ある対数線型モデルの下で得られる最尤推定量は、サンプリングの種類には影響されないといえる。ただし、層別多項分布からの標本の場合、周辺度数の固定の仕方により、設定可能なモデルに制約が生じる点に注意しなければならない。すなわち、(4.33) であれば $u_k^0=0$ ($k=1, \dots, K$) と仮定することは許されない。また、(4.34) の場合は、交互作用

u_{ik}^{AC} がモデルに含まれている必要がある。

2) 効果パラメータの推定

まず, (4.35) より対数尤度関数を求めておこう。

$$\log L = \sum \sum \sum (n_{ijk} \log m_{ijk} - m_{ijk}) - \text{const.}$$

上式は, 飽和モデル (4.22) の下で

$$\begin{aligned} \log L = & -\sum \sum \sum \exp(u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}) \\ & + [n_{+++}u + \sum n_{i++}u_i^A + \sum n_{+j+}u_j^B + \sum n_{++k}u_k^C \\ & + \sum \sum n_{ij+}u_{ij}^{AB} + \sum \sum n_{i+k}u_{ik}^{AC} + \sum \sum n_{+jk}u_{jk}^{BC} \\ & + \sum \sum \sum n_{ijk}u_{ijk}^{ABC}] - \text{const.} \end{aligned} \quad (4.36)$$

となるので, [] に囲まれた部分にあって u パラメータに隣りあう

$$n_{+++}, n_{i++}, n_{+j+}, n_{++k}, n_{ij+}, n_{i+k}, n_{+jk}, n_{ijk}$$

が十分統計量である。これより, 尤度方程式

$$\begin{aligned} \hat{m}_{ijk} &= n_{ijk}, & \hat{m}_{ij+} &= n_{ij+}, & \hat{m}_{i+k} &= n_{i+k}, & \hat{m}_{+jk} &= n_{+jk} \\ \hat{m}_{i++} &= n_{i++}, & \hat{m}_{+j+} &= n_{+j+}, & \hat{m}_{++k} &= n_{++k}, & \hat{m}_{+++} &= n_{+++} \end{aligned}$$

を得る。これで, 飽和モデル (4.22) の下ではどのセルの期待頻度の最尤推定値も観測頻度と一致する ($\hat{m}_{ijk} = n_{ijk}$) ことは明らかであろう。さて, 十分統計量のなかでも, n_{+++}, \dots, n_{+jk} はどれも n_{ijk} を足して求められるので, 最小十分統計量は n_{ijk} ($ijk = 111, \dots, IJK$) ということになる。

不飽和モデルの最小十分統計量を見つける規則は, サンプリングの種類にかかわらずごく簡単なものである。すなわち, モデルの簡易表現に明示される最高次の効果項に対応してつくられる周辺分割表上の観測頻度が, そのモデル下での最小十分統計量というだけのことである (Birch, 1963)。したがって, この規則は三元分割表そのものも周辺分割表の一種とみなせば, 飽和モデルにも適用可能である。たとえば, 3変数の完全な独立性を主張するモデル $[A][B][C]$ ならば, 周辺分割表 $A_{++}, +B_{++}, ++C$ 上の観測頻度 $n_{i++}, n_{+j+}, n_{++k}$ が最小十分統計量である。これはこのモデルの下での尤度関数

$$\begin{aligned} \log L = & -\sum \sum \sum \exp(u + u_i^A + u_j^B + u_k^C) \\ & + [n_{+++}u + \sum n_{i++}u_i^A + \sum n_{+j+}u_j^B + \sum n_{++k}u_k^C] - \text{const.} \end{aligned}$$

から, パラメータ u, u_i^A, u_j^B, u_k^C に隣接する統計量をみれば容易に確認される。尤度方程式は

$$\hat{m}_{i++} = n_{i++}, \quad \hat{m}_{+j+} = n_{+j+}, \quad \hat{m}_{++k} = n_{++k}, \quad \hat{m}_{+++} = n_{+++} \quad (4.37)$$

であるから, 周辺分割表 $A_{++}, +B_{++}, ++C$ 上で期待頻度の最尤推定量が観測値に固定されることは明白である。3変数が互いに完全に独立なら, 前節の (4.28) より, 分割表内部のセル (i, j, k) の期待頻度の最尤推定量について,

$$\hat{m}_{ijk} = \frac{\hat{m}_{i++}\hat{m}_{+j+}\hat{m}_{++k}}{(\hat{m}_{+++})^2}$$

が いえる。この右辺に (4.37) を代入すれば、次式が求まる：

$$\hat{m}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{(n_{+++})^2}$$

変数 C の各水準内で A, B 間の条件付独立性を主張するモデル

$$[AC][BC] : \log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} + u_{jk}^{BC}$$

