Universidade Federal do Rio Grande do Sul
Instituto de Matemática

Uma Nova Medida de Associação para Tabelas de Contingência *RxC* e Aplicações

Jandyra M.G. Fachel

Cadernos de Matemática e Estatística

Série G, nº 03, JUL/90
Porto Alegre, julho de 1990

UMA NOVA MEDIDA DE ASSOCIAÇÃO PARA TABELAS DE CONTINGÊNCIA RxC E APLICAÇÕES

Jandyra M.G. Fachel
INSTITUTO DE MATEMÁTICA
UFRGS
Av. Bento Gonçalves, 9500
91.500 - PORTO ALEGRE - RS

1. INTRODUÇÃO

Uma conhecida medida de associação para tabelas de contingência 2x2 é a razão de produtos cruzados (cross product ratio, odds ratio), a qual é dada por

$\psi = p_{11}p_{22}/p_{21}p_{22}$ onde p_{ij} são as freqüências observadas nas quatro caselas da tabela.

Supomos que exista uma distribuição bivariada contínua associada aos dados da tabela 2x2 e que as dicotomias tenham sido formadas por pontos de corte arbitrários nas distribuições marginais. Seria conveniente que tivéssemos uma distribuição bivariada com a propriedade de que a associação entre as variáveis fosse constante independentemente dos pontos de dicotomia escolhidos.

Pearson e Heron [1] mostraram que é sempre possível construir uma superfície para a qual o parâmetro de associação ψ é constante para toda a divisão formada pelos pontos de dicotomia. Esta distribuição foi denominada "superfície de associação constante". Plackett [2] reintroduziu a mesma distribuição, no contexto, uma família de distribuição bivariada com um parâmetro de associação, cujos membros da família variam conforme a escolha das distribuições marginais.

Mardia [3] mostrou as propriedades desta distribuição e a denominou "Tipo-C" ou distribuição Tipo-Contingência. Esta denominação tem a vantagem de claramente designar todos os membros da família como por exemplo: distribuição Tipo-C-Normal; distribuição Tipo-C-Logística, etc.

O parâmetro de associação (ψ) desta distribuição é estimado pela razão de produtos cruzados para dados apresentados em tabelas de contingência 2x2 . E mais, para qualquer, membro da família tipo-C, o coeficiente de correlação é uma função apenas de ψ .

Considere agora o caso geral, quando os dados são apresentados em uma tabela de contingência RxC , a qual tem variáveis marginais com R e C categorias ordenadas, respectivamente. Como poderíamos estimar o parâmetro de associação (ψ) da distribuição tipo C a qual nós supomos associada aos dados da tabela RxC . Neste trabalho nós apresentamos o método de máxima verossimilhança para estimação do parâmetro ψ , e a solução da equação de verossimilhança é obtida através do processo iterativo denominado Método de escores de Fisher.

2. A DISTRIBUIÇÃO TIPO-C

Suponha duas variáveis aleatórias X e Y com função distribuição F(x) e G(y) respectivamente e função distribuição conjunta H(x,y) . Qualquer distribuição bivariada com f.d. H e f.d. marginais F e G pode ser dicotomizada em um ponto arbitrário (x,y) formando-se uma tabela de contingência 2x2 .

Seja p_{ij} , i,j=1,2 a probabilidade que uma observação esteja na casela (i,j) , a qual é determinada pelas dicotomias no ponto (x,y) . Seja $p_{11}=H$, $p_{12}=F-H$, $p_{21}=G-H$ e $p_{22}=1-F-G+H$, nós temos

	y		
x	H	F-H	F
	G-H	1-F-G+H	1-F
	G	1-G	

A razão de produtos cruzados, ψ , para esta tabela é dada por

$$\psi = \frac{H(1-F-G+H)}{(F-H)(G-H)}$$

então

$$H = \begin{cases} \left[S - \{S^2 - 4\psi(\psi-1)FG\}^{1/2}\right] / \{2(\psi-1)\} & (\psi \neq 1) \\ FG & (\psi = 1) \end{cases}$$

$S=1+(F+G)(\psi-1)$.

Esta expressão define a função distribuição tipo C . Se X e Y são normais a distribuição dadas por esta expressão é a dis-

tribuição tipo C-normal, se $F(x) = x$ e $G(y) = y$ nós temos a distribuição tipo-C uniforme.

