Jorge Manuel Pires Mendonça

ANÁLISE DE TABELAS DE CONTINGÊNCIA INCOMPLETAS E DE TABELAS COM VARIÁVEIS ORDINAIS

Departamento de Estatística e Investigação Operacional
Faculdade de Ciências da Universidade de Lisboa
Fevereiro de 1997

DISSERTAÇÃO DE MESTRADO

PROBABILIDADES E ESTATÍSTICA

À Maria do Carmo

e ao Luís Humberto

Agradecimentos

Reconheço publicamente a todos aqueles que me prestaram o apoio que tornou possível a realização deste trabalho.

À minha orientadora, Drª Margarida Mendes Leal, quero agradecer o apoio, disponibilidade e ajuda prestada nesta dissertação.

À Fernanda Otília pela disponibilidade e incentivo.

Ao Instituto Superior de Engenharia do Porto, pelas facilidades concedidas para a realização deste trabalho.

Ao Fernando Oliveira pela sua extraordinária disponibilidade.

Aos meus colegas de mestrado pelo ânimo, apoio e incentivo.

Aos meus sogros, pela paciência e ânimo que me dedicaram.

Aos meus pais que me apoiaram desde o primeiro momento.

A todos, os meus sinceros agradecimentos.

Índice

Capítulo I

Introdução

1.1 Apresentação do trabalho

Um conjunto de indivíduos de uma população pode ser dividido em classes ou categorias segundo determinados aspectos que pretendemos observar. Estes aspectos que estão na base da classificação são variáveis, cujas categorias devem ser mutuamente exclusivas.

Vamos abordar em particular tabelas que representam contagens do número de indivíduos que satisfazem determinadas combinações de categorias das variáveis que estivermos a estudar. A estas tabelas chamamos "tabelas de contingência".

Frequentemente o cruzamento dessas categorias origina tabelas onde existem células vazias, isto é, células onde alguma combinação de categorias não faz sentido. Por exemplo, a seguinte tabela imaginária apresenta o cruzamento do número de golos marcados por uma equipa de futebol com o respectivo resultado ocorrido num campeonato de futebol

		resultado	
nºde golos marcados	vitória	empate	derrota
0	-------	5	3
1	6	3	1
2	3	0	2
3	7	1	3

Verificamos que a célula vazia da tabela corresponde a uma situação impossível de ocorrer pois uma equipa que não marca golos não ganha. Esta célula constitui um exemplo daquilo se designa por zero estrutural.

Este trabalho é centralizado na análise de tais tabelas incompletas à luz do modelo loglinear.

A razão da escolha do modelo loglinear deve-se ao facto deste não só averiguar a existência ou não de independência entre as variáveis, mas também permitir quantificar os efeitos que as variáveis ou combinações destas exercem sobre os valores observados, o que na análise clássica não acontece. O modelo loglinear relaciona as probabilidades associadas a cada célula com uma função logarítmica de vários parâmetros.

Numa primeira fase analisaremos tabelas bidimensionais incompletas e em seguida tabelas tridimensionais incompletas. Como última fase deste trabalho temos a abordagem de modelos loglineares para variáveis ordinais.

1.2 Nomenclatura

Para uma melhor compreensão das notações existentes neste trabalho, vou aqui indicar o significado das mesmas.

1.2.1 Tabela bidimensional I × J

É uma tabela com I linhas e J colunas que representa todas as combinações possíveis das diversas categorias de duas variáveis, A e B respectivamente. A variável A tem I categorias e a variável B tem J categorias.

	B_1	B_2	....	B_J
A_1	x_{11}	x_{12}	.....	x_{1j}
A_2	x_{21}	x_{22}	.....	x_{2j}
....	....	....	....	....
A_I	x_{I1}	x_{I2}	....	X_{ij}

x_{ij} - frequência observada na célula (i,j) da tabela bidimensional I × J

Total marginal linha

Para cada linha i, temos $x_{i+} = \sum_{j=1}^{J} x_{ij}$, i = 1,..., I

Total marginal coluna

Para cada coluna j, temos $x_{+j} = \sum_{i=1}^{I} x_{ij}$, j = 1,..., J

Ao número total de observações ou dimensão da amostra, chamamos total da tabela que designamos por $N = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}$ ou $N = \sum_{i=1}^{I} x_{i+} = \sum_{j=1}^{J} x_{+j}$

1.2.2 Tabela de contingência tridimensional I × J × K

É uma tabela que resulta da classificação dos N elementos de uma amostra, segundo três variáveis, A, B e C respectivamente. A tem I categorias, B tem J categorias e C tem K categorias. Definimos assim uma tabela de contingência tridimensional I × J × K. Esta tabela é constituída por I linhas, J colunas, K estratos e I × J × K células.

	C_1				...				C_K
	B_1	B_2	...	B_J	...	B_1	B_2	...	B_J
A_1	x_{111}	x_{121}	...	x_{1J1}	...	x_{11K}	x_{12K}	...	x_{1JK}
A_2	x_{211}	x_{221}	...	x_{2J1}	...	x_{21K}	x_{22K}	...	x_{2JK}
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
A_I	x_{I11}	x_{I21}	...	x_{IJ1}	...	x_{I1K}	x_{I2K}	...	x_{IJK}

x_{ijk} - frequência observada da célula (i,j,k)

O total da amostra N é dado por $N = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}$.

Para tabelas tridimensionais há dois tipos de totais marginais:

Totais marginais simples

$$x_{i++} = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk} \text{ , } i = 1,..., I$$

$$x_{+j+} = \sum_{i=1}^{I}\sum_{k=1}^{K} x_{ijk} \text{ , } j = 1,..., J$$

$$x_{++k} = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ijk} \text{ , } k = 1,..., K$$

Totais marginais duplos

$$x_{ij+} = \sum_{k=1}^{K} x_{ijk} \text{ , } i = 1,..., I \quad \text{e} \quad j = 1,..., J$$

$$x_{i+k} = \sum_{j=1}^{J} x_{ijk} \ , i = 1,..., I \ \text{ e } \ k = 1,..., K$$

$$x_{+jk} = \sum_{i=1}^{I} x_{ijk} \ , \ j = 1,..., J \ \text{ e } \ k = 1,..., K$$

1.2.3 Modelos loglineares em tabelas bidimensionais I × J

Sendo m_{ij}, a frequência absoluta esperada da célula (i,j), o modelo loglinear saturado usual define-se por:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)} \ , i = 1,..., I \quad j = 1,..., J$$

Os elementos $\mu_{1(i)}, \mu_{2(j)}$ e $\mu_{12(ij)}$ englobam respectivamente I, J e I × J parâmetros sujeitos às seguintes restrições

$$\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = \sum_i \mu_{12(ij)} = \sum_j \mu_{12(ij)} = 0$$

onde:

$$\mu = \frac{1}{I.J} \sum_{(i,j)} \log m_{ij}$$

$$\mu_{1(i)} = \frac{1}{J} \sum_j \log m_{ij} - \mu \ , i = 1,..., I$$

$$\mu_{2(j)} = \frac{1}{I} \sum_i \log m_{ij} - \mu \ , \quad j = 1,..., J$$

$$\mu_{12(ij)} = \log m_{ij} - \frac{1}{I} \sum_j \log m_{ij} - \frac{1}{J} \sum_i \log m_{ij} + \mu \ , \ i = 1,..., I \ ; j = 1,..., J$$

Estes termos representam respectivamente :

μ - efeito da média global

$\mu_{1(i)}$ - efeito principal da categoria i da variável A

$\mu_{2(j)}$ - efeito principal da categoria j da variável B

$\mu_{12(ij)}$ - efeito de interacção entre os níveis i e j das variáveis A e B

O número de parâmetros independentes de cada um destes termos é determinado pelo tamanho da tabela e pelas restrições impostas aos parâmetros. O quadro seguinte apresenta cada um desses números

Termos	Parâmetros Independentes
μ	1
μ_1	I - 1
μ_2	J -1
μ_{12}	(I - 1).(J - 1)
Total	(I - 1).(J - 1) + (I - 1) + (J - 1) + 1 = I.J

Para além deste modelo podem-se considerar também todos os que dele se obtêm por remoção de alguns dos seus termos.

1.2.4 Modelos loglineares para tabelas tridimensionais I × J × K

Tal como no caso anterior, representando por m_{ijk} a frequência esperada da célula (i,j,k) da tabela tridimensional I × J × K , o modelo loglinear saturado usual define-se da seguinte forma:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)}$$

$i = 1,....I \quad j = 1,...,J \quad k = 1,...,K$

onde: μ - efeito da média global

$\mu_{1(i)}$ - efeito principal da categoria i da variável A

$\mu_{2(j)}$ - efeito principal da categoria j da variável B

$\mu_{3(k)}$ - efeito principal da categoria k da variável C

$\mu_{12(ij)}$ - efeito de interacção entre os níveis i e j das variáveis A e B

$\mu_{13(ik)}$ - efeito de interacção entre os níveis i e k das variáveis A e C

$\mu_{23(jk)}$ - efeito de interacção entre os níveis j e k das variáveis B e C

$\mu_{123(ijk)}$ - efeito de interacção de 2ª ordem

As restrições do modelo são as seguintes:

$$\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = \sum_i \mu_{12(ij)} = \sum_j \mu_{12(ij)} = \sum_i \mu_{13(ik)} = \sum_k \mu_{13(ik)} =$$

$$\sum_j \mu_{23(jk)} = \sum_k \mu_{23(jk)} = \sum_i \mu_{123(ijk)} = \sum_j \mu_{123(ijk)} = \sum_k \mu_{123(ijk)} = 0$$

Podem considerar-se também modelos mais pequenos que se obtêm do modelo saturado por remoção de alguns dos seus termos.

1.2.5 Estatísticas de teste X^2 e Y^2

O objectivo da análise loglinear centra-se em determinar o modelo loglinear, que com o menor número de termos, descreve os dados de um modo satisfatório. Qualquer modelo loglinear não saturado obtem-se a partir do modelo saturado, por remoção de um ou vários dos seus termos. Necessitamos, assim de averiguar quais os termos do modelo que se podem eliminar. Para tal podemos recorrer a testes estatísticos.

Testar a hipótese de que os termos que não figuram num modelo são nulos, equivale a testar a hipótese de independência por ele especificada.

O teste de qualquer destas hipóteses é feito com base nas estatísticas X^2 ou Y^2 que são assintoticamente equivalentes, cujas expressões no caso bidimensional são respectivamente as seguintes:

$$X^2 = \sum_{i,j} \frac{\left(x_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}} \qquad \text{e} \qquad Y^2 = 2\sum_{i,j} x_{ij} \log\left(\frac{x_{ij}}{\hat{m}_{ij}}\right)$$

onde $\hat{m}_{ij}$ representa o estimador da frequência absoluta esperada m_{ij}.

O cálculo do valor destas estatísticas sob cada um dos modelos possíveis não apresenta problemas de maior, considerando para o efeito que os estimadores das frequências esperadas são definidos por:

Tabelas de Contingência Bidimensionais

- Modelo de independência

$$\hat{m}_{ij} = \frac{x_{i+} x_{+j}}{N^2}$$

- Modelo que especifica que o efeito da variável B é nulo:

$$\hat{m}_{ij} = \frac{x_{i+}}{J}$$

- Modelo que especifica que o efeito da variável A é nulo:

$$\hat{m}_{ij} = \frac{x_{+j}}{I}$$

- Modelo que especifica que o efeito da variável A e o efeito da variável B são nulos:

$$\hat{m}_{ij} = \frac{N}{I.J}$$

Tabelas de Contingência Tridimensionais

- Modelo de independência mútua:

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

- Modelo de independência parcial entre (A,B) e C:

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

- Modelo de independência parcial entre (A,C) e B:

$$\hat{m}_{ijk} = \frac{x_{i+k} x_{+j+}}{N}$$

- Modelo de independência parcial entre (B,C) e A:

$$\hat{m}_{ijk} = \frac{x_{+jk} x_{i++}}{N}$$

- Modelo de independência condicional entre A e B dada C:

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

- Modelo de independência condicional entre B e C dada A:

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

- Modelo de não existência de interacção de 2ª ordem

(não há estimação directa)

Os estimadores das frequências esperadas sob os modelos anteriores são todos determinados de forma directa. Esta estimação baseia-se essencialmente em dois resultados de Birch(1963):

- " num dado modelo, todas as frequências marginais observadas, correspondentes aos parâmetros desconhecidos, são estatísticas suficientes e são iguais aos estimadores de máxima verosimilhança das correspondentes frequências esperadas marginais".
- " os estimadores de máxima verosimilhança das frequências esperadas em cada uma das células da tabela, ficam unívocamente determinados através das restrições impostas pelo modelo, juntamente com as condições impostas no primeiro resultado".

Em tabelas tridimensionais, no caso do modelo de não existência de interacção de 2ª ordem isto é

$\mu_{123(ijk)} = 0, \ \forall i,j,k, \ i = 1,...,I \quad j = 1,...,J \quad k = 1,...;K$, a estimação directa não é possível. Por isso, recorre-se ao método iterativo de ajustamento proporcional para tabelas tridimensionais proposto por Deming e Stephen (1940) e descrito por Bishop(1975).

Observando este modelo:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)}$$

verifica-se que, segundo os resultados de Birch, as restrições relativas às marginais são as seguintes:

(1) $m_{ij+} = x_{ij+}, \ \forall i,j$

(2) $m_{i+k} = x_{i+k}, \ \forall i,k$

(3) $m_{+jk} = x_{+jk}, \ \forall j,k$

Como ponto de partida toma-se o conjunto dos valores $\left\{\hat{m}_{ijk}^{(0)}\right\}$ tal que $\hat{m}_{ijk}^{(0)} = 1$, $\forall i,j,k$ e ajustam-se proporcionalmente estes estimadores, por forma que as restrições estabelecidas vão sendo sucessivamente verificadas.

Utilizando a restrição (1) toma-se:

$$\hat{m}_{ijk}^{(1)} = \frac{\hat{m}_{ijk}^{(0)} x_{ij+}}{\hat{m}_{ij+}^{(0)}}$$

Estes estimadores $\hat{m}_{ijk}^{(1)}$ serão ajustados de modo a satisfazerem a restrição (2), obtendo-se:

$$\hat{m}_{ijk}^{(2)} = \frac{\hat{m}_{ijk}^{(1)} x_{i+j}}{\hat{m}_{i+k}^{(1)}}$$

O mesmo se fará a $\hat{m}_{ijk}^{(2)}$ relativamente à restrição (3), obtendo-se:

$$\hat{m}_{ijk}^{(3)} = \frac{\hat{m}_{ijk}^{(2)} x_{+jk}}{\hat{m}_{+jk}^{(2)}}$$

Estes três passos constituem um ciclo que se pode repetir utilizando os valores $\hat{m}_{ijk}^{(3)}$ como valores de partida no 1° passo do ciclo seguinte. O processo termina quando no final de um dado ciclo se constatar que os estimadores não se alteraram significativamente, em relação aos valores obtidos no final do ciclo anterior. Ou seja, o processo termina quando, ao finalizar o v-ésimo ciclo se tiver satisfeita a seguinte condição para um δ fixado: $\left|\hat{m}_{ijk}^{(3v)} - \hat{m}_{ijk}^{(3v-3)}\right| < \delta$.