では、 n_{i+k}, n_{+jk} が最小十分統計量で、尤度方程式

$$\begin{aligned} \hat{m}_{i+k} &= n_{i+k}, & \hat{m}_{+jk} &= n_{+jk}, & \hat{m}_{i++} &= n_{i++} \\ \hat{m}_{+j+} &= n_{+j+}, & \hat{m}_{++k} &= n_{++k}, & \hat{m}_{+++} &= n_{+++} \end{aligned}$$

から周辺分割表 $A_+C, +BC$ 上で期待頻度の最尤推定値は対応する期待頻度に固定されることがわかる。この関係を (4.30 b) に代入すれば、セル (i, j, k) の期待頻度の最尤推定量が求まる：

$$\hat{m}_{ijk} = \frac{n_{i+k}n_{+jk}}{n_{++k}}$$

同時独立性のモデル

$$[AB][C] : \log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB}$$

では、周辺分割表 $AB_+, ++C$ 上で期待頻度の推定値の合計は対応する観測頻度の周辺度数に固定される：

$$\begin{aligned} \hat{m}_{ij+} &= n_{ij+}, & \hat{m}_{++k} &= n_{++k}, & \hat{m}_{i++} &= n_{i++} \\ \hat{m}_{+j+} &= n_{+j+}, & \hat{m}_{+++} &= n_{+++} \end{aligned}$$

ここで (4.30 a) に上式を代入すれば、セル (i, j, k) の期待頻度の推定量が

$$\hat{m}_{ijk} = \frac{n_{ij+}n_{++k}}{n_{+++}}$$

であることは直ちに理解されるであろう。このように、なんらかの独立性を主張するモデルについては、(最小)十分統計量から分割表内部のセルの期待頻度を推定することが可能である。その他の不飽和モデルに関しては、こうした推定式を導くことはできない。期待頻度の推定が目的であれば、モデルの種類を問わず、反復比例あてはめ法 (IPFP : Iterative Proportional Fitting Procedure) を用いればよい。この手法は、先述の Birch (1963) の規則を利用したもので、適当な初期値から始めて、期待頻度の推定値がモデルによって固定される周辺分割表上で観測頻度の合計と等しくなるよう比例調整する。Goodman (1979) や Bishop ら (1975) は、この方法を推奨しているが、推定された \hat{m}_{ijk} から当該モデルの効果パラメータやその分散を別途計算しなければならないことがあり、実際には、ニュートン法を応用して推定することが多いようである (§2.5 参照)。そのため、モデルをデザイン行列を用いた形、 $\theta = X\beta$ 、で表現しておくことと便利である。

飽和モデルについては、 $2 \times 3 \times 2$ の分割表を例に (4.24), (4.26) にデザイン行列 X ならびに効果ベクトル β の構成が示されている。不飽和モデルはこれらを基につくられる。より正確には、行列 X の要素を指定することにより、モデルの種類および効果を測る視点が指定され、それに応じて β の要素が解釈されるというべきであろう。一般に、

期待頻度 m_{111} より m_{IJK} の対数を要素とする列ベクトル

$$\theta = [\log m_{111}, \dots, \log m_{IJK}]'$$

の推定値は、モデルにより異なることがあっても、効果を測る視点によって影響されることはない。条件付独立性を主張するモデル $[AC][BC]$ を取り上げて、 $2 \times 3 \times 2$ 分割表における不飽和モデルのデザイン行列の指定の仕方をみてみよう。

① ダミーコーディング： $\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} + u_{jk}^{BC}$
基準セルを $(1, 1, 1)$ とする。

$$\theta = X\beta$$

θ	X	β
	$u \langle u^A, u^B, u^C \rangle \langle u^{AC}, u^{BC} \rangle$	
$\log m_{111}$	1 0 0 0 0 0 0 0	
$\log m_{121}$	1 0 1 0 0 0 0 0	
$\log m_{131}$	1 0 0 1 0 0 0 0	β_0
$\log m_{211}$	1 1 0 0 0 0 0 0	β_1
$\log m_{221}$	1 1 1 0 0 0 0 0	β_2
$\log m_{231}$	1 1 0 1 0 0 0 0	β_3
$\log m_{112}$	1 0 0 0 1 0 0 0	β_4
$\log m_{122}$	1 0 1 0 1 0 1 0	β_5
$\log m_{132}$	1 0 0 1 1 0 0 1	β_6
$\log m_{212}$	1 1 0 0 1 1 0 0	β_7
$\log m_{222}$	1 1 1 0 1 1 1 0	
$\log m_{232}$	1 1 0 1 1 1 0 1	

u 効果項を要素とする列ベクトル

$$u = [u, u_2^A, u_2^B, u_3^B, u_2^C, u_{22}^{AC}, u_{22}^{BC}, u_{32}^{BC}]'$$

と、上の β とは $u = \beta$ の関係にあることは明白であろう。

② ANOVA コーディング： $\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} + u_{jk}^{BC}$

