Universidade Federal do Rio Grande do Sul
Instituto de Matemática

Coeficientes de Correlação Tipo-Contingência

Jandyra M.G. Fachel

Cadernos de Matemática e Estatística

Série A, nº 10, JAN/90
Porto Alegre, janeiro de 1990

COEFICIENTES DE CORRELAÇÃO TIPO-CONTINGÊNCIA

Jandyra Maria G. Fachel
Instituto de Matemática - DE - UFRGS

RESUMO

A medida de associação para tabelas de contingência 2x2, conhecida como a razão de produtos cruzados é generalizada para tabelas de contingência RxC, utilizando-se um método iterativo. Supomos que a distribuição associada aos dados da tabela é a distribuição hipergeométrica central Tipo-C e estimamos o parâmetro de associação desta distribuição pelo método de máxima verossimilhança. Medidas de correlação baseadas nesta medida de associação são então denominadas coeficientes de correlação "tipo-contingência". Estas medidas substituem com vantagens algumas medidas de correlação até então utilizadas para dados apresentados em tabelas de contingência. A utilização destes coeficientes em aplicações práticas é exemplificada.

1. INTRODUÇÃO

Para tabelas de contingência 2x2 várias medidas de correlação têm sido sugeridas na literatura e muitos destes coeficientes são funções da razão de produtos cruzados ("cross product ratio", "odds ratio") dada por

$$\psi = \frac{p_{11}\, p_{22}}{p_{12}\, p_{21}} \quad ,$$

onde p_{ij}, i, j=1,2 é a probabilidade de uma observação estar na célula (i,j).

O coeficiente de associação Q de Yule e o coeficiente de coligação r_y de Yule são exemplos destes coeficientes:

$$Q = (\psi - 1)/(\psi + 1)$$

$$r_y = (\psi^{1/2} - 1)/(\psi^{1/2} + 1)$$

O coeficiente Q_g de Pearson, o qual é uma aproximação ao coeficiente de correlação tetracórica é:

$$Q_g = \cos(\pi/(\psi^{1/2} + 1))$$

Bishop, Fienberg and Holland (1975, cap. 11) apresentam as propriedades básicas da razão de produtos cruzados, ψ, e consideram medidas de associação baseadas em ψ, dadas pela fórmula geral:

$$g(\psi) = \frac{f(\psi) - 1}{f(\psi) + 1}$$

Qualquer destas medidas de associação ou correlação para tabelas 2x2 baseadas em ψ podem ser associadas a uma distribuição bivariada contínua a qual seria a "geradora" dos dados da tabela de contingência formada por uma dupla dicotomia das variáveis contínuas marginais.

Esta distribuição é denominada a distribuição tipo-C ou distribuição tipo contingência (Mardia, 1970). A distribuição tipo-C foi introduzida por Pearson e Heron (1913) como uma superfície de associação constante.

2. DISTRIBUIÇÃO TIPO-C

Suponha duas variáveis aleatórias X e Y com função distribuição F(x) e G(y) respectivamente e com função distribuição conjunta H(x,y). As distribuições marginais podem ser dicotomizadas em pontos arbitrários x e y formando uma tabela de contingência 2x2, como segue:

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

F; 1 - F; G; 1 - G

A razão de produtos cruzados para esta tabela é dada por

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

ou

$$(\psi-1)H^2 - (1+(F-G)(\psi-1))H + \psi FG = 0 \qquad \psi > 0$$

Mardia mostrou que a única raiz possível da equação quadrática acima é dada por

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

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

Esta expressão define a função distribuição tipo-C. Se X e Y são variáveis normais, a distribuição dada por esta expressão é denominada distribuição tipo-C-normal. Se as marginais são uniforme, temos a distribuição tipo-C-uniforme. Se

X e Y são logística, temos a distribuição tipo-C-logística. O parâmetro de associação para qualquer membro desta família de distribuições bivariadas contínuas é ψ que é a razão de produtos cruzados. Para dados apresentados em tabelas 2x2, este parâmetro é estimado pela razão de produtos cruzados amostral dado por $\hat{\psi} = \frac{m_{11}m_{22}}{m_{12}m_{21}}$ onde m_{ij} são as freqüências observadas.