A distribuição assintótica tanto de X^2 como de Y^2 é a de um χ^2 cujo número de graus de liberdade depende do modelo a ajustar. O número de graus de liberdade pode ser determinado por um dos seguintes processos:

- calcular a diferença entre o número total de células da tabela e o número total de parâmetros independentes do modelo.

- calcular o número de parâmetros englobados pelos termos que não figuram no modelo.

Com base no primeiro processo apresenta-se no quadro seguinte, o número de graus de liberdade associado às estatísticas de teste sob cada um dos modelos em causa. Em primeiro lugar para tabelas de contingência bidimensionais I × J:

Modelo	Número de graus de liberdade
$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}$	(I - 1).(J - 1)
$\log m_{ij} = \mu + \mu_{1(i)}$	I.(J - 1)
$\log m_{ij} = \mu + \mu_{2(j)}$	J.(I - 1)
$\log m_{ij} = \mu$	I.J - 1

Seguindo a mesma metodologia para tabelas de contingência tridimensionais temos:

Modelo	Número de graus de liberdade
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ij)} + \mu_{23(jk)}$	(I - 1).(J - 1).(K - 1)
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ij)}$	I.(J - 1).(K -1)
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{23(jk)}$	J.(I - 1).(K - 1)
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ij)} + \mu_{23(jk)}$	K.(I - 1).(J - 1)
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)}$	(I.J - 1).(K - 1)
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ij)}$	(I.K - 1).(J - 1)
$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{23(jk)}$	(J.K - 1).(I - 1)
$\log m_{ijk} = \mu$	I.J.K - I - J - K + 2

CAPÍTULO II

Tabelas incompletas

1. Introdução

As tabelas de contingência são tabelas de frequências que condensam toda a informação relativamente à classificação cruzada dos elementos de uma população. Frequentemente encontramos tabelas com células cujas frequências observadas são nulas ou cujas frequências observadas são simplesmente omitidas.

Estes zeros podem ser de dois tipos:

- zeros "**à priori**" ou zeros "**estruturais**", que ocorrem quando há determinadas combinações de categorias que são impossíveis de se verificar (a probabilidade associada a essas combinações é nula)
- zeros "**amostrais**", que ocorrem devido à amostragem, por exemplo, quando se recolhe uma amostra demasiado pequena em relação ao número de células da tabela ou quando há variáveis com categorias demais.

Para exemplificar a definição do primeiro tipo de zeros, consideramos a seguinte tabela resultante de uma amostra recolhida para um dado grupo de adolescentes. Esta tabela foi extraída de B. S. Everitt(1977).

		Sexo				
		Masculino			feminino	
idade		12-15	16-17		12-15	16-17
problemas de saúde relacionados com	S	4	2		9	17
	M	------	------		4	8
	H	42	7		19	10
	N	57	20		71	31

S- reprodução sexual

M- problemas menstruais

H- perfeitamente saudável

N- respostas que não se enquadram nas anteriores

Naturalmente os rapazes não são afectados por problemas menstruais. A análise de tabelas com este tipo de células nulas constituiu o objecto principal do nosso estudo.

O segundo tipo de zeros pode ser exemplificado pela seguinte tabela referente a uma amostra de resultados dos alunos do 12°ano no exame nacional de Matemática:

	Classificação Matemática			
	0-5	5-10	10-15	15-20
rapaz	0	3	5	2
rapariga	4	2	7	0

Em qualquer destes dois tipos de tabelas é impossível ajustar o modelo loglinear saturado, pois os estimadores das frequências esperadas, são dados por combinações lineares dos logaritmos das frequências e o logaritmo de zero não está definido.

Relativamente à tabela com zeros amostrais, um dos processos de resolver este problema consiste em aumentar a dimensão da amostra e refazer a tabela de contagens. No entanto, por vezes, tal não é possível.

Um outro processo, sugerido por diversos autores (por exemplo, Bishop (1975)) e que se tem mostrado relativamente eficaz, consiste em aumentar o valor da frequência de cada célula da tabela, adicionando-lhe meia unidade, antes de prosseguir a análise.

As tabelas com zeros "à priori" ou "estruturais", designam-se por tabelas "incompletas" e a sua análise apresenta problemas específicos.

Neste capítulo analisaremos este tipo de tabelas sob duas perspectivas diferentes:

- análise loglinear de tabelas com zeros estruturais de um modo geral

- análise loglinear à luz do modelo de Quasi-Independência que se baseia sobretudo na análise das partes das tabelas que não contêm zeros estruturais. Eventualmente podem ser analisadas tabelas completas que se transformam em tabelas incompletas com a exclusão de algumas das suas células. Esta exclusão é muitas vezes justificada quando queremos analisar determinadas tabelas em condições especiais cuja estrutura inicial não é a de tabela incompleta.

As células com zeros estruturais podem ocorrer em vários contextos. As observações para algumas células numa tabela de contingência são muitas vezes truncadas ou não referidas (Goodman 1968). Outras vezes, certas combinações são impossíveis, e a probabilidade zero é atribuída a estas células (Kastenbaum 1958).

2. Tabelas com zeros estruturais. Caso geral

Os modelos loglineares a ajustar a estas tabelas não devem incluir parâmetros correspondentes às células nulas, pois sabe-se à partida que esses parâmetros irão ser nulos. Para tais modelos, a estimação dos valores esperados deve ser feita por um processo iterativo, tomando-se como valor de partida para as células não nulas o valor um e para as células nulas o valor zero. Assim, a tabela dos valores esperados estimados também terá nulas as células correspondentes aos zeros à priori. A medida de ajustamento a utilizar pode ser X^2 ou Y^2.

No caso de tabelas completas o número de graus de liberdade associado a cada modelo é dado pela diferença entre o total de células da tabela e o número de parâmetros independentes a estimar incluídos no modelo. Para as tabelas incompletas ter-se-á de subtrair ainda o número de células com zeros. No entanto, é necessário algum cuidado na determinação do número de parâmetros independentes a estimar, pois os que correspondem a células com frequências nulas não necessitam de ser estimados, por serem já conhecidos; há pois que descontar estes do total de parâmetros independentes a estimar. Se designarmos por N_1, N_2 (descontando já o número de parâmetros correspondentes às células nulas), N_3 as quantidades acima referidas, o número de graus de liberdade do modelo é $N_1 - N_2 - N_3$.

Como exemplo do cuidado que devemos ter no cálculo do número de graus de liberdade, analisaremos de seguida a tabela já anteriormente apresentada:

		Sexo				
		masculino			feminino	
idade		12-15	16-17		12-15	16-17
problemas de saúde relacionados com	S	4	2		9	17
	M	------	------		4	8
	H	42	7		19	10
	N	57	20		71	31

Observamos nesta tabela que estão em estudo três variáveis. Consideramos:

- idade como variável linha com duas categorias: 12-15 e 16-17
- sexo como variável coluna com duas categorias: masculino e feminino
- "problemas de saúde relacionados com" como variável estrato com quatro categorias: S, M, H e N.

Observamos ainda que nesta tabela existem dois zeros estruturais, nas células (1,1,2) e (2,1,2) respectivamente.

Suponhamos que pretendemos ajustar o modelo de independência mútua entre as variáveis, ou seja:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} \ , i = 1,2 \quad j = 1,2 \quad k = 1,2,3,4$$

Este ajustamento faz-se através do teste da hipótese

$$Ho: \ \mu_{12(ij)} = \mu_{13(ik)} = \mu_{23(jk)} = \mu_{123(ijk)} = 0 \ , \ \forall i,j,k.$$

Vamos calcular os estimadores das frequências esperadas utilizando o método iterativo de ajustamento proporcional de Deming-Stephen. Começamos por assumir os valores iniciais dos estimadores

$$\hat{m}_{ijk}^{(0)} = \begin{cases} 1, \text{ se } (i,j,k) \text{ não contem um zero estrutural} \\ 0, \text{ se } (i,j,k) \text{ contem um zero estrutural} \end{cases}$$

e recordando que terão de ser satisfeitas as restrições seguintes:

(1) $m_{i++} = x_{i++}, \ \forall i$

(2) $m_{+j+} = x_{+j+}, \ \forall j$

(3) $m_{++k} = x_{++k}, \ \forall k$

Utilizando a restrição (1) toma-se:

$$\hat{m}_{ijk}^{(1)} = \frac{\hat{m}_{ijk}^{(0)} x_{i++}}{\hat{m}_{i++}^{(0)}}$$

Estes estimadores $\hat{m}_{ijk}^{(1)}$ serão ajustados de modo a satisfazerem a restrição (2), obtendo-se:

$$\hat{m}_{ijk}^{(2)} = \frac{\hat{m}_{ijk}^{(1)} x_{+j+}}{\hat{m}_{+j+}^{(1)}}$$

O mesmo se fará a $\hat{m}_{ijk}^{(2)}$ relativamente à restrição (3), obtendo-se:

$$\hat{m}_{ijk}^{(3)} = \frac{\hat{m}_{ijk}^{(2)} x_{++k}}{\hat{m}_{++k}^{(2)}}$$

Estes três passos constituem um ciclo que se pode repetir utilizando os valores $\hat{m}_{ijk}^{(3)}$ como valores de partida no 1° passo do ciclo seguinte. O processo termina quando no final de um dado ciclo se constatar que os estimadores não se alteraram significativamente, em relação aos valores obtidos no final do ciclo anterior. Ou seja, o processo termina quando, ao finalizar o v-ésimo ciclo se tiver satisfeita a seguinte condição para um δ fixado: $\left|\hat{m}_{ijk}^{(3v)} - \hat{m}_{ijk}^{(3v-3)}\right| < \delta$.

No caso da tabela em causa ao fim de cinco ciclos esta última condição é satisfeita para $\delta = 0.01$. Obtemos a seguinte tabela de estimativas das frequências esperadas:

		Sexo				
		masculino			feminino	
idade		12-15	16-17		12-15	16-17
problemas de saúde relacionados com	S	7.37	3.04		8.21	3.39
	M	------	------		8.49	3.51
	H	26.12	10.78		29.09	12.00
	N	59.95	24074		66.76	27.55

A estatística de teste toma o valor $X^2 = 28.24$ e o número de graus de liberdade associado é g.l. = 16 - 2 - (1 + 1 +1 + 3) = 8. A um nível de significância $\alpha = 0.05$ (o valor do quantil de probabilidade 0.05 do $\chi^2_{(8)} = 15.507$), o que nos leva a rejeitar a hipótese de independência mútua entre as três variáveis.

Suponhamos agora que pretendiamos ajustar o modelo de não existência de interacção de 2ª ordem dado por

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)}$$

$$i = 1,2 \quad j = 1,2 \quad k = 1,2,3,4$$

Seguindo o mesmo processo iterativo,onde as restrições a satisfazer são :

(1) $m_{ij+} = x_{ij+}, \forall i,j$

(2) $m_{i+k} = x_{i+k}, \forall i,k$

(3) $m_{+jk} = x_{+jk}, \forall j,k$

as estimativas das frequências esperados ao fim de quatro ciclos são os seguintes:

		Sexo				
		masculino			feminino	
idade		12-15	16-17		12-15	16-17
problemas de saúde relacionados com	S	4.03	1.97		8.97	7.03
	M	------	------		4.00	8.00
	H	39.81	9.19		21.19	7.81
	N	59.16	17.84		68.84	33.16

O valor da estatística de teste é $X^2 = 2.03$. Relativamente ao cálculo do número de graus de liberdade, tendo em conta o que foi referido, tem-se: $N_1 = 16$, $N_2 = 12$ e $N_3 = 2$, pois tendo de contar as células correspondentes aos zeros estruturais às quais correspondem os parâmetros nulos, tem-se:

Parâmetros	Número de parâmetros independentes
μ	1
$\mu_{1(i)}$	1
$\mu_{2(j)}$	1
$\mu_{3(k)}$	3
$\mu_{12(ij)}$	1
$\mu_{13(ik)}$	3
$\mu_{23(jk)}$	2
total	12

Sendo assim o número de graus de liberdade é g.l. = 2. Então a um nível de significância de 0.05 somos levados a concluir que não existe interacção de 2ª ordem.

3. Modelo de Quasi-Independência

3.1 Apresentação do modelo

O termo "quasi-independência" foi formalmente introduzido na literatura de tabelas de contingência por Goodman(1968) em tabelas bidimensionais, mas o seu uso remonta a 1961/63/64/65 quando Goodman realizou o seu trabalho sobre o modelo "mover-stayer", fluxos de transacção e análise de tabelas da mobilidade social. As extensões a tabelas multidimensionais foram desenvolvidas com mais detalhe por Fienberg(1972) e Haberman(1974).

Vamos começar por analisar tabelas bidimensionais incompletas e o modelo loglinear da quasi-independência.

Seja m_{ij} a frequência absoluta esperada de indivíduos, de uma amostra de dimensão N correspondente à célula (i,j) da tabela bidimensional I×J.

As frequências esperadas marginais definem-se por:

$$\begin{aligned} m_{i.} &= \sum_{j=1}^{J} m_{ij}, i = 1,...,I \\ m_{.j} &= \sum_{i=1}^{I} m_{ij}, j = 1,...,J \\ m_{..} &= \sum_{i=1}^{I}\sum_{j=1}^{J} m_{ij} \end{aligned} \qquad (3.1\text{-}1)$$

O modelo geral de independência entre as variáveis correspondentes à 'linha' e à 'coluna', respectivamente, é dado por

$$m_{ij} = a_i b_j \qquad (3.1\text{-}2)$$

onde a_i e b_j são constantes positivas para $i = 1,...,I$ e $j = 1,...,J$.

Esta definição de independência é equivalente à que é usada geralmente e que consiste em dizer que as classificações linha e coluna são independentes se as probablidades associadas às células verificam a condição:

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad (3.1\text{-}3)$$

onde π_{ij} é a probabilidade de um indíviduo de uma tabela bidimensional I×J ser classificado na célula (i,j) , $\pi_{i.}$ e $\pi_{.j}$ são respectivamente as probabilidades de um indíviduo ser classificado na i-ésima categoria da 1ª variável e na j-ésima categoria da segunda variável.