主効果、交互作用のなかで添字が、 $i=1, j=1, k=1$ のいずれかに該当するものを推定上、冗長なパラメータとする。

$$\begin{array}{c} \theta \\ \log m_{111} \\ \log m_{121} \\ \log m_{131} \\ \log m_{211} \\ \log m_{221} \\ \log m_{231} \\ \log m_{112} \\ \log m_{122} \\ \log m_{132} \\ \log m_{212} \\ \log m_{222} \\ \log m_{232} \end{array} = \begin{array}{c} X \\ \mathbf{u} \quad \langle \mathbf{u}^A, \mathbf{u}^B, \mathbf{u}^C \rangle \quad \langle \mathbf{u}^{AC}, \mathbf{u}^{BC} \rangle \\ \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 0 & -1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 & -1 & 1 & 0 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 0 & -1 & -1 & -1 & 0 \\ 1 & 1 & 0 & 1 & -1 & -1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 & 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} \beta \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{array}$$

\mathbf{u} 効果項を要素とする列ベクトル

$$\mathbf{u} = [\mathbf{u}, \mathbf{u}_2^A, \mathbf{u}_2^B, \mathbf{u}_3^C, \mathbf{u}_2^C, \mathbf{u}_{22}^{AC}, \mathbf{u}_{22}^{BC}, \mathbf{u}_{32}^{BC}]'$$

と、上の β とは $\mathbf{u} = \beta$ の関係にある。

3) モデルの選択

三元分割表には飽和モデル $[ABC]$ から一様性のモデル $[]$ まで 19 ものモデルが候補となるわけであるから ‘よい’ モデルを見つけ出すのは容易なことではない。ただし、層別多項分布に従う標本の場合は、サンプリングの条件から設定できるモデルの種類はこれよりは少なくなる。たとえば、すももの生育状況を調べるためのデータ(表 4.9)では、周辺分割表 A_+C の頻度が固定されるわけで、この条件を考慮したモデルは

$$[AC], [AC][B], [AC][AB], [AC][BC], [AC][AB][BC], [ABC]$$

の 6 種類に限られる。しかし、AIC やカイ二乗尤度比統計量 (G^2) の数理的基準にのみ頼った結果、実質的な解釈に困難をきたすようでは分析の目的は果たせない。モデル選択の方針を定める上でとりわけ重要なのは、本書で何度も強調してきた専門領域での理論との整合性である。整合性がある程度みだされた ‘よさそうな’ モデルのなかから最終候補を絞り込む場合には、解釈の容易さ

を優先させるべきであろう。その意味では独立性に言及できるモデル(条件付、同時、完全)が好まれる。節約の原則(principle of parsimony)の意味をパラメータの少なさに限定していると、採用したモデルから実質的な結論を導くの
に苦しみられるおそれがある。よって、AIC その他の数理的基準は、こうした方針を効率的に運用するための指針であると捉えておくのが賢明であろう。たとえ、‘AIC 最小’の考えに従うにしても、どれほどの差をもって二つのモデルの優劣を決定できるのかは明確ではない。有意性検定についても絶対的な有意水準が存在しない以上、結局最後は主観的な判断に委ねられることになる。このような本書の立場を踏まえた上で、表 4.7~4.9 のデータに対して、なんらかの独立性を意味するモデルを中心に検討してみよう。(その他のモデルを含めてテストしても結論に変わりはない。)

表 4.10

モデル	G^2	df	p	AIC (パラメータ数)
データ：表 4.7				
[ABC]	—	—	—	60.978 (8)
[AB][AC][BC]	0.149	1	0.699	59.127 (7)
[AC][BC]*	2.026	2	0.363	59.004 (6)
[AB][C]	24.429	3	0.000	79.407 (5)
[A][B][C]	25.037	4	0.000	78.015 (4)
データ：表 4.8				
[ABC]	—	—	—	40.864 (8)
[AB][AC][BC]	0.043	1	0.835	38.907 (7)
[AC][BC]*	0.082	2	0.960	36.946 (6)
[AB][C]	205.870	3	0.000	240.734 (5)
[A][B][C]	211.482	4	0.000	244.346 (4)
データ：表 4.9				
[ABC]*	—	—	—	38.742 (8)
[AB][AC][BC]	2.294	1	0.130	39.036 (7)
[AB][AC]	105.182	2	0.000	139.924 (6)
[AC][BC]	53.440	2	0.000	88.182 (6)
[AC][B]	151.019	3	0.000	183.761 (5)

* は採択されたモデルである。

表 4.10 に示す AIC および尤度比カイ二乗統計量(G^2)によれば、表 4.7, 4.8 では要因 C の水準に関して条件付独立性を主張するモデル [AC][BC] の AIC が最も小さい。解釈の容易さを考えれば、これより複雑なモデル [AB][AC][BC] や [ABC] を好んで選ぶ理由はないだろう。よって、トカゲの分布生態