Como Pearson e Heron (1913) mostraram, é sempre possível construir esta superfície de associação constante para a qual o parâmetro de associação ψ independe dos pontos de dicotomia nas marginais.

Por uma analogia com o nome da distribuição nós denominaremos todos os coeficientes de correlação que são funções de ψ como "Coeficientes de Correlação tipo-Contingência".

3. A RAZÃO DE PRODUTOS CRUZADOS GLOBAL PARA O CASO DE TABELAS DE CONTINGÊNCIA RxC

Considere o caso geral quando os dados são dados em uma tabela de contingência RxC, com variáveis marginais com R e C categorias ordenadas respectivamente, isto é, quando as variáveis manifestas (observadas) são politômicas em vez de dicotômicas. Como estimaríamos o parâmetro ψ da distribuição tipo-C? A idéia agora é obter uma "razão de produtos cruzados global" para dados politômicos.

Nosso problema pode ser resumido como segue: nós observamos duas variáveis ordinais U e V. Estas são classificadas em R e C categorias respectivamente. Nós supomos que subjacentes às variáveis U e V existam as variáveis latentes contínuas X e Y as quais têm a distribuição bivariada contínua tipo-C. A relação entre U e X, por exemplo pode ser assim especificada:

$U = 1$ se $X < x_1$

$U = 2$ se $x_1 \leq X < x_2$

$U = 3$ se $x_2 \leq X < x_3$

...

$U = R$ se $x_{R-1} \leq X$

o analogamente para V e Y.

O problema de estimar ψ para este caso é resolvido utilizando-se o método de estimação de máxima verossimilhança.

Os dados são dados em uma tabela de contingência RxC com freqüências observadas n_{ij}, $i=1,\ldots,R$; $j=1,\ldots,C$. As probabilidades p_{ij} são obtidas em função de H, a função distribuição tipo-C dada na seção anterior. Então a função de verossimilhança é dada por

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

onde C é uma constante que não depende do parâmetro ψ e $\sum_i \sum_j p_{ij}=1$.

Usando logaritimos temos:

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

e então

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

A solução da equação de verossimilhança é obtida pelo processo iterativo denominado e método de escores para parâmetros (Kendall and Stuart, 1979, pg. 52). Detalhes dos procedimentos computacionais bem como o formulário necessário para solução das equações de máxima verossimilhança são obtidos em Fachel (1986). Um programa de computador para obtenção do estimador de máxima verossimilhança do parâmetro ψ está disponível em FORTRAN. O programa solicita como dados de entrada as dimensões R e C da tabela de contingência e as freqüências observadas. Os resultados do programa consistem do estimador da razão de produtos cruzados global, ψ, da variância assintótica do estimador (e erro padrão) e do valor da derivada da função de verossimilhança no ponto de máximo. O número de iterações e as freqüências esperadas são também fornecidos.

4. COEFICIENTES DE CORRELAÇÃO TIPO-CONTINGÊNCIA PARA DADOS POLITÔMICOS

Muito freqüentemente nas ciências de comportamento, os dados são observados como variáveis ordinais. Exemplos são as escalas de Likert, escalas de atitude, etc. Comumente usamos valores inteiros para cada categoria, por exemplo, 1, 2 e 3 ou 1 até 5.

Olsson (1979) introduziu o coeficiente de correlação policórico, o qual é uma generalização do coeficiente tetracórico para dados binários. Olsson propõe um método para obter o estimador de máxima verossimilhança do coeficiente de correlação de uma distribuição normal bivariada a qual se supõe associada aos dados apresentados em uma tabela de contingência RxC.

Nas seções anteriores nós nos referimos ao método de máxima verossimilhança para estimar o parâmetro ψ da distribuição tipo-C para dados em tabelas de contingência RxC. O parâmetro de associação ψ é a razão de produtos cruzados constante para as (R-1) (C-1) tabelas 2x2 que podem ser formadas da tabela maior RxC e pode ser definido como a razão de produtos cruzados global. Então, as medidas de correlação como função da razão de produtos cruzados podem ser utilizadas como estimadores do coeficiente de correlação latente para tabelas RxC. Estes são os coeficientes tipo contingência para dados politômicos.