Seja S o conjunto das células de uma tabela incompleta bidimensional que não contem zeros estruturais. Suponhamos que a condição 3.1-2 se verifica para todos os m_{ij} de células de S . Em tais casos dizemos que as linhas e colunas de S são quasi-independentes.

Vamos considerar o modelo loglinear saturado usual:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}, \; i = 1,...,I \quad j = 1,...,J \qquad (3.1\text{-}4)$$

para as células de S com as seguintes restrições:

$$\sum_{i=1}^{I} \delta_i^{(2)} \mu_{1(i)} = \sum_{j=1}^{J} \delta_j^{(1)} \mu_{2(j)} = 0 \qquad (3.1\text{-}5)$$

$$\sum_{i=1}^{I} \delta_{ij} \mu_{12(ij)} = \sum_{j=1}^{J} \delta_{ij} \mu_{12(ij)} = 0$$

onde

$$\delta_{ij} = \begin{cases} 1, \text{ se } (i,j) \in S \\ 0, \text{ caso contrario} \end{cases}$$

$$\delta_i^{(2)} = \begin{cases} 1 \text{ se } \delta_{ij} = 1 \text{ para algum } j \\ 0, \text{caso contrario} \end{cases}$$

$$\delta_j^{(1)} = \begin{cases} 1, \text{se } \delta_{ij} = 1 \text{ para algum } i \\ 0, \text{ caso contrario} \end{cases}$$

Definimos então o modelo da quasi-independência para a subtabela S, considerando $\mu_{12(ij)} = 0, \forall i,j$, no modelo saturado usual, pelo que se obtém :

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} \qquad (3.1\text{-}6)$$

3.2 Interpretação do modelo

Podemos encarar a interpretação do modelo da quasi-independência numa tabela bidimensional incompleta, de uma maneira análoga à interpretação do modelo de independência em tabelas completas.

A quasi-independência implica que as proporções de indivíduos, nas células correspondentes de duas linhas (colunas) são as mesmas, desde que não consideremos nessas linhas (colunas) aquelas que têm zeros estruturais. Assim a quasi-independência é uma forma de independência condicional com a restrição da nossa atenção a uma porção incompleta da tabela.

Consideremos, como exemplo uma tabela bidimensional 3×3 onde a célula (1,1) é omitida. S consiste no conjunto das restantes oito células. Se a quasi-independência se aplicar a estas, podemos escrever as respectivas frequências na forma multiplicativa $m_{ij} = a_i b_j$, $(i,j) \in S$. Os valores esperados das frequências ficam assim escritos na seguinte tabela:

------	a_1b_2	a_1b_3
a_2b_1	a_2b_2	a_2b_3
a_3b_1	a_3b_2	a_3b_3

Se eliminarmos a primeira linha ficamos com

a_2b_1	a_2b_2	a_2b_3
a_3b_1	a_3b_2	a_3b_3

uma tabela bidimensional 2×3 que satisfaz a independência usual. O mesmo acontece se eliminarmos a primeira coluna

a_1b_2	a_1b_3
a_2b_2	a_2b_3
a_3b_2	a_3b_3

ficamos com uma tabela bidimensional 3×2 que também satisfaz a independência.

Há no entanto algumas configurações de S que não satisfazem a independência usual mesmo eliminando linhas ou colunas. Por exemplo, numa tabela bidimensional 3×3 em que estão omitidas as células da diagonal principal e onde a quasi-independência se aplica às restantes células, a tabela dos valores esperados é

------	a_1b_2	a_1b_3
a_2b_1	-----	a_2b_3
a_3b_1	a_3b_2	-----

Ora, neste caso, a eliminação de linhas ou colunas não reduz o estudo a uma tabela que satisfaz a independência usual. Isto é um exemplo de como a quasi-independência se reduz à independência se pudermos decompôr a tabela.

3.3 Modelos de Amostragem

Os esquemas de amostragem que podem estar subjacentes a uma determinada tabela são, entre outros:

1- Poisson, com variáveis de Poisson independentes para cada célula.

2- Multinomial.

3- Produto multinomial, isto é, um conjunto de multinomiais independentes para linhas ou para colunas.

Para tabelas incompletas, tal como para tabelas completas, os estimadores de máxima verosimilhança dos valores esperados sob o modelo da quasi-independência têm os mesmos resultados assintóticos para os três tipos de esquema de amostragem.

Vamos considerar o esquema de Poisson.

Assume-se que x_{ij} , o valor observado da célula (i,j) $\in$ S, é uma observação de uma variável de Poisson com valor médio m_{ij} e que estas I× J variáveis de Poisson são mutuamente independentes.

A função de verosimilhança é dada por $$L = \Pi^{*} m_{ij}^{x_{ij}} \frac{e^{-m_{ij}}}{x_{ij}!} \qquad (3.3\text{-}1)$$

onde Π^{*} representa o produtório ao longo das células (i,j) de S.

O logaritmo neperiano desta função é

$$LnL = \sum \delta_{ij} \left[x_{ij} \log(m_{ij}) - m_{ij} - \log(x_{ij}!) \right] \qquad (3.3\text{-}2)$$

onde $\delta_{ij} = \begin{cases} 1, & (i,j) \in S \\ 0, & \text{caso contrario} \end{cases}$

Sob o modelo 3.1.2, vem $\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}$, tem-se então

$$LnL = \sum \delta_{ij} \left[x_{ij} \left(\mu + \mu_{1(i)} + \mu_{2(j)} \right) - m_{ij} - \log\left(x_{ij}! \right) \right] =$$

$$= \sum \delta_{ij} x_{ij} \mu + \sum \delta_{ij} x_{ij} \mu_{1(i)} + \sum \delta_{ij} x_{ij} \mu_{2(j)} - \sum \delta_{ij} m_{ij} - \sum \delta_{ij} \log\left(x_{ij}! \right) =$$

$$= x_{++}\mu + \sum_i x_{i+}\mu_{1(i)} + \sum_j x_{+j}\mu_{2(j)} - m_{++} - \sum \delta_{ij} \log(x_{ij}) \qquad (3.3\text{-}3)$$

3.4 Estimação sob o modelo de quasi-independência

Vamos agora ver como se calculam os estimadores de m_{ij} para depois efectuarmos o ajustamento do modelo. No que segue vamos desenvolver dois métodos iterativos equivalentes para o cálculo dos estimadores.

Consideramos uma expressão equivalente a 2.1-2:

$$m_{ij} = \delta_{ij} \cdot a_i \cdot b_j \quad \begin{matrix} i = 1,...,I \\ j = 1,...,J \end{matrix} \qquad (3.4\text{-}1)$$

onde $\delta_{ij} = \begin{cases} 1, & (i,j) \in S \\ 0, & \text{caso contrario} \end{cases}$

Os processos que vamos utilizar recorrem as restrições que constituem os resultados de Birch:

1) $\hat{m}_{i+} = x_{i+}$, $i = 1,...,I$

2) $\hat{m}_{+j} = x_{+j}$, $j = 1,...,J$ (3.4-2)

O primeiro destes processos constitui uma aplicação do método iterativo de ajustamento proporcional de Deming-Stephan.

Convencionamos o passo 0 como:

$$\hat{m}_{ij}^{(0)} = \delta_{ij} \qquad i \in \{1,...,I\}, j \in \{1,...,J\} \qquad (3.4\text{-}3)$$

Ajustando $\hat{m}_{ij}$ de modo que a restrição (1) seja satisfeita, o passo seguinte dá-nos

$$\hat{m}_{ij}^{(1)} = \frac{\hat{m}_{ij}^{(0)} \cdot x_{i+}}{\hat{m}_{i+}^{(0)}} \qquad (3.4\text{-}4)$$

Estes estimadores $\hat{m}_{ij}^{(1)}$ são agora ajustados de modo a satisfazerem a restrição (2). Daí o passo seguinte:

$$\hat{m}_{ij}^{(2)} = \frac{\hat{m}_{ij}^{(1)} . x_{+j}}{\hat{m}_{+j}^{(1)}} \tag{3.4-5}$$

Estes dois passos constituem um ciclo que se pode repetir utilizando os valores $\hat{m}_{ij}^{(2)}$ como valores de partida no 1° passo, em vez de $\hat{m}_{ij}^{(0)}$. O processo termina quando no fim de um dado ciclo, se verificar que os estimadores não se alteraram significativamente, em relação aos valores obtidos no final do ciclo anterior.

Ou seja, se no v-ésimo ciclo da iteração temos:

$$\hat{m}_{ij}^{(2v-1)} = \frac{\hat{m}_{ij}^{(2v-2)} x_{i+}}{\hat{m}_{i+}^{(2v-2)}} \tag{3.4-6}$$

$$\hat{m}_{ij}^{(2v)} = \frac{\hat{m}_{ij}^{(2v-1)} x_{+j}}{\hat{m}_{+j}^{(2v-1)}} \tag{3.4-7}$$

o processo termina no fim deste ciclo se se tiver para um δ fixado: $\left| m_{ij}^{(2v)} - m_{ij}^{(2v-2)} \right| < \delta$.

O segundo processo também constitui um método iterativo onde recorremos à estimação dos parâmetros multiplicativos a_i (i = 1,...,I) e b_j (j = 1,...,J) para depois obtermos os estimadores de máxima verosimilhança dos m_{ij} correspondentes.

Reescrevendo as equações de máxima verosimilhança segundo as relações da quasi-independência, vem:

$$\hat{a}_i \sum_{j=1}^{J} \delta_{ij} \hat{b}_j = x_{i+} \quad , \; i = 1,...,I \tag{3.4-8}$$

$$\hat{b}_j \sum_{i=1}^{I} \delta_{ij} \hat{a}_i = x_{+j} \quad , \; j = 1,...,J \tag{3.4-9}$$

donde

$$\hat{a}_i = \frac{x_{i+}}{\sum_{j=1}^{J} \delta_{ij} \hat{b}_j} \tag{3.4-10}$$

$$\hat{b}_j = \frac{x_{+j}}{\sum_{i=1}^{I} \delta_{ij} \hat{a}_i} \tag{3.4-11}$$

Esta segunda representação sugere o seguinte processo iterativo para estimar os $\{a_i\}$ e os $\{b_j\}$.

Começando por $b_j^{(0)} = 1$, $j = 1,...,J$ (3.4-12)

temos em seguida:

$$a_i^{(1)} = \frac{x_{i+}}{\sum_{j=1}^{J} \delta_{ij} b_j^{(0)}}, \quad i = 1,...,I \tag{3.4-13}$$

$$b_j^{(1)} = \frac{x_{+j}}{\sum_{i=1}^{I} \delta_{ij} a_i^{(1)}} \quad , j = 1,...,J \tag{3.4-14}$$

Continuamos até ao v-ésimo ciclo da iteração onde obtemos:

$$a_i^{(v)} = \frac{x_{i+}}{\sum_{j} \delta_{ij} b_j^{(v-1)}} \quad , i = 1,...,I \tag{3.4-15}$$

$$b_j^{(v)} = \frac{x_{+j}}{\sum_{i} \delta_{ij} a_i^{(v)}} \quad , j = 1,...,J \tag{3.4-16}$$

Depois do v-ésimo ciclo os estimadores de m_{ij} são dados por

$$\hat{m}_{ij}^{(2v)} = \delta_{ij} a_i^{(v)} b_j^{(v)} \tag{3.4-17}$$

O processo pára quando a diferença entre os estimadores obtidos no final de dois ciclos consecutivos fôr suficientemente pequena.

Prova-se que as expressões (3.4-7) e (3.4-17) conduzem a resultados iguais, conforme vamos ver de seguida.

Sabendo que $\hat{m}_{ij}^{(0)} = \delta_{ij}$

$$\hat{m}_{ij}^{(2v-1)} = \frac{\hat{m}_{ij}^{(2v-2)} x_{i+}}{\hat{m}_{i+}^{(2v-2)}} \quad \text{e} \quad \hat{m}_{ij}^{(2v)} = \frac{\hat{m}_{ij}^{(2v-1)} x_{+j}}{\hat{m}_{+j}^{(2v-1)}} \quad , \quad \text{conforme}$$

espressões (3.4-6) e (3.4-7) respectivamente, e que

$$b_j^{(0)} = 1$$

$$a_i^{(v)} = \frac{x_{i+}}{\sum_j \delta_{ij} b_j^{(v-1)}} \quad , \quad b_j^{(v)} = \frac{x_{+j}}{\sum_i \delta_{ij} a_i^{(v)}} \quad \text{e} \quad \hat{m}_{ij}^{(2v)} = \delta_{ij} a_i^{(v)} b_j^{(v)}$$

conforme expressões (3.4-15) , (3.4-16) e (3.4-17), respectivamente.

Pretende-se provar a igualdade das espressões (3.4-7) e (3.4-17) isto é

$$\frac{\hat{m}_{ij}^{(2v-1)} x_{+j}}{\hat{m}_{+j}^{(2v-1)}} = \delta_{ij} a_i^{(v)} b_j^{(v)}, \quad v = 1,2,\ldots,n$$

Utilizando o método de indução matemática temos:

1) Para $v = 1$ $\quad \dfrac{\hat{m}_{ij}^{(1)} x_{+j}}{\hat{m}_{+j}^{(1)}} = \delta_{ij} a_i^{(1)} b_j^{(1)}$

Ora $$\frac{\hat{m}_{ij}^{(1)} x_{+j}}{\hat{m}_{+j}^{(1)}} = \frac{\hat{m}_{ij}^{(0)} x_{i+}}{\hat{m}_{i+}^{(0)}} \frac{x_{+j}}{\hat{m}_{+j}^{(1)}} = \frac{\hat{m}_{ij}^{(0)} x_{i+}}{\hat{m}_{i+}^{(0)}} \frac{x_{+j}}{\sum_{i=1}^{I} \frac{\hat{m}_{ij}^{(0)} x_{i+}}{\hat{m}_{i+}^{(0)}}}$$

$$= \frac{\delta_{ij} x_{i+}}{\delta_{i+}} \frac{x_{+j}}{\sum_{i=1}^{I} \delta_{ij} \frac{x_{i+}}{\delta_{i+}}} = \delta_{ij} \frac{x_{i+}}{\delta_{i+}} \frac{x_{+j}}{\sum_{i=1}^{I} \delta_{ij} \frac{x_{i+}}{\delta_{i+}}}$$

$$= \delta_{ij} a_i^{(1)} b_j^{(1)}$$

2) Admitindo a igualdade válida para v = n, vamos provar que também é válida para v = n + 1, isto é

$$\frac{\hat{m}_{ij}^{(2n-1)}x_{+j}}{\hat{m}_{+j}^{(2n-1)}} = \delta_{ij}a_{ij}^{(n)}b_j^{(n)} \Rightarrow \frac{\hat{m}_{ij}^{(2n+1)}x_{+j}}{\hat{m}_{+j}^{(2n+1)}} = \delta_{ij}a_i^{(n+1)}b_j^{(n+1)}$$

Pretendemos provar $\frac{\hat{m}_{ij}^{(2n+1)}x_{+j}}{\hat{m}_{+j}^{(2n+1)}} = \delta_{ij}a_i^{(n+1)}b_j^{(n+1)}$ o que é o mesmo que provar

que $\frac{\hat{m}_{ij}^{(2n+1)}x_{+j}}{\hat{m}_{+j}^{(2n+1)}} = \delta_{ij}\frac{x_{i+}}{\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}\frac{x_{+j}}{\sum_{i=1}^{I}\delta_{ij}\frac{x_{i+}}{\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}}$.

