Análise de Correspondência: bases teóricas na interpretação de dados categóricos em Ciências da Saúde

Correspondence Analysis: a theoretical basis for categorical data interpretation in Health Sciences

Análisis de Correspondencia: bases teóricas para la interpretación de datos categóricos en Ciencias de Salud

Antonio Fernando Catelli Infantosi [1]
João Carlos da Gama Dias Costa [1]
Renan Moritz Varnier Rodrigues de Almeida [1]

[1] Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Correspondência
A. F. C. Infantosi
Programa de Engenharia Biomédica, Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro. C. P. 68510, Cidade Universitária, Rio de Janeiro, RJ 21941-972, Brasil.
afci@peb.ufrj.br

Abstract

Categorical variables are common in the biomedical field, and many descriptive methods have been proposed for revealing intrinsic patterns in data. Correspondence Analysis is an especially useful method for categorical data analysis of large contingency tables. Although numerous studies have been published on this method, most Portuguese-language articles have failed to explore its full potential, focusing only on graphical interpretation. The current paper reviews the method, showing that graphical analysis can be enriched by the right statistics. The article presents the mathematical basis for correspondence analysis and its most frequently used statistics. The procedure has shown that such statistics enrich symmetric map evaluation, that a low relative frequency category can be represented by supplementary category points, and that inertia contributions are highly related to residual analysis of contingency tables, not easily visualized by symmetric maps. Correspondence Analysis has proven advantageous when compared to principal components analysis.

Multivariate Analysis; Health Care Quality Indicators; Statistical Data Interpretation

Resumo

Na área biomédica, a ocorrência de dados categóricos é comum, e métodos de análise específicos para este tipo de dado são usados para revelar padrões existentes. A Análise de Correspondência é uma dessas técnicas, utilizada na análise de tabelas de contingência de grande porte. A maioria dos trabalhos publicados em periódicos brasileiros foca apenas na sua interpretação gráfica, não abordando outras potencialidades da técnica. O objetivo do trabalho é mostrar a técnica não limitada à análise gráfica, mas também utilizar estatísticas que permitem sua análise quantitativa. Exemplo mostra que a análise gráfica é enriquecida com a utilização dessas estatísticas, e que a inclusão de uma categoria com baixa ocorrência pode ser considerada como categoria suplementar devido à sua baixa contribuição à inércia. Assim, diminui-se a subjetividade na análise, sendo possível revelar a relação entre as categorias com a análise de resíduos, aspecto este não facilmente observado graficamente. Comparação com a Análise de Componentes Principais mostrou a vantagem da técnica.

Análise Multivariada; Indicadores de Qualidade em Assistência à Saúde; Interpretação Estatística de Dados

Introdução

Na área de saúde, a ocorrência de variáveis qualitativas é comum, o que torna importante a aplicação de técnicas estatísticas próprias para a análise deste tipo de dado. Por exemplo, variáveis como gênero ou a ocorrência ou não de determinado atributo, ou, ainda, variáveis que possuem mais de um estado, tais como aquelas que indicam a severidade de uma doença (leve, moderado ou grave), são denominadas categóricas [1]. Dada a grande quantidade de informação armazenada nos bancos de dados atuais, métodos multivariados têm sido propostos para a obtenção de informação relevante, de maneira rápida e confiável.

Um desses métodos é a Análise de Correspondência (AC). A AC permite a visualização gráfica das categorias das variáveis em uma tabela de contingência e, assim, verificar o grau de interação entre as mesmas. Os conceitos principais da AC são os perfis de linha ou coluna e a distância qui-quadrado [2,3]. Em breves palavras, a AC é um método de análise gráfica de tabelas de contingência, e seus conceitos principais foram descritos em 1940 por Fisher [4], que os exemplificou com uma análise de associação entre cor dos olhos e tipos de cabelo de habitantes da cidade escocesa de Caithness [4,5,6]. Posteriormente, já na década de 1960, Benzécri e seu grupo de colaboradores iniciaram a divulgação da AC e suas variantes sob a alcunha de *L'Analyse des Données*, levando ao desenvolvimento da técnica na França [2,6,7]. De modo independente e em paralelo, o desenvolvimento da AC teve início nos Países Baixos e no Japão, sob títulos diferentes, como Análise de Homogeneidade e Escalonamento Dual [8]. A partir de Lebart et al. [3], no início da década de 1980, a AC e sua variante, a Análise de Correspondência Múltipla (ACM), foram difundidas mais fortemente em países anglófonos.