Os coeficientes de correlação tipo contingência são definidos de acordo com as fórmulas escolhidas para transformar a medida de associação ψ em um coeficiente de correlação com valores entre -1 e 1. Para margens uniformes, o coeficiente de correlação da distribuição tipo-C é dado por

$$\rho_U(\psi) = \frac{\psi + 1}{\psi - 1} - \frac{2\psi \log\psi}{(\psi - 1)^2}$$ (Ver Mardia, 1970).

Chambers (1982) mostra que algumas medidas de correlação podem ser convenientemente aproximadas pela seguinte expressão, a qual é uma generalização dos coeficientes de Yule:

$$r_\upsilon = (\psi^\upsilon - 1)/(\psi^\upsilon + 1)$$

Para $\upsilon = 2/3$, nós obtemos uma aproximação do coeficiente $\rho_U(\psi)$;

para υ = 0.64, nós obtemos um estimador do coeficiente de correlação da distribuição tipo-C normal. Para υ = 0.61, obtemos uma aproximação do coeficiente $\rho_L(\psi)$ o qual é o coeficiente de correlação da distribuição tipo-C logística (ver Fachel, 1986).

Para estimar o coeficiente de correlação de uma distribuição normal bivariada quando os dados são apresentados em tabelas de contingência 2x2, Chambers sugere o coeficiente r_υ com υ = 0.74, isto é,

$$r_{0.74} = (\psi^{0.74} - 1)/(\psi^{0.74} + 1)$$

Para dados em tabelas RxC, nós usamos o estimador de máxima verossimilhança de ψ e calculamos $r_{0.74}(\hat{\psi})$ o qual é então utilizado para estimar o parâmetro ρ da normal bivariada.

O programa de computador PSIØ fornece a variância assintótica destes estimadores, denominados coeficientes de correlação tipo contingência.

5. EXEMPLOS NUMÉRICOS

Apresentamos alguns exemplos numéricos comparando coeficientes de correlação tipo-contingência com outros métodos.

Tabela 1 - (a) Dados artificiais: Distribuição normal bivariada para ρ = 0.5 ajustada para unidades nas caselas (de Pearson e Heron, 1913, p.220)

	1	2	3	4	5+6	7	8	TOTAL
1	7	20	5	2	-	-	-	34
2	21	145	79	36	10	9	1	301
3	6	94	85	54	19	22	4	284
4	2	32	39	31	12	17	4	137
5+6	-	18	28	25	11	18	5	105
7	-	11	22	24	12	22	7	98
8	-	2	6	8	5	13	7	41
TOTAL	36	322	264	180	69	101	28	1000

(b) Freqüências esperadas para tabelas (a) supondo a distribuição tipo-C (dados obtidos pelo programa PSI0)

	1	2	3	4	5+6	7	8	TOTAL
1	4.5	19.4	5.9	2.4	0.7	0.9	0.2	34
2	20.0	153.3	74.4	30.7	8.9	11.0	2.7	301
3	7.0	89.9	92.7	53.2	16.4	20.0	4.8	284
4	2.0	27.1	39.9	34.3	12.8	16.8	4.1	137
5+6	1.2	15.9	25.2	27.2	12.3	18.3	4.9	105
7	0.9	12.0	19.1	23.5	12.6	22.8	7.1	98
8	0.4	4.4	6.8	8.7	5.3	11.2	4.2	41
TOTAL	36	322	264	180	89	101	28	1000

(c) Razão de produtos global para tabela (a)

$\psi = 4.486$

(d) Coeficientes de correlação tipo contingência para dados apresentados na tabela (a)

Coeficiente tipo Contingência $r_U(\psi)$	0.465
Coeficiente tipo Contingência $r_{0.74}(\psi)$	0.504
Valor do parâmetro de correlação	0.500