Ora $\frac{\hat{m}_{ij}^{(2n+1)}x_{+j}}{\hat{m}_{+j}^{(2n+1)}} = \frac{\hat{m}_{ij}^{(2n)}x_{i+}}{\hat{m}_{i+}^{(2n)}}\frac{x_{+j}}{\hat{m}_{+j}^{(2n+1)}} =$

$$= \frac{\delta_{ij}a_i^{(n)}b_j^{(n)}x_{i+}}{\sum_{j=1}^{J}\delta_{ij}a_i^{(n)}b_j^{(n)}}\frac{x_{+j}}{\sum_{i=1}^{I}\frac{\delta_{ij}a_i^{(n)}b_j^{(n)}x_{i+}}{\sum_{j=1}^{J}\delta_{ij}a_i^{(n)}b_j^{(n)}}} =$$

$$= \frac{\delta_{ij}a_i^{(n)}b_j^{(n)}x_{i+}}{a_i^{(n)}\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}\frac{x_{+j}}{\sum_{i=1}^{I}\frac{\delta_{ij}a_i^{(n)}b_j^{(n)}x_{i+}}{a_i^{(n)}\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}} = \delta_{ij}\frac{x_{i+}}{\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}\frac{b_j^{(n)}x_{+j}}{b_j^{(n)}\sum_{i=1}^{I}\delta_{ij}\frac{x_{i+}}{\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}}$$

$$= \delta_{ij}\frac{x_{i+}}{\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}\frac{x_{+j}}{\sum_{i=1}^{I}\delta_{ij}\frac{x_{i+}}{\sum_{j=1}^{J}\delta_{ij}b_j^{(n)}}} = \delta_{ij}a_i^{(n+1)}b_j^{(n+1)}$$

Portanto estes dois processos são equivalentes. O segundo processo torna-se mais vantajoso se pretendermos estimar os parâmetros $\{a_i\}$ e $\{b_j\}$, ainda que estes

estimadores também se possam obter através do primeiro método. No entanto, em determinadas situações o primeiro processo apresenta mais vantagens, nomeadamente no que se refere às facilidades de programação computacional. Além disso, o segundo processo normalmente requer mais cálculos.

3.5 Ajustamento do modelo

Tal como acontece com os modelos loglineares usuais, o ajustamento do modelo da quasi-independência faz-se através de testes estatísticos. Podemos utilizar o clássico teste do χ^2, pois a teoria assintótica que lhe serve de base é válida também sob o modelo de quasi-independência, seja qual fôr o esquema de amostragem em causa (Poisson ou multinomial). Assim sendo, a estatística de teste será dada por:

$$X^2 = \sum_{(i,j) \in S} \frac{\left(x_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}} \qquad (3.5\text{-}1)$$

Sob a hipótese nula da Quasi-Independência X^2 tem uma distribuição χ^2 cujo número de graus de liberdade é determinado conforme vamos ver a seguir.
Além deste teste pode-se utilizar também o teste de razão de verosimilhanças cuja estatística de teste é dada por

$$Y^2 = 2 \sum_{(i,j) \in S} x_{ij} \log\left(\frac{x_{ij}}{\hat{m}_{ij}}\right) \qquad (3.5\text{-}2)$$

Graus de liberdade para a Quasi-Independência

No modelo de quasi-independência não podemos aplicar a regra geral de cálculo do número de graus de liberdade para um modelo loglinear não saturado, que nos diz simplesmente para calcular a diferença entre o número total de células

estimadas e o número total de parâmetros independentes. Ora, neste modelo verificamos que o número de células estimadas é dado por $N_1 - N_3$ onde N_1 representa as $I \times J$ células da tabela e N_3 representa o número de células que contêm zeros estruturais. Novamente, consideramos $N_2 = 1+(I-1)+(J-1)$ parâmetros independentes tal como em modelos não saturados para tabelas completas, cuja expressão é dada por $\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}$. Portanto seguindo a regra de cálculo temos g.l. = $N_1 - N_2 - N_3$.

3.6 Conectividade e Separabilidade

Quando consideramos a estimação de m_{ij} numa tabela incompleta sob o modelo da quasi-independência e calculamos o número de graus de liberdade a ele associado, há situações como as que vamos descrever de seguida em que devemos considerar o subconjunto incompleto S em partes separadas. Para descrever tais situações precisamos de introduzir os conceitos de conectividade e separabilidade.

Numa tabela de contingência bidimensional duas células dizem-se associadas se não contêm zeros estruturais e se estão ou na mesma linha ou na mesma coluna. Um conjunto de células que não contêm zeros estruturais diz-se conectado se qualquer par de células pode ser ligado por uma cadeia de células onde dois quaisquer membros consecutivos desta cadeia devem estar associados. Finalmente uma tabela diz-se conectada se o seu conjunto de células que não contêm zeros estruturais é conectado. Um tabela incompleta que não é conectada diz-se separável, desde que após cada permutação apropriada de linhas e colunas possamos dividir as células não vazias de uma tabela, no mínimo, em duas tabelas separadas (subrectângulos), onde cada uma das subtabelas não tem linha ou coluna comum com outra.

Como exemplo do que acabou de ser referido consideremos a seguinte tabela incompleta bidimensional 4×4

a	------	-------	b
-----	c	d	-------
-----	e	f	-------
g	------	-------	h

que se separa em duas subtabelas

a	**b**		
g	**h**		
		f	e
		d	c

Foram efectuados os seguintes movimentos pela ordem indicada:

- troca de 4ª com 2ª coluna
- troca de 2ª com 4ª linha

A vantagem principal da separabilidade é a redução do número de graus de liberdade assim como a possibilidade de estudarmos as novas tabelas resultantes, cujo cálculo dos valores esperados das células se torna também bastante mais simples.

Se tivermos uma tabela incompleta separável em k rectângulos, o conjunto S de células não vazias é constituído por k subconjuntos disjuntos $S_1,S_2,...,S_k$, cada um dos quais contendo células não vazias de um dos k subrectângulos separáveis. Sempre que os subconjuntos Sv e Sw ($v \neq w$), não tenham linhas ou colunas comuns, podemos escrever o modelo da quasi-independência 2.1-2 de forma que os valores esperados

das células de Sv não tenham parâmetros comuns com os valores esperados de Sw. Isto quer dizer que temos um modelo de quasi-independência para cada um dos subconjuntos S_l, l = 1,....,k.

Sob os modelos de amostragem que aqui vamos considerar, a função de verosimilhança pode ser escrita como um produto de funções cada uma das quais envolvendo parâmetros de um e um só dos subconjuntos S_l. Por isso, os estimadores de máxima verosimilhança dos valores esperados para um dado subconjunto S_l, são baseados somente nos dados desse subconjunto. Como resultado, tanto para a estimação como para testes, é possível considerar cada subconjunto (subrectângulo) da tabela separadamente .

3.7 Métodos de Estimação Directa para a Quasi-Independência

No que vimos anteriormente não encontramos nenhum método directo para o cálculo dos estimadores de máxima verosimilhança numa subtabela incompleta S. Contudo, há certos tipos de tabelas incompletas onde a configuração dos zeros estruturais é tal que, é possível uma estimação directa para as restantes células. Vamos analisar neste capítulo algumas dessas configurações.

Começamos por descrever duas regras para localizar células não interactivas (células que não produzem efeito na independência ou Quasi-independência). Normalmente o estimador de máxima verosimilhança para estas células é o próprio valor observado (frequência da célula). Se estas células são facilmente localizáveis então podemos removê-las da tabela enquanto analisamos a quasi-independência e recolocamo-las quando a análise estiver completa. A segunda regra que vamos analisar fornece um método para decompor algumas tabelas incompletas em subtabelas cujos estimadores de máxima verosimilhança podem ser determinados separadamente. Descreveremos ainda outras duas regras que nos fornecem métodos directos para o cálculo dos estimadores de máxima verosimilhança (EMVs) em dois tipos diferentes de tabelas incompletas.

Através do uso conjunto das quatro regras podemos calcular directamente os EMVs para todo o tipo de configurações de tabelas incompletas. Vamos então descrever as quatro regras.

No que se segue continuamos a considerar

$$\delta_{ij} = \begin{cases} 1, & (i,j) \in S \\ 0, & \text{caso contrario} \end{cases}$$ onde S é conjunto de células que não contêm zeros estruturais.

Regra 1(Células isoladas)

Se $\delta_{ij}^{..} = 1$ para alguma célula (i,j) mas os restantes $\delta_{ij}^{..}$ da mesma linha ou coluna são todos zeros então $\hat{m}_{ij}^{..} = x_{ij}^{..}$, uma vez que os EMVs são unicamente determinados a partir da preservação dos totais marginais. Diz-se que a célula (i,j) é não interactiva e neste caso simplesmente eliminamos a i-ésima linha (j-ésima coluna) e continuamos o processo de estimação.

Na seguinte tabela bidimensional 3×3, a célula (3,3) é isolada

m_{11}	m_{12}	m_{13}
m_{21}	m_{22}	-----
-----	-----	m_{33}

Se eliminarmos a terceira linha a célula (1,3) fica isolada

m_{11}	m_{12}	m_{13}
m_{21}	m_{22}	-----

e portanto temos ainda uma célula não interactiva que podemos eliminar. Ficamos com uma tabela bidimensional 2×2

m_{11}	m_{12}
m_{21}	m_{22}

Os estimadores de máxima verosimilhança das células desta tabela são agora determinados a partir da expressão referida no capítulo 1 quando é ajustado o modelo de independência em tabelas bidimensionais completas.

Regra 2 (Semiseparabilidade)

Mantel (1970) definiu uma tabela incompleta inseparável como sendo semiseparável se puder ser separada em duas ou mais subtabelas, pela simples remoção de uma linha ou coluna.

Suponhamos agora que uma tabela é semiseparável pela remoção de uma coluna. Particionamos esta tabela em conjuntos de linhas em que a cada um corresponde exactamente uma das subtabelas que resultam da respectiva remoção. Podemos de seguida estimar os valores esperados das células em cada um destes subconjuntos de linhas segundo o modelo da quasi-independência, da mesma maneira que faríamos se, após a eliminação de colunas vazias, cada conjunto fosse uma tabela separável.

Consideremos como exemplo a seguinte tabela bidimensional 5×5

m_{11}	m_{12}			
m_{21}	m_{22}			m_{25}
		m_{33}	m_{34}	m_{35}
		m_{43}	m_{44}	
		m_{53}	m_{54}	m_{55}

Esta tabela é semiseparável pois a simples eliminação da 5ª coluna transforma-a em duas tabelas separáveis:

m_{11}	m_{12}
m_{21}	m_{22}

m_{33}	m_{34}
m_{43}	m_{44}
m_{53}	m_{54}

Vamos analisar as duas subtabelas que resultam do conjunto formado pelas duas primeiras linhas e do conjunto formado pelas restantes linhas. Ficamos assim com duas subtabelas reduzidas:

m_{11}	m_{12}	
m_{21}	m_{22}	m_{25}

(1)

m_{33}	m_{34}	m_{35}
m_{43}	m_{44}	
m_{53}	m_{54}	m_{55}

(2)

Na tabela (1) m_{25} é uma célula isolada pelo que a eliminação desta célula através da regra 1 transforma esta tabela numa tabela bidimensional 2×2 cujos EMVs são facilmente calculáveis. A tabela (2) admite EMVs usando as regras 3 ou 4 que vamos descrever a seguir.

Regra 3 (tabelas Bloco-triangular)

São exemplos de tabelas "bloco-triangular" as seguintes tabelas incompletas:

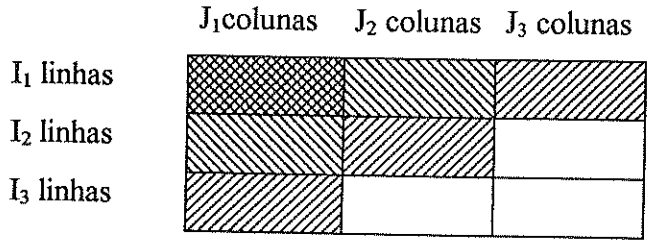
m_{11}	m_{12}	m_{13}	m_{14}
m_{21}	m_{22}	m_{23}	....
m_{31}	m_{32}	....	....
m_{41}	....	....	....

....	....	m_{13}	m_{14}
....	....	m_{23}	m_{24}
m_{31}	m_{32}	m_{33}	m_{34}
m_{41}	m_{42}	m_{43}	m_{44}

m_{11}	m_{12}	m_{13}	m_{14}
....	m_{22}	m_{23}	m_{24}
....	m_{32}	m_{33}	m_{34}
....	....	m_{43}	m_{44}
....	....	m_{53}	m_{54}

Chamamos a este tipo de tabela "bloco-triangular" porque as células que não contêm zeros estruturais depois de permutações de linhas e/ou colunas formam um triângulo rectângulo com blocos de células alinhadas ao longo da hipotenusa. Por conveniência de notação chamamos tabelas "bloco-triangular" às que têm os seus zeros estruturais num dos cantos da tabela. No que se segue vamos calcular os EMVs para um caso particular de tabelas "bloco-triangular". De um modo geral, uma tabela bidimensional I×J, é bloco-triangular se depois da permutação de linhas ou colunas , $\delta_{ij}^{..} = 0$ implica $\delta_{kl} = 0$ para todos $k \geq i$ e $l \geq j$.

Vamos agora deduzir os EMVs para uma tabela bidimensional com I linhas e J colunas, onde $I = I_1 + I_2 + I_3$ linhas e $J = J_1 + J_2 + J_3$ colunas.



Os zeros estruturais ficam nos dois blocos de células completamente especificados colocando $\delta_{ij} = 0$ para $i \geq I_1 + 1$ e $j \geq J_1 + J_2 + 1$ e para $i \geq I_1 + I_2 + 1$ e $j \geq J_1 + 1$.