3. PROCESSO ITERATIVO PARA ESTIMAÇÃO DO COEFICIENTE DE ASSOCIAÇÃO ψ

Suponha que as variáveis U e V são classificadas em R e C categorias respectivamente, formando-se a tabela de contingência RxC . Assumimos que associadas às variáveis U e V estão as variáveis contínuas X e Y com distribuição bivariada contínua tipo-C. Suponha que as formas de F e G são conhecidas. Utilizando-se o método de estimação de máxima verossimilhança temos

$$L = C \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

então

$$\ln L = \ln C + \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij} \ln p_{ij}$$

$$\frac{\partial \ell}{\partial \psi} = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{p_{ij}} \frac{\partial p_{ij}}{\partial \psi} \qquad (1)$$

Na Tabela 1 apresentamos as fórmulas necessárias para estimar ψ a partir de dados apresentados em tabelas RxC .

n_{ij}	p_{ij}	$\frac{\partial p_{ij}}{\partial \psi}$
n_{11}	$p_{11} = H_{11}$	H'_{11}
n_{12}	$p_{12} = H_{12} - H_{11}$	$H'_{12} - H'_{11}$
n_{21}	$p_{21} = H_{21} - H_{11}$	$H'_{21} - H'_{11}$
n_{ab} $a = 2,\ldots,r-1$ $b = 2,\ldots,c-1$ $r,c > 2$	$p_{ab} = H_{ab} - H_{a-1,b} - H_{a,b-1} + H_{a-1,b-1}$	$\frac{\partial p_{ab}}{\partial \psi}$ $H'_{ab} - H'_{a-1,b} - H'_{a,b-1} + H'_{a-1,b-1}$
n_{a1}	$p_{a1} = H_{a1} - H_{a-1,1}$	$H'_{a1} - H'_{a-1,1}$
n_{1b}	$p_{1b} = H_{1b} - H_{1,b-1}$	$H'_{1b} - H'_{1b-1}$
n_{ac}	$p_{ac} = f_a - H_{a,c-1} + H_{a-1,c-1}$	$-H'_{a,c-1} + H'_{a-1,c-1}$
n_{rb}	$p_{rb} = g_b - H_{r-1,b} + H_{r-1,c-1}$	$-H'_{r-1,b} + H'_{r-1,c-1}$
n_{1c}	$p_{1c} = f_1 - H_{1,c-1}$	$-H'_{1,c-1}$
n_{r1}	$p_{r1} = g_1 - H_{r-1,1}$	$-H'_{r-1,1}$
n_{rc}	$p_{rc} = 1 - \sum_{i=1}^{r-1} f_i - \sum_{j=1}^{c-1} g_j + H_{r-1,c-1}$	$H'_{r-1,c-1}$

A solução da equação (1) é obtida pelo método iterativo conhecido como método de escores para parâmetros (Fisher's Scoring method; in Kendall and Stuart [4]). O procedimento iterativo é dado por:

$$\hat{\psi} = t-\left(\frac{\partial \ell nL}{\partial \psi}\right)_t \Big/ \left[E\left(\frac{\partial^2 \ell nL}{\partial \psi^2}\right)\right]_t = t+\left(\frac{\partial \ell nL}{\partial \psi}\right)_t (\text{var } \hat{\psi})_t$$

onde var $\hat{\psi}$ é a variância assintótica. O processo inicia com algum valor inicial t e é repetido até a convergência da estimativa para um desejado grau de precisão. Foi observado empiricamente que o valor inicial não é crítico e que o processo converge rapidamente para diferentes valores iniciais.

4. A EXPANSÃO EM SÉRIE DA FUNÇÃO DISTRIBUIÇÃO TIPO-C

Nesta seção nós apresentamos uma expressão alternativa para a função distribuição tipo-C dada por $H(x,y)$, obtida pela expansão binomial da fórmula original. A razão principal pela procura de uma expressão aproximada para H foi evitar problemas numéricos que apareceram no processo iterativo para valores de ψ na vizinhança de 1. ($\psi = 1$ quando as variáveis são independentes)

Após expandir a função $H(x,y)$ e após simplificações algébricas temos:

$$H = FG + \frac{\alpha F\bar{G}\bar{F}G}{1+\alpha(F\bar{G}+\bar{F}G)} + \frac{\alpha^3(F\bar{G}\bar{F}G)^2}{[1+\alpha(F\bar{G}+\bar{F}G)]^3} + \frac{2\alpha^5(F\bar{G}\bar{F}G)^3}{[1+\alpha(F\bar{G}+\bar{F}G)]^5}$$
$$+ \frac{5\alpha^7(F\bar{G}\bar{F}G)^4}{[1+\alpha(F\bar{G}+\bar{F}G)]^7} + \frac{14\alpha^9(F\bar{G}\bar{F}G)^5}{[1+\alpha(F\bar{G}+\bar{F}G)]^9} + \frac{42\alpha^{11}(F\bar{G}\bar{F}G)^6}{[1+\alpha(F\bar{G}+\bar{F}G)]^{11}} +$$
$$+ \frac{132\alpha^{13}(F\bar{G}\bar{F}G)^7}{[1+\alpha(F\bar{G}+\bar{F}G)]^{13}} + \ldots \qquad |\alpha| < \frac{1}{(F\bar{G}+\bar{F}G)}$$

Por causa das vantagens numéricas da expressão aproximada acima ela é usada no processo iterativo para $0.98 \leq \psi \leq 1.02$. Na figura 1 nós apresentamos um exemplo onde a aproximação foi utilizada com sensível melhoria nos resultados. O gráfico mostra a derivada da função de verossimilhança na vizinhança de $\psi = 1$ antes e depois de usar a aproximação para H.