(表 4.7) については、トカゲの種類 (C) さえ区別すれば、棲息する枝の高さ (A) と太さ (B) の間には特に関係があるとは認められない。同じく、表 4.8 からはクリニック (C) の違いを踏まえると、出産前の看護期間 (A) と新生児の産後 1 か月以内の生死は独立であるといえる。最後に、すももの実験データ (表 4.9) について、AIC が最小となるのは飽和モデル $[ABC]$ である。このモデルによると、切断する根の長さ (A) はこの植物の生育 (B) に影響を与え、かつその影響は植樹の時期により左右されるということになる。ちなみに、Bishop ら (1975, p. 89) は、AIC を利用しないで有意性検定のみ用いているので、二次交互作用 u_{ABC}^{ABC} を 0 とするモデル $[AB][AC][BC]$ を採択している。

データ分析はモデル選択で終わるわけではない。残された課題の一つに結論の一般性があげられる。トカゲの棲息状況 (表 4.7) を例に考えても、調査地域に棲息するトカゲがこの 2 種類に限定されているなら先の結論は一般性をもつ。しかし、棲息場所が大きく異なるトカゲが何種類か存在するなら問題は複雑である。棲息状況を幅広く扱うために、カテゴリーを増やしたのでは必然的に 0 の観測値が多数生じるであろう。この問題の統計的処理には 5 章で述べるような工夫が要求される。それとは別に現実的な処理として、分類基準を違えた分割表をいくつか用意しそれぞれについてモデル検証を行い結論を比較することが考えられる。すべての表について同じモデルが採用されれば結論の一般性は高いし、そうでなければ得られた知識は限定的なものとなる。表 4.11 は Schoener (1968) のトカゲの生態調査の結果をやはり Fienberg (1980) が再編成したものであるが、比較するトカゲの種類 (A) と、変数 B, C の分類基準が表 4.7 と異なっている。各変数の主効果が必ず含まれるモデルに限定してテストしてみたところ (表 4.12)、条件付独立性を仮定するモデル $[AC][BC]$ のあてはまりが良好であった。しかし、わずかながらモデル $[AB][AC][BC]$ の AIC が小さいので、数理的基準を重視するなら表 4.7 とは異なった結論が導かれることになる。幸い、両モデルの AIC の差はきわめて小さいので、モデル採択に続くパラメータ解釈の負担や、結論の一般性を考慮すると条件付独立性を認める $[AC][BC]$ を採択しても問題はないだろう。したがって、少なくとも表 4.7, 4.11 のデータに関する限り、 $[AC][BC]$ は共通のモデルとなりうるわけである。

表 4.11

C:種類	A:枝の高さ	B:枝の直径 (inches)		計
		1) ≤ 2.5	2) > 2.0	
1) sagrei	1) > 5.00 feet	15	18	33
	2) ≤ 5.00	48	84	132
2) angusticeps	1) > 5.00 feet	21	1	22
	2) ≤ 5.00	3	2	5
合	計	87	105	192

資料: Schoener (1968)

表 4.12

モデル	G^2	df	p	AIC (パラメータ数)
[ABC]	—	—	—	49.774 (8)
[AB][AC][BC]	2.706	1	0.100	50.480 (7)
[AC][BC]*	4.882	2	0.087	50.656 (6)
[AB][C]	57.387	3	0.000	101.161 (5)
[A][B][C]	70.080	4	0.000	111.854 (4)

* は採択されたモデルである。

第2の課題は、推定されたパラメータの実質的な解釈である。基本的には推定値を §3.4 に紹介したような方法で比較しまとめていくわけだが、特殊な比較の例としていわゆる従属変数に注目した扱いを次節で説明する。

4.5 反応変数の分析

分割表を構成する変数は一般に、説明変数(explanatory variable)とよばれるものと被説明変数または**反応変数**(response variable)とよばれるものによって分けられる。おもしろいのは、反応変数の一つとは限らず、それ以上複数個あっても許されることであろう。前節の表 4.8 や表 4.9 は一つの反応変数 B と説明変数 A, C からできている。これに対し、表 4.7 および表 4.11 の例はトカゲの種類を説明変数としてみれば、棲息する枝の高さと、太さが反応変数と考えられる。しかし、常に説明変数と反応変数の2種類が要求されるわけではなく、いずれか一方の変数によってのみ表がつけられる場合がある。(もっともこのとき、両者の区別は意味を失う。) 消費者の意識態度測定ではいくつかの指標を用いるが、こうした指標間の関連をみるのに使われる分割表は反応変数だけで構成されている。もし、意識や態度が何かの行動を規定すると捉えら