Recorda-se que, $\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}$

Escrevendo os m_{ij} na forma multiplicativa,observando o bloco diagonal definido por $i = I_1 + 1, I_1 + 2, \ldots, I_1 + I_2$ e $j = J_1 + 1, J_1 + 2, \ldots, J_1 + J_2$, e notando que :

$$m_{i+}m_{+j} = \left(\sum_{l=1}^{J_1+J_2} e^{\mu+\mu_{1(i)}+\mu_{2(l)}} \right) \left(\sum_{k=1}^{I_1+I_2} e^{\mu+\mu_{1(k)}+\mu_{2(j)}} \right)$$

$$= e^{\mu+\mu_{1(i)}+\mu_{2(j)}} \left(\sum_{l=1}^{J_1+J_2} \sum_{k=1}^{I_1+I_2} e^{\mu+\mu_{1(k)}+\mu_{2(l)}} \right)$$

$$= m_{ij} \left(\sum_{k=1}^{I_1+I_2} \sum_{l=1}^{J_1+J_2} m_{kl} \right) \qquad (3.7\text{-}1)$$

Visto que $\sum_{k=1}^{I_1+I_2} \sum_{l=1}^{J_1+J_2} m_{kl} = m_{++} - \sum_{i=I_1+I_2+1}^{I_1+I_2+I_3} m_{i+} - \sum_{j=J_1+J_2+1}^{J_1+J_2+J_3} m_{+j}$ (3.7-2)

podemos escrever os m_{ij} para este bloco como funções directas dos marginais totais e como se verificam as igualdades:

$m_{i+} = x_{i+}$
$m_{+j} = x_{+j}$ os EMVs são dados por

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}}{\sum_{k=1}^{I_1+I_2} \sum_{l=1}^{J_1+J_2} x_{kl}} \qquad i = I_1+1,\ldots,I_1+I_2\ ,\ j = J_1+1,\ldots,J_1+J_2 \qquad (3.7\text{-}3)$$

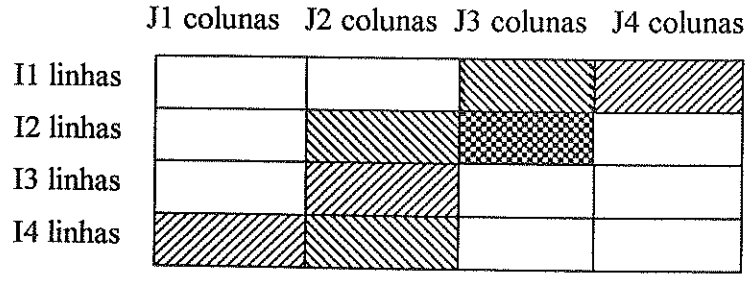
Calculamos para os restantes dois blocos diagonais os EMVs usando fórmulas análogas às anteriores. Desde que tenhamos os EMVs dos blocos da diagonal calculamos de seguida os totais marginais para as restantes células ainda não

estimadas, subtraindo os EMVs dos três blocos diagonais aos totais marginais originais e procedendo como se as células já estimadas fossem zeros estruturais. Isto deixa-nos com uma pequena tabela bloco triangular $(I_1+I_2)\times(J_1+J_2)$ onde as células com zeros estruturais ficam completamente especificadas pela condição $\delta_{ij}=0$ para $i=I_1+1$ e $j=J_1+1$. Procedemos de seguida com a estimação directa para as restantes células efectuando as necessárias alterações nos totais marginais e nas fórmulas (3.7-1) , (3.7-2) e (3.7-3) .

Regra 4 (tabelas com blocos-escada)

Uma tabela incompleta inseparável diz-se "bloco-escada" se depois de permutar linhas e colunas, podemos dividir a tabela em conjuntos de linhas e colunas, cada um dos quais contendo um rectângulo de células não nulas e tal que cada rectângulo partilhe colunas com os rectângulos imediatamente acima ou abaixo. Os EMVs dos valores esperados não nulos são determinados a partir de um modelo de quasi-independência. Quando um dos rectângulos tem somente uma linha ou coluna, essa linha ou coluna pode ter células isoladas ou a tabela pode ser semiseparável.

Vamos de seguida deduzir os EMVs conforme Bishop (1975) para uma tabela 'bloco-escada' bidimensional com I linhas e J colunas onde $I = I_1+I_2+I_3+I_4$ e $J = J_1+J_2+J_3+J_4$.



Em particular vamos deduzir os EMVs para um dos sete blocos de células. Em primeiro lugar consideramos a zona marcada a quadriculado. Temos $i = I_1 + 1, \ldots I_1+I_2$, $j = J_1+J_2+1, \ldots, J_1+J_2+J_3$

Para a célula (i,j) observamos os respectivos produtos dos totais marginais, isto é.,

$$m_{i+}m_{+j} = \left(\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(i)}+\mu_{2(l)}}\right)\left(\sum_{k=1}^{I_1+I_2} e^{\mu+\mu_{1(k)}+\mu_{2(j)}}\right) =$$

$$= e^{\mu+\mu_{1(i)}+\mu_{2(j)}}\left(\sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}}\right)$$

$$= m_{ij}\left[\sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}}\right] \qquad (3.7\text{-}4)$$

e, portanto

$$m_{ij} = \frac{m_{i+}.m_{+j}}{\sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}}} \qquad (3.7\text{-}5)$$

Somando em todas as células (i,j) do bloco vem,

$$\frac{\sum_{i=I_1+1}^{I_1+I_2}\sum_{j=J_1+J_2+1}^{J_1+J_2+J_3} m_{i+}m_{+j}}{\sum_{i=I_1+1}^{I_1+I_2}\sum_{j=J_1+J_2+1}^{J_1+J_2+J_3} m_{ij}} = \sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}} \qquad (3.7\text{-}6)$$

Substituindo em (2.7-5), obtemos os EMVs para as células deste bloco da seguinte forma:

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}\left(\sum_{k=I_1+1}^{I_1+I_2}\sum_{l=J_1+J_2+1}^{J_1+J_2+J_3} x_{kl}\right)}{\sum_{k=I_1+1}^{I_1+I_2}\sum_{l=J_1+J_2+1}^{J_1+J_2+J_3} x_{k+}x_{+l}} \qquad (3.7\text{-}7)$$

$i = I_1 + 1, \ldots, I_1+I_2 \quad j = J_1+J_2+1, \ldots, J_1+J_2+J_3$

Em particular vamos deduzir os EMVs para um dos sete blocos de células. Em primeiro lugar consideramos a zona marcada a quadriculado. Temos $i = I_1 + 1 ,...I_1+I_2$, $j = J_1+J_2+1,..., J_1+J_2+J_3$

Para a célula (i,j) observamos os respectivos produtos dos totais marginais, isto é.,

$$m_{i+}m_{+j} = \left(\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(i)}+\mu_{2(l)}}\right)\left(\sum_{k=1}^{I_1+I_2} e^{\mu+\mu_{1(k)}+\mu_{2(j)}}\right) =$$

$$= e^{\mu+\mu_{1(i)}+\mu_{2(j)}}\left(\sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}}\right)$$

$$= m_{ij}\left[\sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}}\right] \qquad (3.7\text{-}4)$$

e, portanto

$$m_{ij} = \frac{m_{i+}.m_{+j}}{\sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}}} \qquad (3.7\text{-}5)$$

Somando em todas as células (i,j) do bloco vem,

$$\frac{\sum_{i=I_1+1}^{I_1+I_2}\sum_{j=J_1+J_2+1}^{J_1+J_2+J_3} m_{i+}m_{+j}}{\sum_{i=I_1+1}^{I_1+I_2}\sum_{j=J_1+J_2+1}^{J_1+J_2+J_3} m_{ij}} = \sum_{k=1}^{I_1+I_2}\sum_{l=J_1+1}^{J_1+J_2+J_3} e^{\mu+\mu_{1(k)}+\mu_{2(l)}} \qquad (3.7\text{-}6)$$

Substituindo em (2.7-5), obtemos os EMVs para as células deste bloco da seguinte forma:

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}\left(\sum_{k=I_1+1}^{I_1+I_2}\sum_{l=J_1+J_2+1}^{J_1+J_2+J_3} x_{kl}\right)}{\sum_{k=I_1+1}^{I_1+I_2}\sum_{l=J_1+J_2+1}^{J_1+J_2+J_3} x_{k+}x_{+l}} \qquad (3.7\text{-}7)$$

$i = I_1 +1 ,...,I_1+I_2 \quad j = J_1+J_2+1,..., J_1+J_2+J_3$

3.8 Exemplo

Numa escola foram realizadas duas provas numa determinada disciplina. Os resultados dessas provas foram distribuídos por cinco níveis de classificação. O critério de atribuição do nível a cada aluno foi o seguinte: prevalência do nível mais elevado. A tabela seguinte representa uma tabela bidimensional incompleta 5×5, onde as linhas representam as categorias da variável: primeira prova e as colunas representam as categorias da variável: segunda prova. Em cada uma das células figura o número de alunos com a combinação de categorias correspondente.

			2ªprova			
1ªprova	**cinco**	**quatro**	**três**	**dois**	**um**	**total**
um	3	2	8	2	3	18
dois	11	7	4	4	----	26
três	9	10	12	----	----	31
quatro	6	2	----	----	----	8
cinco	5	----	----	----	----	5
total	34	21	24	6	3	88

A disposição da tabela permite-nos verificar que se trata de uma tabela bloco-triangular. Estamos em condições de aplicar a regra três.

Naturalmente, pretendemos testar a hipótese de independência entre as duas provas. Logo o modelo a ajustar é dado por:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} \quad , i = 1,..., 5 \text{ e } j = 1,..., 5$$

Vamos calcular os estimadores das frequências esperadas conforme a regra 4. Então começamos pelas células da diagonal onde $\hat{m}_{55} = 5$ e $\hat{m}_{15} = 3$, porque são células isoladas e pela regra 1, têm valor esperado igual ao valor observado. Obtemos para as restantes células da diagonal:

$$\hat{m}_{42} = \frac{8 \times 21}{88 - (24 + 6 + 3 + 5)} = 3.36$$

$$\hat{m}_{33} = \frac{31 \times 24}{88 - (6+3+5+8)} = 11.27$$

$$\hat{m}_{24} = \frac{26 \times 6}{88 - (3+5+8+31)} = 3.8$$

De seguida subtraímos aos totais marginais estes valores e obtemos a tabela:

			2ªprova		
1ªprova	**cinco**	**quatro**	**três**	**dois**	**total**
um	3	2	8	2	15
dois	11	7	4	----	22.2
três	9	10	----	----	19.73
quatro	6	----	----	----	4.64
total	29	17.64	12.73	2.2	61.57

Calculamos as estimativas das células da diagonal usando o mesmo raciocínio:

$$\hat{m}_{41} = 4.64 \ ; \ \hat{m}_{32} = 8.29 \ ; \ \hat{m}_{23} = 8.07 \ ; \ \hat{m}_{14} = 2.2$$

Subtraímos novamente aos totais marginais estes valores e obtemos a tabela:

		2ªprova		
1ªprova	**cinco**	**quatro**	**três**	**total**
um	3	2	8	12.8
dois	11	7		14.13
três	9		----	11.44
total	24.36	9.35	4.66	38.37

As estimativas da diagonal principal são dadas por:

$$\hat{m}_{31} = 11.44 \ ; \ \hat{m}_{22} = 5.93 \ ; \ \hat{m}_{13} = 4.66$$

Subtraíndo estes valores aos totais marginais obtemos a seguinte tabela:

		2ªprova	
1ªprova	**cinco**	**quatro**	**total**
um	3	2	8.14
dois	11		8.2
total	12.92	3.42	16.34

As estimativas das células da diagonal são dadas por: $\hat{m}_{12} = 3.42$; $\hat{m}_{21} = 8.2$.

Subtraindo-as aos totais marginais obtemos a seguinte tabela:

	2ªprova	
1ªprova	cinco	total
um	3	4.72
total	4.72	

Obtemos a última estimativa $\hat{m}_{11} = 4.72$

A tabela completa das estimativas das frequências esperadas fica assim completa:

			2ªprova			
1ªprova	cinco	Quatro	três	dois	um	total
um	4.72	3.42	4.66	2.2	3	18
dois	8.2	5.93	8.07	3.8	----	26
três	11.44	8.29	11.27	----	----	31
quatro	4.64	3.36	----	----	----	8
cinco	5	----	----	----	----	5
total	34	21	24	6	3	88

O valor da estatística de teste é dado por $X^2 = 8.71$. O número de graus de liberdade é dado por g.l = 25 - (1 + 4 + 4) - 10 = 6. Considerando o nível de significância α = 0.05 (o valor do quantil de probabilidade 0.05 do $\chi^2_{(6)} = 12.592$), verificamos que não podemos rejeitar a hipótese de independência entre as duas provas.

4. Tabelas Incompletas Multidimensionais

4.1 Apresentação do modelo

Depois de termos analisado tabelas bidimensionais incompletas, vamos de seguida analisar tabelas incompletas com dimensão superior a dois. Concretamente analisaremos tabelas incompletas tridimenionais.

Consideremos, novamente o conjunto S das células de uma tabela incompleta $I \times J \times K$ que é constituído por todas as células que não contêm zeros estruturais. Sendo m_{ijk} o número esperado de indíviduos na célula (i,j,k), tem-se $m_{ijk} = 0$ para $(i,j,k) \notin S$.

Consideramos o modelo loglinear saturado usual:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)} \quad (4.1\text{-}1)$$

$$i = 1,...,I \quad j = 1,...,J \quad k = 1,...,K$$

para as células de S com as seguintes condições:

$$\sum_i \delta_i^{(23)} \mu_{1(i)} = \sum_i \delta_{ij}^{(3)} \mu_{12(ij)} = \sum_i \delta_{ik}^{(2)} \mu_{13(ik)} = \sum_i \delta_{ijk} \mu_{123(ijk)} = 0 \quad (4.1\text{-}2)$$

$$\sum_j \delta_j^{(13)} \mu_{2(j)} = \sum_j \delta_{jk}^{(1)} \mu_{23(jk)} = \sum_k \delta_k^{(12)} \mu_{3(k)} = 0 \quad (4.1\text{-}3)$$

com $$\delta_{ijk} = \begin{cases} 1, & (i,j,k) \in S \\ 0, & (i,j,k) \notin S \end{cases} \quad (4.1\text{-}4)$$

$$\delta_{ij}^{(3)} = \begin{cases} 1, & \delta_{ijk} = 1 \text{ para algum } k \\ 0, & \text{caso contrario} \end{cases} \quad (4.1\text{-}5)$$

$$\delta_i^{(23)} = \begin{cases} 1, & \delta_{ijk} = 1 \text{ para algum } (j,k) \\ 0, & \text{caso contrario} \end{cases} \quad (4.1\text{-}6)$$

$$\delta_{ik}^{(2)} = \begin{cases} 1, & \delta_{ijk} = 1 \text{ para algum } j \\ 0, & \text{caso contrario} \end{cases} \quad (4.1\text{-}7)$$

$$\delta_{jk}^{(1)} = \begin{cases} 1, \ \delta_{ijk} = 1 \text{ para algum } i \\ 0, \text{ caso contrario} \end{cases} \qquad (4.1\text{-}8)$$

$$\delta_{j}^{(13)} = \begin{cases} 1, \ \delta_{ijk} = 1 \text{ para algum } (i,k) \\ 0, \text{ caso contrario} \end{cases} \qquad (4.1\text{-}9)$$

$$\delta_{k}^{(12)} = \begin{cases} 1, \ \delta_{ijk} = 1 \text{ para algum } (i,j) \\ 0, \text{ caso contrario} \end{cases} \qquad (4.1\text{-}10)$$

4.2 Separabilidade a três dimensões

Antes de passarmos ao desenvolvimento do modelo que acabamos de enunciar faz aqui sentido retomar o tema da separabilidade em tabelas incompletas. No caso das tabelas incompletas bidimensionais era relativamente simples lidar com os conceitos de conectividade e separabilidade pois só tinhamos duas dimensões e um só modelo. No que vamos desenvolver a seguir não é assim tão simples, uma vez que os conceitos referidos dependem da dimensão da tabela e do modelo a ajustar.