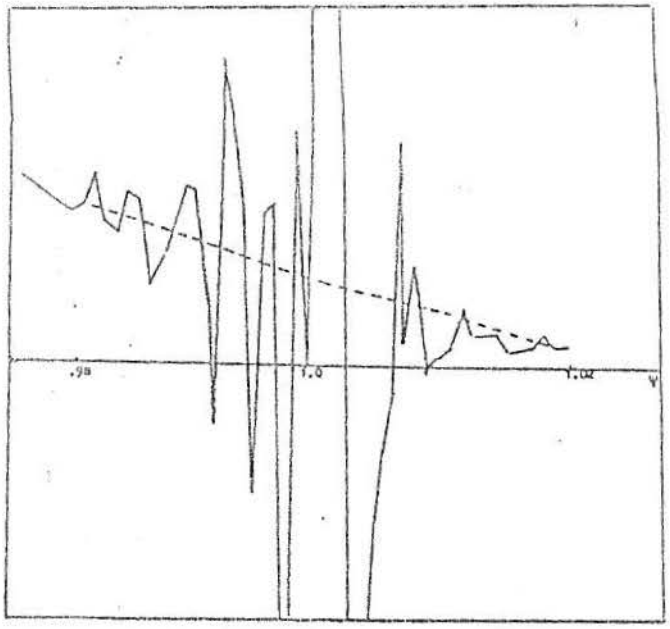

Fig.1-Derivada da função de verossimilhança na vizinhança de $\psi = 1$

5. APLICAÇÕES

São diversas as aplicações deste coeficiente de associação (ψ) para tabelas de contingência RxC , obtido pelo método de máxima verossimilhança, o qual poderíamos denominar "razão de produtos cruzados global", significando uma generalização do mesmo coeficiente para tabelas 2x2 . A interpretação de ψ para tabelas RxC é a mesma do que para tabelas 2x2 ($0 \leq \psi \leq \infty$) e o coeficiente pode ser usado para medir associação entre duas variáveis categorizadas ordinais. Uma das aplicações desta medida é a utilização de coeficientes de correlação baseados em ψ para técnicas estatísticas multivariadas, como por exemplo Análise Fatorial para dados categorizados.

BILBIOGRAFIA

[1] Pearson, K. and Heron, D. - On theories of association. Biometrika, 1913, 9, 159-315.

[2] Plackett, R.L. - A class of bivariante distributions. Journal of the American Statistical Association, 1965, 60, 516-522.

[3] Mardia, K.V. - Families of bivariate distributions. London:Charles Griffin.

[4] Kendall, M. and Sturart A. - The advanced theory of statistics (Vol.2, 4th edition). London: Charles Griffin & Co.

Publicações do Instituto de Matemática da UFRGS
Cadernos de Matemática e Estatística

Série G: Textos para Discussão

1. Carlos A. Crusius - Econometria e Verificabilidade de Teorias Econômicas - JAN/90.
2. Dinara W. Xavier Fernandez - Esperanças dos Quadrados Médios na Análise de Variância em Modelos com População Finita - JAN/90.
3. Jandyra M. Guimarães Fachel - Uma Nova Medida de Associação para Tabelas de Contingência RxC e Aplicações - JUL/90.

Universidade Federal do Rio Grande Sul
Reitor: Professor Tuiskon Dick

Instituto de Matemática
Diretor: Professor Aron Taitelbaum
Núcleo de Atividades Extra Curriculares
Coordenador: Professora Jandyra G. Fachel
Secretária: Rosaura Monteiro Pinheiro

Os Cadernos de Matemática e Estatística publicam as seguintes séries:
Série A: Trabalho de Pesquisa
Série B: Trabalho de Apoio Didático
Série C: Colóquio de Matemática SBM/UFRGS
Série D: Trabalho de Graduação
Série E: Dissertações de Mestrado
Série F: Trabalho de Divulgação
Série G: Textos para Discussão

Toda correspondência com solicitação de números publicados e demais informações deverá ser enviada para:
NAEC - Núcleo de Atividades Extra Curriculares
Instituto de Matemática - UFRGS
Av. Bento Gonçalves, 9500
91.500 - Agronomia - POA/RS
Telefone: 36.11.59 ou 36.17.85 Ramal: 252