No Brasil, a técnica é mais utilizada nos estudos de ecologia e, em menor escala, na área da saúde. Pesquisa realizada pelos autores na base de dados SciELO, utilizando-se o termo "análise de correspondência", identificou que 156 artigos redigidos em língua portuguesa foram publicados até 2012. Desses, 52,6% foram publicados em periódicos pertencentes às Ciências Exatas e da Terra, 20,5% às Ciências Agrárias e 13,5% às Ciências da Saúde. Outros 7,7% e 5,8% foram publicados em Ciências Sociais Aplicadas e Ciências Humanas, respectivamente. O resultado da pesquisa mostrou que a AC não é muito utilizada, sendo a aplicação da técnica maior nas Ciências Exatas e da Terra havendo, porém, um crescimento evidente de seu emprego nas Ciências da Saúde no último triênio (10 artigos, de um total de 21). O primeiro trabalho publicado em periódico brasileiro na área da saúde (1992) pode ser creditado a Carvalho & Struchiner [9], que abordam a associação entre unidades de vacinação constantes do Plano Nacional de Imunização (PNI).

O objetivo deste trabalho é mostrar a AC como uma ferramenta de análise de dados categóricos não apenas limitada à análise gráfica, mas abordando também estatísticas que podem ser utilizadas na interpretação. Para isso, a teoria da AC é apresentada na seção *Teoria*, bem como as expressões algébricas do método, a discussão sobre a dualidade existente entre a análise dos perfis de linha e dos perfis de coluna, enquanto a seção *Estatísticas* contempla as estatísticas mais comumente usadas para a interpretação da AC. O artigo é concluído com um exemplo ilustrativo e a discussão sobre a contribuição que as estatísticas fornecem para a análise dos dados, sendo também realizada uma comparação do método com outra técnica, a Análise de Componentes Principais (ACP). Alguns livros-textos [2,3,5,6,10] que abordam aspectos algébricos e geométricos da AC serviram de base para este trabalho, mas que, a menos de alguma citação específica, não mais serão citados.

Teoria

Considerando-se uma tabela de contingência **N** de dimensões I x J, cuja soma de todas as suas células seja n_{++}, define-se cada elemento do perfil de linha i em relação às categorias dispostas nas colunas j como:

$$r_{ij} = \frac{n_{ij}}{n_{i+}} \quad (1)$$

em que n_{ij} é o elemento da célula i, j, e n_{i+} é a soma total da i-ésima linha, ou seja,

$$n_{i+} = \sum_{j=1}^{J} n_{ij}$$

A menos que definido em contrário, entende-se que $1 \leq i \leq I$ e $1 \leq j \leq J$, $I, J \in \mathbf{N}$ para todo o texto. Assim, o perfil da i-ésima linha pode ser considerado um vetor no espaço J - dimensional, cujas coordenadas são dadas por cada elemento r_{ij}, ou seja, pelo vetor $\mathbf{r_i} = [r_{i1}; r_{i2}; ...; r_{iJ}]$. Analogamente, para os perfis de coluna, tem-se:

$$c_{ij} = \frac{n_{ij}}{n_{+j}} \quad (2)$$

em que

$$n_{+j} = \sum_{i=1}^{I} n_{ij} \text{ e } \mathbf{c_j} = [c_{1j}; c_{2j}; ... c_{Ij}]$$

e, assim, pode-se construir a matriz $\mathbf{A}(i,:) = [\mathbf{r_i}]$ composta de perfis de linha e a matriz $\mathbf{B}(:,j) = [\mathbf{c_j}]$ composta de perfis de coluna. Os vetores $\mathbf{c} = [n_{+1};n_{+2};...n_{+J}]^t$ e $\mathbf{r} = [n_{1+};n_{2+};...n_{I+}]$ são os vetores de totais de coluna e linha, respectivamente. Na nomenclatura da AC, os centroides de linha (ou perfil médio de linha) e de coluna (perfil médio de coluna) podem ser expressos respectivamente por:

$$\mathbf{r}_0 = [r_{01};r_{02};...;r_{0J}] = [r_{0j}] =$$

$$\left[\frac{n_{+1}}{n_{++}};\frac{n_{+2}}{n_{++}};...;\frac{n_{+j}}{n_{++}}\right] = \frac{\mathbf{c}}{n_{++}} \quad (3)$$

e

$$\mathbf{c}_0 = [c_{01};c_{02};...;c_{0I}]^t = [c_{0i}]^t =$$

$$\left[\frac{n_{1+}}{n_{++}};\frac{n_{2+}}{n_{++}};...;\frac{n_{I+}}{n_{++}}\right]^t = \frac{\mathbf{r}}{n_{++}} \quad (4)$$

Os elementos de um centroide refletem a importância relativa de uma determinada categoria em relação às demais e, uma vez que são calculados em relação ao total da tabela, são chamados de massas. A Tabela 1 ilustra os perfis de linha de uma AC simples.