Vamos ilustrar em seguida a separabilidade a três dimensões. Para dimensões superiores as ideias básicas são as mesmas.

Definimos a célula (i,j,k) na tabela $I \times J \times K$ tendo 1ª coordenada i, 2ª coordenada j e 3ª coordenada k. Seja D um subconjunto não vazio do conjunto de inteiros $\{1,2,3\}$. Dizemos que duas células são D-associadas se não contêm zeros estruturais e se as suas coordenadas correspondentes ao subconjunto D coincidem. Dizemos então que um conjunto de células que não contêm zeros estruturais é (D_1,D_2,D_3) conectado se qualquer célula pode ser ligada a outra por uma cadeia de células, devendo quaisquer dois membros consecutivos desta cadeia estar D_1-associados, D_2-associados ou D_3-associados. Qualquer conjunto de células que não contêm zeros estruturais é (D_1, D_2, D_3) separável se não é (D_1, D_2, D_3) - conectado.

Considerando um conjunto (D_1, D_2, D_3) separável, podemos dividi-lo em subconjuntos onde cada um é (D_1, D_2, D_3) conectado. Mas quando dois destes subconjuntos são combinados deixam de ser conectados. Estes subconjuntos constituem as chamadas componentes separáveis do conjunto original.

De uma forma geral, qualquer tabela completa é por si própria separável, e pode ser repartida em subtabelas. Então o modelo loglinear para a tabela completa pode ser separado em modelos loglineares paralelos para cada uma das subtabelas.

Portanto uma tabela que por exemplo é $(\{1,3\},\{2\})$ separável, quer dizer que o conjunto das células da tabela que não contêm zeros estruturais é separado em dois conjuntos onde no primeiro, considerado qualquer par de células, estas têm a 1ª e 3ª coordenadas comuns, enquanto que no segundo conjunto, considerado qualquer par de células, estas têm a 2ª coordenada comum.

Ora neste caso faz sentido que os termos μ_{12} , μ_{23} e μ_{123} sejam removidos do modelo loglinear saturado usual, o que equivale a afirmar que estamos a ajustar um modelo através de testes estatísticos, cuja hipótese nula é H_0: $\mu_{12} = \mu_{23} = \mu_{123} = 0$.

Alguns exemplos de separabilidade em tabelas tridimensionais são a seguir apresentados, onde os traços representam zeros estruturais:

1 °

m_{111}	m_{121}	-------		m_{112}	m_{122}	-------		-------	-------	-------
m_{211}	m_{221}	-------		m_{212}	m_{222}	-------		-------	-------	-------
-------	-------	-------		-------	-------	-------		-------	-------	m_{333}

É uma tabela $(\{1\},\{2\},\{3\})$ separável e por isso separável para o modelo especificado por $\mu_{12} = \mu_{13} = \mu_{23} = \mu_{123} = 0$.

2°

m_{111}	m_{121}	-------		m_{112}	m_{122}	-------		-------	-------	-------
m_{211}	m_{221}	-------		m_{212}	m_{222}	-------		-------	-------	-------
-------	-------	m_{331}		-------	-------	m_{332}		-------	-------	m_{333}

Esta tabela é tanto $(\{1,3\},\{2\})$ separável como $(\{2,3\},\{1\})$ separável , e por isso separável para o modelos especificados por $\mu_{12} = \mu_{23} = \mu_{123} = 0$ e por $\mu_{12} = \mu_{13} = \mu_{123} = 0$.

3°

m_{111}	m_{121}	-------		m_{112}	m_{122}	-------		-------	-------	m_{133}
m_{211}	m_{221}	-------		m_{212}	m_{222}	-------		-------	-------	m_{233}
-------	-------	m_{331}		-------	-------	m_{332}		-------	-------	m_{333}

Esta tabela é $(\{1,3\},\{2\})$ separável mas não é $(\{2,3\},\{1\})$ nem $(\{1,2\}.\{3\})$ separável.

4.3 Estimadores de máxima verosimilhança e graus de liberdade

As condições gerais de existência dos estimadores de máxima verosimilhança não são muito diferentes do caso bidimensional conforme vamos ver de seguida.

Numa tabela tridimensional $I \times J \times K$ sob o modelo loglinear não saturado

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \qquad (4.3\text{-}1)$$

os estimadores de máxima verosimilhança vão ser obtidos tomando em conta as seguintes restrições :

$$\hat{m}_{ij+} = x_{ij+}$$
$$\hat{m}_{i+k} = x_{i+k} \qquad (4.3\text{ -}2)$$
$$\hat{m}_{+jk} = x_{+jk}$$

$$i = 1,\ldots, I \quad j = 1,\ldots, J \quad k = 1,\ldots, K$$

4.3.1 Método iterativo de estimação

Seguindo um caminho semelhante ao caso bidimensional, calculamos os estimadores de máxima verosimilhança aplicando o método iterativo de ajustamento proporcional de Deming-Stephan, que desenvolvemos da maneira seguinte:

1. No passo 0, tomamos $m_{ijk}^{(0)} = \delta_{ijk}$ (4.3.1-1)

onde $\delta_{ijk} = \begin{cases} 1, & (i,j,k) \in S \\ 0, & \text{caso contrario} \end{cases}$

2. No v-ésimo ciclo temos:

$$m_{ijk}^{(3v-2)} = \frac{m_{ijk}^{(3v-3)} x_{ij+}}{\sum_k m_{ijk}^{(3v-3)}}$$

$$m_{ijk}^{(3v-1)} = \frac{m_{ijk}^{(3v-2)} x_{+jk}}{\sum_i m_{ijk}^{(3v-2)}} \qquad (4.3.1\text{-}2)$$

$$m_{ijk}^{(3v)} = \frac{m_{ijk}^{(3v-1)} x_{i+k}}{\sum_j m_{ijk}^{(3v-1)}}$$

3. Continua-se o processo iterativo até atingir a aproximação desejada.

3.3.2 Graus de liberdade

A regra de cálculo dos graus de liberdade para tabelas incompletas segue um raciocínio semelhante ao caso bidimensional.

Considerando novamente uma tabela tridimensional $I \times J \times K$ sob o modelo (4.3-1), procedemos da seguinte forma:

1° - Contamos o número de parâmetros independentes estimados - E

2° - Calculamos $V = I.J.K - E$

3° - Contamos o número de zeros estruturais - Ze

4° - Calculamos $Z_p = Z_{12} + Z_{23} + Z_{13}$ onde

$$Z_{12} = I.J - \sum_{ij} \delta_{ij}^{(3)}$$

$$Z_{23} = J.K - \sum_{jk} \delta_{jk}^{(1)}$$

$$Z_{13} = I.K - \sum_{ik} \delta_{ik}^{(2)}$$

Z_{12}, Z_{23} e Z_{13} representam respectivamente os números de zeros das configurações esperadas marginais $C_{12}^{*} = \{m_{ij+}\}$, $C_{13}^{*} = \{m_{i+k}\}$ e $C_{23}^{*} = \{m_{+jk}\}$.

Finalmente calculamos o n° de graus de liberdade para o modelo em causa:

$$G.L. = V - Z_e + Z_p \qquad (4.3.2\text{-}1)$$

Além disso, seja

$$\sum_{i} \delta_{i}^{(23)} = I,\ \sum_{j} \delta_{j}^{(13)} = J \quad \text{e} \quad \sum_{k} \delta_{k}^{(12)} = K \qquad (4.3.2\text{-}2)$$

porque sendo o modelo (3.3-1) a tabela correspondente tem pelo menos respectivamente: um par (j,k) em cada linha i, um par (i,k) em cada coluna j e um par (i,j) em cada estrato k, correspondente a uma célula que não constitui zero estrutural.

Então substituindo em (3.3.2-1) o número de graus de liberdade é dado por

$$(I-1)(J-1)(K-1) - Z_e + (Z_{12} + Z_{23} + Z_{13}) \qquad (4.3.2\text{-}3)$$

Se a mesma tabela é inseparável para o modelo $\mu_{12} = \mu_{123} = 0$,então efectuando novamente a substituição em (3.3.2-1) o número de graus de liberdade é dado por

$$K(I-1)(J-1) - Z_e + (Z_{12} + Z_{23} + Z_{13}) \qquad (4.3.2\text{-}4)$$

Se há um estrato nulo na tabela devemos adicionar um grau de liberdade a (4.3.2-3) e (4.3.2-4). Por outro lado, se há uma linha vazia na tabela, adicionamos um grau de liberdade a (4.3.2-3) mas não a (4.3.2-4).

4.4 Exemplo

Para ilustrar o que acabamos de desenvolver vamos considerar a seguinte tabela apresentada no 3° exemplo da página 50:

m_{111}	m_{121}	-----	m_{112}	m_{122}	-------	--------	------	m_{133}
m_{211}	m_{221}	-----	m_{212}	m_{222}	------	--------	------	m_{233}
------	--------	m_{331}	-----	------	m_{332}	-------	------	m_{333}

Para este caso observamos uma tabela $(\{1,3\},\{2\})$-separável para o modelo:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ik)}$$

Para este modelo o número de zeros na configuração marginal esperada $C_{13}^{*} = \{m_{i+k}\}$ é $Z_{13} = 0$, e há $Z_e = 14$ zeros estruturais. Se não considerarmos a separabilidade da tabela o número de graus de liberdade é dado por:

$$IJK - Z_e - [1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) - Z_{13}] = 2 \qquad (4.4\text{-}1)$$

Há, no entanto, duas componentes separáveis da tabela para este modelo:

m_{111}	m_{121}
m_{211}	m_{221}

m_{112}	m_{122}
m_{212}	m_{222}

e

------	--------	m_{133}
------	--------	m_{233}
m_{331}	m_{332}	m_{333}

A primeira componente é uma tabela completa $2 \times 2 \times 2$ e tem três graus de liberdade. Pois, considerando para esta componente o modelo

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ik)} ,$$

$$i = 1,2 \quad j = 1,2 \quad k = 1,2$$

verificamos que o número de graus de liberdade é dado por g.l. = I . J . K - (1 + (I-1) + (J-1) + (k-1) + (I-1)(k-1)) = 3 (I = 2, J = 2, K = 2) .

A segunda componente é bidimensional e tem zero graus de liberdade, uma vez que μ_{13} é o termo de interacção para a tabela bidimensional. Pois, considerando o mesmo modelo, o número de graus de liberdade é zero porque observando a seguinte tabela:

Parâmetros	Número de parâmetros independentes
μ	1
$\mu_{1(i)}$	2
$\mu_{3(k)}$	2
$\mu_{13(ik)}$	0

verificamos que o número total de parâmetros independentes é 5. Logo g.l. = 9 - 5 - 4

Portanto o número total de graus de liberdade é 3 + 0 = 3 e não 2.

4.5 Ajustamento do modelo

O ajustamento do modelo deve sempre tomar em consideração as características da tabela de forma que ao estudarmos a sua separabilidade, poderemos numa primeira fase tentar ajustar um determinado modelo.

Sendo a tabela separável em componentes a análise do modelo é feita em duas fases distintas: análise da primeira componente e depois a análise da segunda componente.

Retomando o exemplo anterior 3.4, temos em análise duas tabelas: uma tabela tridimensional completa $2 \times 2 \times 2$ e uma tabela bidimensional 3×3 incompleta com quatro zeros estruturais.

O modelo a ajustar na primeira tabela é dado por:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ik)}$$

$$i = 1,2 \quad j = 1,2 \quad k = 1,2$$

A hipótese a testar neste modelo é $H_0: \mu_{12} = \mu_{23} = \mu_{123} = 0$. Temos em causa um modelo de independência parcial entre as duas variáveis (variável de níveis i, variável de níveis k) e a variável de níveis j. Neste caso os estimadores das frequências esperadas são obtidos de forma directa através da fórmula:

$$\hat{m}_{ijk} = \frac{x_{i+k} x_{+j+}}{N}$$

O modelo a ajustar à segunda tabela é dado por:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{13(ik)}$$

$$i = 1,2,3 \quad j = 3 \quad k = 1,2,3$$

Trata-se neste caso de um modelo loglinear saturado para uma tabela bidimensional. Os estimadores das frequências esperadas são dados por:

$\hat{m}_{133} = x_{13+}$; $\hat{m}_{233} = x_{23+}$; $\hat{m}_{331} = x_{+31}$; $\hat{m}_{332} = x_{+32}$;

$\hat{m}_{333} = N - \hat{m}_{133} - \hat{m}_{233}$ ou $\hat{m}_{333} = N - \hat{m}_{331} - \hat{m}_{332}$.

Os quatro primeiros estimadores são obtidos à custa dos totais marginais, uma vez que na tabela existem quatro zeros estruturais. Depois de calculados os estimadores

para as duas tabelas calculamos a estatística X^2 que segue uma distribuição χ^2 com três graus de liberdade.

CAPÍTULO III

Modelos loglineares para variáveis ordinais

1. Introdução

Os modelos loglineares usuais tratam todas as variáveis como se estas fossem nominais. No entanto, por vezes, deparamos com situações onde as categorias de algumas dessas variáveis estão naturalmente ordenadas em consonância com alguma escala subjacente. O modelo loglinear não consegue traduzir concisamente a informação contida nestas variáveis e na relação entre as suas categorias. Isto, devido à falta de flexibilidade adequada para descrever associações e interacções inerentes à natureza ordinal das variáveis.

Neste capítulo consideraremos, de uma forma introdutória, alguns modelos para tabelas bidimensionais e tridimensionais que exploram a ordem das variáveis através da atribuição de 'scores ' às respectivas categorias. Este processo de ordenação quantitativa das variáveis objectiva explicitar uma escala intervalar subjacente, com os 'scores' escolhidos reflectindo distâncias assumidas entre pontos médios nessa escala.