Tabela 2 - (a) Dados artificiais obtidos por Olsson (1979)

x/y	1	2	3	TOTAL
1	13	6	0	19
2	69	113	22	204
3	41	132	104	277
TOTAL	123	251	126	500

(b) Freqüências esperadas usando o modelo da distribuição tipo-C

x/y	1	2	3	TOTAL
1	10.83	6.76	1.41	19
2	76.52	103.20	24.28	204
3	35.65	141.04	100.31	277
TOTAL	123	291	126	500

(c) Razão de produtos cruzados global

$$\psi = 4.361$$

(d) Coeficientes de correlação para os dados da tabela (a)

Coeficiente Policórico de Olsson	0.490 (*)
Coeficiente tipo contingência $r_{0.74}(\psi)$	0.497
Valor do parâmetro de Correlação	0.500 (*)

(*) de Olsson (1979, pg. 456)

Os exemplos numéricos desta seção mostram que: 1) o método de máxima verossimilhança para estimar o parâmetro ψ da distribuição tipo-C nos fornece um novo método para calcular coeficientes de correlação para dados politômicos. 2) o coeficiente de correlação tipo contingência $r_{0.74}(\psi)$ é um bom estimador do coeficiente de correlação de uma distribuição normal bivariada para dados apresentados em tabelas de contingência RxC. 3) A distribuição tipo C-normal é similar à distribuição normal bivariada nos exemplos apresentados nesta seção.

Aplicação dos coeficientes de correlação tipo contingência em métodos de Análise Multivariada, como Análise Fatorial para variáveis politômicas são apresentados em Fachel (1986).

BIBLIOGRAFIA

1. Bishop, Y.M.; Fienberg, S.E. and Holand, P.W. (1975). Discrete Multivariate Analysis. Theory and practice. Cambridge, Mass. the MIT Press.

2. Chambers, R.G. (1982). Correlation coefficients from 2x2 tables and from biserial data. British Journal of Mathematical and Statistical Psychology, 35, 216-227.

3. Fachel, J.M.G. (1986). The C-Type distribution as an underlying model for categorical data and its use in Factor Analysis. PhD Thesis. London School of Economics and Political Science. University of London.

4. Kendall, M. and Stuart, A. (1979). The Advanced theory of statistics (Vol.2, 4th ed.). London: Charles Griffin & Co.

5. Mardia, K.V. (1970). Families of bivariate distributions. London: Charles Griffin & Co.

6. Pearson, K. and Heron, D. (1913). On theories of association. Biometrika, 9 159-315.

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

Série A: Trabalho de Pesquisa

1. Marcos Sebastiani - Transformation des Singularités - MAR/89.
2. Jaime Bruck Ripoll - On a Theoremof R. Langevin About Curvature and Complex Singularities - MAR/89.
3. Eduardo Cisneros, Miguel Ferrero e Maria Inés Gonzales - Prime Ideals of Skew Polynomial Rings and Skew Laurent Polynomial Rings - ABR/89.
4. Oclide José Dotto - ϵ - Dilations - JUN/89.
5. Jaime Bruck Ripoll - A Characterization of Helicoids - JUN/89.
6. Mark Thompson, V. B. Moscatelli - Asymptotic Distribution of Liusternik-Schnirelman Eigenvalues for Elliptic Nonlinear Operators -JUL/89.
7. Mark Thompson - The Formula of Weyl for Regions with a Self- Similar Fractal Boundary - JUL/89.
8. Jaime Bruck Ripoll - A Note on Compact Surfaces with Non Zero Constant Mean Curvature - OUT/89.
9. Jaime Bruck Ripoll - Compact ϵ - Convex Hypersurfaces - NOV/89.
10. Jandyra Maria G. Fachel - Coeficientes de Correlação Tipo -Contingência - JAN/90.
11. Jandyra Maria G. Fachel - The Probability of Ocurrence of Heywood Cases - JAN/90.
12. Jandyra Maria G. Fachel - Heywood Cases in Unrestricted Factor Analysis - JAN/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