Na prática a escolha dos 'scores' é feita na perspectiva da simplificação da interpretação do modelo para a descrição dos dados ordinais, por isso é usual

escolher para os 'scores', inteiros definidos pelos próprios índices das categorias ordenados de forma crescente.

No que se segue, analisaremos teoricamente o ajustamento de modelos loglineares a tabelas bidimensionais e de seguida a tabelas tridimensionais conforme Agresti(1984).

2. Tabelas bidimensionais

No caso das tabelas bidimensionais, vamos abordar duas situações: (a_1) uma variável é ordinal e a outra é nominal e (a_2) as duas variáveis são ordinais.

Consideramos novamente o modelo loglinear saturado definido por:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)} \ , \ i = 1,...,I \quad j = 1,...,J \qquad (2\text{-}1)$$

(a_1) Consideremos o caso de uma tabela bidimensional $I \times J$ onde uma das variáveis X ou Y, é ordinal. Comecemos por considerar Y ordinal e designemos por $\{v_j\}$ os 'scores' atribuídos às suas categorias. Normalmente consideram-se $\{v_j\}$ inteiros e ordenados de forma que $v_1 < v_2 < ... < v_J$, para reflectir a ordem das colunas.

No modelo loglinear saturado o termo de interacção de 1ª ordem engloba $(I-1)(J-1)$ parâmetros independentes, que se designam por $\mu_{12(ij)}$. Vamos tomar em consideração este termo para reflectir a natureza ordinal de Y, traduzindo-o da seguinte forma: $\mu_{12(ij)} = \tau_i (v_j - \bar{v})$

onde $\bar{v} = \dfrac{\sum_j v_j}{J}$ e τ_i são parâmetros relacionados com as categorias de X, satisfazendo $\sum_i \tau_i = 0$. Obtemos assim o modelo loglinear

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \tau_i(v_j - \bar{v}) \quad i = 1,...,I \ ; \ j = 1,...,J \qquad (2\text{-}2)$$

com $\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = \sum_i \tau_i = 0$.

Este modelo possui (I-1)(J-2) graus de liberdade pois o número de parâmetros independentes é 1+ (I-1) + (J-1) + (I-1) = 1 + 2(I-1) + (J-1) e por isso o número total de graus de liberdade associado ao modelo é dado por:

$$g.l. = IJ - [1 + 2(I-1) + (J-1)] = (I-1)(J-2) \ .$$

Portanto, possui menos (I-1)(J-2) parâmetros independentes que o modelo loglinear saturado e mais (I-1) parâmetros independentes que o modelo de independência.

Os $\{\tau_i\}$ representam parâmetros adicionais que estão ligados directamente aos efeitos linha. Por isso se designa este modelo pomodelo de efeitos de linha

Explicitando para uma linha particular i, o desvio do $\log(m_{ij})$ face ao modelo de independência é uma função linear de Y com declive τ_i, pois considerando o modelo de independência definido por:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} \quad i = 1,...,I \ ; \ j = 1,...,J$$

e o modelo de efeitos de linha definido por:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \tau_i(v_j - \bar{v}) \quad i = 1,...,I \ ; \ j = 1,...,J$$

o desvio do $\log(m_{ij})$ é dado por $d = \tau_i(v_j - \bar{v})$ que constitui uma função linear de Y uma vez que esta expressão é dependente dos seus scores centrados. Se $\tau_i > 0$, a probabilidade de classificar um elemento da linha i em categorias acima de $\bar{v}$ tem um valor ao que teria no caso das variáveis serem independentes.

Para compreendermos claramente este modelo teremos de interpretar os parâmetros τ_i. Consideremos um par arbitrário de linhas (a,b) com a < b e um par arbitrário de colunas (c,d) com c < d, então estes parâmetros τ_i podem ser estimados a partir de: $\log\left(\frac{\hat{m}_{ac}.\hat{m}_{bd}}{\hat{m}_{ad}.\hat{m}_{bc}}\right) = (\hat{\tau}_b - \hat{\tau}_a)(v_d - v_c)$ (2-3)

Consideremos agora o caso em que a variável X das linhas é ordinal. Sejam $\{\lambda_i\}$ os 'scores' atribuídos às categorias de X. O modelo fica então:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \tau_j(\lambda_i - \bar{\lambda}) \quad i = 1,...,I \; ; \; j = 1,...,J \qquad (2\text{-}4)$$

Este modelo é designado por modelo de efeitos de coluna, cuja análise é análoga à do modelo anterior efectuando a troca de linhas por colunas.

(a_2) Por último consideremos o caso em que ambas as variáveis X e Y são ordinais. Naturalmente vamos impôr uma estrutura diferente ao termo de interacção de primeira ordem $\mu_{12(ij)}$. Com efeito, sejam $\{\lambda_i\}$ e $\{v_j\}$ os 'scores' associados às categorias de X e Y. Sabemos, através do modelo de efeitos de linha, que o desvio da independência é uma função linear de Y para X fixo e, através do modelo de efeitos de coluna, que o desvio da independência é uma função linear de X para Y fixo. Ligando estes dois aspectos obtemos o seguinte modelo:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \beta(\lambda_i - \bar{\lambda})(v_j - \bar{v}) \quad i = 1,...,I \quad j = 1,...,J \qquad (2\text{-}5)$$

onde $\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = 0$. Relativamente ao modelo loglinear saturado o termo de interacção $\mu_{12(ij)}$ é substituído por $\beta(\lambda_i - \bar{\lambda})(v_j - \bar{v})$.

Este modelo designa-se por modelo de associação linear por linear. Ele tem apenas mais um parâmetro independente (β) que o modelo de independência. Assim o número de graus de liberdade associado é (I-1)(J-1)-1 = I.J - I - J .

O parâmetro β designa-se por parâmetro de associação entre X e Y. Se $\beta > 0$, as frequências esperadas das células do canto superior esquerdo da tabela e inferior direito da mesma, são maiores do que seriam se X e Y fossem independentes.

Tal como no caso anterior é necessário compreender o significado do parâmetro β. Consideremos um par arbitrário de linhas (a,b) com a < b e um par

arbitrário de colunas (c,d) com c < d, então este parâmetro pode ser estimado a partir de: $\log\left(\frac{\hat{m}_{ac}\hat{m}_{bd}}{\hat{m}_{ad}\hat{m}_{bc}}\right) = \hat{\beta}(\lambda_b - \lambda_a)(v_d - v_c)$ (2-6)

2.1 Estimação do modelo

Infelizmente não há um método directo para o cálculo dos EMVs $\{\hat{m}_{ij}\}$ de $\{m_{ij}\}$ no modelo anterior. Sob as condições normais de amostragem, os estimadores satisfazem as equações de máxima verosimilhança :

$$\begin{aligned} &\hat{m}_{i+} = x_{i+}, i = 1,\ldots,I \\ &\hat{m}_{+j} = x_{+j}, j = 1,\ldots,J \\ &\sum_i \sum_j \lambda_i v_j \hat{m}_{ij} = \sum_i \sum_j \lambda_i v_j x_{ij} \end{aligned} \quad (2.1\text{-}1)$$

As duas primeiras equações indicam que $\{\hat{m}_{ij}\}$ têm os mesmos totais linha e coluna que os $\{x_{ij}\}$. Sabemos que $p_{ij} = \frac{x_{ij}}{n}$ e $\hat{\pi}_{ij} = \frac{\hat{m}_{ij}}{n}$ representam os estimadores da probabilidade π_{ij} para os dados observados e segundo o modelo de associação linear por linear, respectivamente. Então a terceira equação de máxima verosimilhança pode ser expressa por $\sum_i \sum_j \lambda_i v_j \hat{\pi}_{ij} = \sum_i \sum_j \lambda_i v_j p_{ij}$.

Para o cálculo destes estimadores Goodman (1979) fornece uma solução iterativa usando uma aplicação unidimensional do método de Newton-Raphson. Um método alternativo é seguido como uma aplicação do corolário 2 do teorema 1 de Darroch and Ratcliff (1972) . Este último é o que aqui vamos aplicar.

Para o modelo de associação linear por linear um ciclo simples tem três passos . Se $m_{ij}^{(t)}$ representa uma aproximação de $\hat{m}_{ij}$ num certo estádio, então os três

passos são (para todo i e j)

$$\hat{m}_{ij}^{(t+1)} = \left(\frac{x_{i+}}{\hat{m}_{i+}^{(t)}}\right)\hat{m}_{ij}^{(t)} \qquad (2.1\text{-}2)$$

$$\hat{m}_{ij}^{(t+2)} = \left(\frac{x_{+j}}{\hat{m}_{+j}^{(t+1)}}\right)\hat{m}_{ij}^{(t+1)}$$

$$\hat{m}_{ij}^{(t+3)} = \left(\frac{\sum_{(a,b)}\lambda_a^* v_b^* x_{ab}}{\sum_{(a,b)}\lambda_a^* v_b^* \hat{m}_{ab}^{(t+2)}}\right)^{\lambda_i^* v_j^*} \left(\frac{\sum_{(a,b)}\left(1-\lambda_a^* v_b^*\right)x_{ab}}{\sum_{(a,b)}\left(1-\lambda_a^* v_b^*\right)\hat{m}_{ab}^{(t+2)}}\right)^{1-\lambda_i^* v_j^*} \hat{m}_{ij}^{(t+2)}$$

onde $\{\lambda_i^*\}$ e $\{v_j^*\}$ são reescalonamentos lineares dos scores linha e coluna que satisfazem as condições $0 \le \lambda_i^* \le 1$ e $0 \le v_j^* \le 1$. Para scores inteiros $\{\lambda_i = i\}$ e $\{v_j = j\}$, no terceiro passo podemos tomar $\lambda_i^* = \frac{i-1}{I-1}$ e $v_j^* = \frac{j-1}{J-1}$. Este método é relativamente simples, mas o processo de convergência pode ser bastante lento. Para grandes tabelas pode levar algumas centenas de ciclos até se atingir a convergência adequada.

Para o modelo de efeitos linha dado pela expressão:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \tau_i(v_j - \bar{v}) \quad i = 1,...,I \;;\; j = 1,...,J$$

os estimadores das frequências esperadas deverão satisfazer as seguintes equações de verosimilhança:

$$\begin{aligned} &\hat{m}_{i+} = x_{i+} \;,\; i = 1,...,I \\ &\hat{m}_{+j} = x_{+j} \;,\; j = 1,...,J \\ &\sum_j v_j \hat{m}_{ij} = \sum_j v_j x_{ij} \;,\; i = 1,...,I \end{aligned} \qquad (2.1\text{-}3)$$

Tal como no modelo anterior não existe uma estimação directa, por isso, usa-se novamente o método iterativo sugerido por Darroch e Ractcliff(1972).

Vamos aqui, ilustrar a aplicação deste método para este modelo.

Para tabelas bidimensionais temos o seguinte desenvolvimento.

Consideremos k conjuntos de restrições onde o l-ésimo conjunto tem a seguinte forma:

$$\sum_i \sum_j a_{s(ij)}^{(l)} \hat{m}_{ij} = h_s^{(l)} \quad , s = 1,..., d(l) \tag{2.1-4}$$

onde , $\sum_{s=1}^{d(l)} a_{s(ij)}^{(l)} = 1$, $\sum_{s=1}^{d(l)} h_s^{(l)} = 1$ e $a_{s(ij)}^{(l)} \geq 0$, $h_s^{(l)} > 0$.

$$\text{Define-se } \hat{m}_{ij}^{(t+1)} = \hat{m}_{ij}^{(t)} \prod_{s=1}^{d(t_k)} \left(\frac{h_s^{(t_k)}}{h_s^{(t_k,t)}} \right)^{a_{s(ij)}^{(t_k)}} , t \geq 0 \tag{2.1-5}$$

onde t_k é o resto da divisão inteira de t por k e $h_s^{(l,t)} = \sum_i \sum_j a_{s(ij)}^{(l)} \hat{m}_{ij}^{(t)}$.

Vamos ilustrar este método usando o modelo de efeitos de linha.

O primeiro conjunto de restrições representa o primeiro conjunto de equações de máxima verosimilhança e define-se por:

$$\sum_i \sum_j a_{s(ij)}^{(1)} \hat{m}_{ij} = h_s^{(1)} \quad , s = 1,..., I \tag{2.1-6}$$

onde $a_{s(ij)}^{(1)} = 1$, $i = s$, $j = 1,..., J$ e $h_s^{(1)} = x_{s+}$.

$= 0$, caso contrário

Há $d(1) = I$ equações neste primeiro conjunto.

O segundo conjunto de equações de verosimilhança é expresso a partir de:

$$\sum_i \sum_j a_{s(ij)}^{(2)} \hat{m}_{ij} = h_s^{(2)} \quad , s = 1,..., J \tag{2.1-7}$$

onde $a_{s(ij)}^{(2)} = 1$, $j = s$, $i = 1,..., I$ e $h_s^{(2)} = x_{+s}$.

$= 0$, caso contrário

Há $d(2) = J$ equações neste segundo conjunto.

O terceiro conjunto de equações de verosimilhança é expresso a partir de:

$$\sum_I \sum_J a^{(3)}_{s(ij)} \hat{m}_{ij} = h^{(3)}_s \quad , s = 1,..., I \tag{2.1-8}$$

onde $a^{(3)}_{s(ij)} = v^*_j$, $i = s$, $j = 1,..., J$ e $h^{(3)}_s = \sum_j x_{sj} v^*_j$.

$= 0$, caso contrário

Os $\left\{v^*_j\right\}$ são reescalonamentos lineares dos $\left\{v_j\right\}$ no intervalo $[0,1]$. Como $\sum_s a^{(3)}_{s(ij)} = 1$ e $\sum_s h^{(3)}_s = x_{++}$, adicionamos ao terceiro conjunto de equações, I equações e obtemos a seguinte expressão:

$$\sum_i \sum_j a^{(3)}_{s(ij)} \hat{m}_j = h^{(3)}_s \quad , \; s = I+1,..., 2I \tag{2.1-9}$$

onde $a^{(3)}_{s(ij)} = 1 - v^*_j$, $i = s - I$, $j = 1,..., J$ e $h^{(3)}_s = \sum_j x_{sj}\left(1 - v^*_j\right)$

$= 0$, caso contrário

Há neste terceiro conjunto $d(3) = 2I$ equações.

Substituindo estas expressões em (2.1-5), obtemos as três seguintes expressões que constitem um ciclo.

$$\hat{m}^{(t+1)}_{ij} = \left(\frac{x_{i+}}{\hat{m}^{(t)}_{i+}}\right) \hat{m}^{(t)}_{ij}$$

$$\hat{m}^{(t+2)}_{ij} = \left(\frac{x_{+j}}{\hat{m}^{(t+1)}_{+j}}\right) \hat{m}^{(t+1)}_{ij} \tag{2.1-10}$$

$$\hat{m}^{(t+3)}_{ij} = \left(\frac{\sum_b v^*_b x_{ib}}{\sum_b v^*_b \hat{m}^{(t+2)}_{ib}}\right)^{v^*_b} \left(\frac{\sum_b \left(1 - v^*_b\right) x_{ib}}{\sum_b \left(1 - v^*_b\right) \hat{m}^{(t+2)}_{ib}}\right)^{1 - v^*_j} . \hat{m}^{(t+2)}_{ij}$$

onde $v_b^* = \frac{b-1}{J-1}$ e $v_j^* = \frac{j-1}{J-1}$.

3. Tabelas Tridimensionais

Os modelos anteriormente analisados para tabelas bidimensionais podem ser generalizados a tabelas com dimensão superior. Vamos ilustrar esta situação através da análise de tabelas tridimensionais $I \times J \times K$.

Consideramos uma tabela tridimensional $I \times J \times K$, cujas frequências esperadas designamos por m_{ijk}.

O modelo loglinear saturado usual é dado por:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)}$$

$$i = 1,....I \quad j = 1,...,J \quad k = 1,...,K$$

Nos modelos que vamos desenvolver a seguir temos uma ou mais variáveis ordinais. Os termos de interacção do modelo usual são substituídos por termos constituídos por parâmetros relacionados com as categorias da(s) variáveis e com os "scores" que reflectem a natureza ordinal das variáveis.

No que se segue iremos analisar os modelos a ajustar aos diferentes tipos de tabelas que correspondem respectivamente, às situações: três variáveis ordinais ("ordinal × ordinal × ordinal"), duas variáveis ordinais e uma nominal ("ordinal × ordinal × nominal") e uma variável ordinal e duas variáveis nominais ("ordinal × nominal × nominal").

3.1 Modelo "ordinal × ordinal × ordinal"

Começamos pelo tabela em que todas as variáveis X, Y e Z são ordinais. Designamos os "scores" das variáveis X, Y e Z, respectivamente por $\{\lambda_i\}$, $\{v_j\}$ e $\{\theta_k\}$ sempre ordenados de forma crescente .

Uma generalização do modelo de independência mútua entre as variáveis X, Y e Z, onde para cada par de variáveis se utiliza a sua natureza ordinal, é dada por:

$$\begin{aligned} \log m_{ijk} &= \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \beta_{12}(\lambda_i - \bar{\lambda})(v_j - \bar{v}) \\ &+ \beta_{13}(\lambda_i - \bar{\lambda})(\theta_k - \bar{\theta}) + \beta_{23}(v_j - \bar{v})(\theta_k - \bar{\theta}) \\ i &= 1,\ldots,I \quad j = 1,\ldots,J \quad k = 1,\ldots,K \end{aligned} \tag{3.1-1}$$

Os parâmetros β_{12}, β_{13} e β_{23} reflectem as interacções entre as variáveis (X,Y) , (X,Z) e (Y,Z) respectivamente. Este modelo tem mais três parâmetros independentes que o modelo de independência mútua pois o seu número de graus de liberdade é dado por g.l. = I.J.K - (1 + (I-1) +(J-1) + (K-1) + 3) = I.J.K - I - J - K - 1.

Tal como em tabelas bidimensionais, o desvio dos $\log m_{ijk}$ da independência é, respectivamente, uma função linear de X uma vez fixadas as variáveis Y e Z, uma função linear de Y uma vez fixadas as variáveis X e Z e uma função linear de Z uma vez fixadas as variáveis X e Y.

Os parâmetros β_{12}, β_{13} e β_{23} são estimados a partir de

$$\begin{aligned} \hat{\beta}_{12}(\lambda_{i+1} - \lambda_i)(v_{j+1} - v_j) &= \log\left(\frac{\hat{m}_{ijk}\hat{m}_{i+1,j+1,k}}{\hat{m}_{i,j+1,k}\hat{m}_{i+1,j,k}}\right) \\ \hat{\beta}_{13}(\lambda_{i+1} - \lambda_i)(\theta_{k+1} - \theta k) &= \log\left(\frac{\hat{m}_{ijk}\hat{m}_{i+1,j,k+1}}{\hat{m}_{i,j,k+1}\hat{m}_{i+1,j,k}}\right) \\ \hat{\beta}_{23}(v_{j+1} - v_j)(\theta_{k+1} - \theta_k) &= \log\left(\frac{\hat{m}_{ijk}\hat{m}_{i,j+1,k+1}}{\hat{m}_{i,j,k+1}\hat{m}_{i,j+1,k}}\right) \end{aligned} \tag{3.1-2}$$

Os estimadores $\hat{m}_{ijk}$ das frequências esperadas são determinados a partir das seguintes equações de verosimilhança conforme é sugerido por Agresti(1984):

$$\begin{aligned}
\sum_i \sum_j \lambda_i v_j \hat{m}_{ij+} &= \sum_i \sum_j \lambda_i v_j x_{ij+} \\
\sum_i \sum_k \lambda_i \theta_k \hat{m}_{i+k} &= \sum_i \sum_k \lambda_i \theta_k x_{i+k} \\
\sum_j \sum_k v_j \theta_k \hat{m}_{+jk} &= \sum_j \sum_k v_j \theta_k x_{+jk}
\end{aligned} \qquad (3.1\text{-}3)$$

3.2 Modelo "nominal × ordinal × ordinal"

Consideremos a tabela em que a variável X é nominal e as restantes Y e Z são ordinais. O modelo neste caso define-se por:

$$\log m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \tau_i^{12}\left(v_j - \bar{v}\right) + \tau_i^{13}\left(\theta_k - \bar{\theta}\right) + \beta_{23}\left(v_j - \bar{v}\right)\left(\theta_k - \bar{\theta}\right)$$

$$\text{onde } \sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = \sum_k \mu_{3(k)} = \sum_i \tau_i^{12} = \sum_i \tau_i^{23} = 0 \qquad (3.2\text{-}1)$$

O número de graus de liberdade deste modelo é dado por

g.l. = I.J.K - (1 + (I-1) + (J-1) + (K-1) + (I-1) + (I-1) + 1) = I.J.K -3I - J - K + 3

Os parâmetros $\{\tau_i^{12}\}$ representam os efeitos da variável X na associação (X,Y) que são homogéneos ao longo dos níveis da variável Z. Os parâmetros $\{\tau_i^{13}\}$ representam os efeitos da variável Z na associação (X,Z) que são homogéneos ao longo dos níveis da variável Y.

Os parâmetros β_{23} , τ_i^{12} , τ_i^{13} estimam-se a partir de:

$$\hat{\beta}_{23}\left(v_{j+1}-v_j\right)\left(\theta_{k+1}-\theta_k\right)=\log\left(\frac{\hat{m}_{ijk}\hat{m}_{i,j+1,k+1}}{\hat{m}_{i,j+1,k}\hat{m}_{i,j,k+1}}\right)$$

$$\left(\hat{\tau}_{i+1}^{12}-\hat{\tau}_i^{12}\right)\left(v_{j+1}-v_j\right)=\log\left(\frac{\hat{m}_{ijk}\hat{m}_{i+1,j+1,k}}{\hat{m}_{i+1,j,k}\hat{m}_{i,j+1,k}}\right) \qquad (3.2\text{-}2)$$

$$\left(\hat{\tau}_{i+1}^{13}-\hat{\tau}_i^{13}\right)\left(\theta_{k+1}-\theta_k\right)=\log\left(\frac{\hat{m}_{ijk}\hat{m}_{i+1,j,k+1}}{\hat{m}_{i+1,j,k}\hat{m}_{i,j,k+1}}\right)$$

Também neste modelo Agresti(1984) sugere que os estimadores $\hat{m}_{ijk}$ das frequências esperadas sejam determinados a partir das seguintes equações de verosimilhança:

$$\sum_j v_j\hat{m}_{ij+}=\sum_j v_j x_{ij+} \qquad , i=1,\ldots,I$$

$$\sum_j\sum_k v_j\theta_k\hat{m}_{+jk}=\sum_j\sum_k v_j\theta_k x_{+jk} \qquad (3.2\text{-}3)$$

$$\sum_k \theta_k\hat{m}_{i+k}=\sum_k \theta_k x_{i+k} \qquad , i=1,\ldots,I$$

3.3 Modelo "nominal × nominal × ordinal"

Consideremos X e Y duas variáveis nominais e Z a variável ordinal. Deste modo temos o seguinte modelo:

$$\log m_{ijk}=\mu+\mu_{1(i)}+\mu_{2(j)}+\mu_{3(k)}+\mu_{12(ij)}+\tau_i^{13}\left(\theta_k-\overline{\theta}\right)+\tau_j^{23}\left(\theta_k-\overline{\theta}\right)$$

$$i=1,\ldots,I \quad j=1,..,J \quad k=1,\ldots,K \qquad (3.3\text{-}1)$$

onde $\sum_i \mu_{1(i)}=\sum_j \mu_{2(j)}=\sum_k \mu_{3(k)}=\sum_i \tau_i^{13}=\sum_j \tau_j^{23}=\sum_i \mu_{12(ij)}=\sum_j \mu_{12(ij)}=0$.

Os termos $\left\{\mu_{12(ij)}\right\}$ representam a associação entre as variáveis nominais X e Y. Os parâmetros $\left\{\tau_i^{13}\right\}$ e $\left\{\tau_j^{23}\right\}$ representam respectivamente os efeitos da variável

Z na associação (X,Z) que são homogéneos ao longo dos níveis de Y e os efeitos da variável Y na associação (Y,Z) que são homogéneos ao longo dos níveis de X.

Os parâmetros τ_i^{13} e τ_j^{23} estimam-se a partir de

$$\left(\hat{\tau}_{i+1}^{13} - \hat{\tau}_i^{13}\right)\left(\theta_{k+1} - \theta_k\right) = \log\left(\frac{\hat{m}_{ijk}\hat{m}_{i+1,j,k+1}}{\hat{m}_{i+1,j,k}\hat{m}_{i,j,k+1}}\right) \qquad (3.3\text{-}2)$$

$$\left(\hat{\tau}_{j+1}^{23} - \hat{\tau}_j^{23}\right)\left(\theta_{k+1} - \theta_k\right) = \log\left(\frac{\hat{m}_{ijk}\hat{m}_{i,j+1,k+1}}{\hat{m}_{i,j+1,k}\hat{m}_{i,j,k+1}}\right)$$

Novamente para este modelo Agresti(1984) sugere que os estimadores $\hat{m}_{ijk}$ das frequências esperadas sejam determinados a partir das seguintes equações de verosimilhança:

$$\begin{aligned}
&\hat{m}_{ij+} = x_{ij+} && , i = 1,\ldots, I \quad , \quad j = 1,\ldots, J \\
&\sum_k \theta_k \hat{m}_{i+k} = \sum_k \theta_k x_{i+k} && , i = 1,\ldots, I \qquad (3.3\text{-}3) \\
&\sum_k \theta_k \hat{m}_{+jk} = \sum_k \theta_k x_{+jk} && , j = 1,\ldots, J
\end{aligned}$$

A seguinte tabela apresenta de uma forma sumária os modelos aqui apresentados para tabelas tridimensionais.

Variáveis ordinais	termos (X,Y)	de (X,Z)	associação (Y,Z)	graus de liberdade
X, Y, Z	$\beta_{12}\left(\lambda_i - \bar{\lambda}\right)\left(v_j - \bar{v}\right)$	$\beta_{13}\left(\lambda_i - \bar{\lambda}\right)\left(\theta_k - \bar{\theta}\right)$	$\beta_{23}\left(v_j - \bar{v}\right)\left(\theta_k - \bar{\theta}\right)$	I.J.K - I - J - K - 1
Y, Z	$\tau_i^{12}\left(v_j - \bar{v}\right)$	$\tau_i^{13}\left(\theta_k - \bar{\theta}\right)$	$\beta_{23}\left(v_j - \bar{v}\right)\left(\theta_k - \bar{\theta}\right)$	I.J.K -3I-J-K+3
Z	$\mu_{12(ij)}$	$\tau_i^{13}\left(\theta_k - \bar{\theta}\right)$	$\tau_j^{23}\left(\theta_k - \bar{\theta}\right)$	I.J.K-I.J-I-J-K+3

Todos os modelos apresentados neste capítulo sugerem um tratamento mais aprofundado pois simplesmente foi feita uma abordagem bastante sumária.

Bibliografia

- Agresti, A. (1984) *Analysis of Ordinal Categorical Data*. John Wiley & Sons.

- Birch, M.W. (1963) -Maximum Likehood in Three-way Contingency Tables. *J.Roy. Statist. Soc.*, B, 25, 220-233.

- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975) - *Discrete Multivariate Analysis*. Massachusetts Institute of Technology Press.

- Darroch, J. N. and D. Ractcliff (1972) - Generalized Iterative Scaling for Loglinear Models. *Ann. Math. Statist.*, 43, 1470-1480.

- Deming, W. E. and Stephan, F. F. (1940) - On a Least Squares Adjustment of a Sampled Frequency Table When The Expected Marginal Totals Are Known. *Ann. Math. Statist.*, 11, 427-444.

- Everitt, B.S. (1977) - *The Analysis of Contingency Tables*. Chapman and Hall, London.

- Everitt, B.S., Dunn, G. (1991) - *Applied Multivariate Data Analysis*. Edward Arnold, Cambridge

- Fienberg (1972) - The Analysis of Incomplete Multy-way Contingency Tables. *Biometrics*, 28, 177-202.

- Goodman, L. A. (1968) - The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interaction in Contingency Tables With Or Without Missing Cells. *J. Amer. Statist. Assoc.*, 63, 1091-1131.

- Goodman, L. A. (1979) - Simple Models for the Analysis of Association in Cross-Classification Having Ordered Categories. *J. Amer. Statist. Assoc.*, 74 , 537-552.

- Haberman. S. J. (1974) - *The Analysis of Frequency Data.* Chicago,.University of Chicago Press.

- Hildebrand, D., Lang,J., Rosenthal H. (1977) - *.Analysis of ordinal data.* Sage Publications / Beverly Hills / London

- Jobson, J.D. (1992) - *Applied Multivariate Data Analysis. Volume II : Categorical and Multivariate methods.* Springer-Verlag, New York.

- Kastenbaun, M. A. (1958) - Estimation of relative frequencies of four sperm types in Drosophila melanogaster. *Biometrics.* 14, 223-228.

- Mantel (1970) - Incomplete Contingency Tables, *Biometrics*, 26, 291-304.

- Upton, G. J. G. (1978) - *The Analysis of Cross-tabulated Data.* John Wiley & Sons, Chichester.