Pode-se verificar que o centroide de linha é a média ponderada entre os perfis de linha e suas respectivas massas, ou seja:

$$\sum_{i=1}^{I} c_{0i}\mathbf{r_i} = \sum_{i=1}^{I} \frac{n_{i+}}{n_{++}} \cdot \frac{n_{ij}}{n_{i+}} =$$

$$\frac{1}{n_{++}} \sum_{i=1}^{I} n_{ij} = \frac{n_{+j}}{n_{++}} = r_{0j} \quad (5)$$

Similarmente, para os perfis de coluna:

$$\sum_{j=1}^{J} r_{0j}\mathbf{c_j} = \sum_{j=1}^{J} \frac{n_{+j}}{n_{++}} \cdot \frac{n_{ij}}{n_{+j}} = \frac{1}{n_{++}} \sum_{j=1}^{J} n_{ij} = \frac{n_{i+}}{n_{++}} = c_{0i} \quad (6)$$

Define-se a distância qui-quadrado entre os perfis das linhas i e i' como:

$$D_{qui}^{i,i'} = \sqrt{\sum_{j=1}^{J} \frac{1}{r_{0j}} \cdot (r_{ij} - r_{i'j})^2} \quad (7)$$

Utilizando-se (1), a distância qui-quadrado entre quaisquer dois perfis de linha pode ser reescrita como:

$$D_{qui}^{i,i'} = \sqrt{\sum_{j=1}^{J} \frac{1}{r_{0j}} \cdot \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i+}}\right)^2} =$$

$$= \sqrt{\sum_{j=1}^{J} \left(\frac{n_{ij}}{\sqrt{r_{0j}} \cdot n_{++} \cdot c_{0i'}} - \frac{n_{i'j}}{\sqrt{r_{0j}} \cdot n_{++} \cdot c_{0i'}}\right)^2} \quad (8)$$

resultando que a distância qui-quadrado entre perfis de linha é a distância Euclidiana em um espaço Euclidiano com novas coordenadas. Denotando

$$\frac{n_{ij}}{n_{++}} = p_{ij} \text{ e } \frac{p_{ij}}{\sqrt{r_{0j}} \cdot c_{0i}} = s_{ij}^l,$$

tem-se

$$D_{euc}^{i,i'} = \sqrt{\sum_{j=1}^{J} \left(s_{ij}^l - s_{i'j}^l\right)^2}$$

Sendo $\mathbf{D_c} = diag(\mathbf{c}_0)$ e $\mathbf{D_r} = diag(\mathbf{r}_0)$ as matrizes diagonais dos centroides de coluna (I x I) e linha (J x J), respectivamente, tem-se:

$$\mathbf{S^I} = \mathbf{D_c^{-1}} \times \mathbf{P} \times \mathbf{D_r^{-0,5}} \quad (9)$$

em que S^l é a nova matriz com a padronização da matriz de perfis de linha (I x J), e P a matriz com elementos p_{ij}. Similarmente, utilizando-se a distância qui-quadrado entre os perfis de coluna, obtém-se

$$D_{qui}^{j,j'} = \sqrt{\sum_{i=1}^{I} \frac{1}{c_{0i}} \cdot (c_{ij} - c_{ij'})^2} \text{ e então:}$$

$$\mathbf{S^c} = \mathbf{D_c^{-0,5}} \times \mathbf{P} \times \mathbf{D_r^{-1}} \quad (10)$$

Em relação ao centroide, a distância qui-quadrado do perfil de linha i pode ser representada por uma distância Euclidiana com as coordenadas:

$$D_{qui}^{i,r_0} = \sqrt{\sum_{j=1}^{J} \frac{1}{r_{0j}} \cdot (r_{ij} - r_{0j})^2} =$$

$$= \sqrt{\sum_{j=1}^{J} \left(\frac{r_{ij}}{\sqrt{r_{0j}}} - \frac{r_{0j}}{\sqrt{r_{0j}}}\right)^2} =$$

$$= \sqrt{\sum_{j=1}^{J} \frac{1}{r_{0j}} \cdot (\tilde{r}_{ij} - \tilde{r}_{0j})^2} \quad (11)$$

em que $\tilde{r}_{ij} = \frac{r_{ij}}{\sqrt{r_{0j}}}$ e $\tilde{r}_{0j} = \frac{r_{0j}}{\sqrt{r_{0j}}}$

Com base nas definições de perfil de linha e coluna, é possível verificar o princípio distributivo, no qual dois perfis semelhantes podem ser unidos em um perfil único com massa igual à soma das massas individuais relativas aos perfis considerados [2,3,11]. De fato, sejam os perfis de linha $\mathbf{r_i}$ e $\mathbf{r_{i'}}$, com massas n_{i+}/n_{++} e $n_{i'+}/n_{++}$, respectivamente, tal que $\mathbf{r_i} \approx \mathbf{r_{i'}}$ e, consequentemente, $n_{i+} \approx n_{i'+}$ e $n_{ij} \approx n_{i'j}$. Somando-se os elementos dos perfis de linha, coluna a coluna, tem-se um novo perfil de linha, com elementos $[(r_{i1}+r_{i'1}); (r_{i2} + r_{i'2});...;(r_{iJ} + r_{i'J})]$ e massa $(n_{i+} + n_{i'+})/n_{++}$.

Tabela 1

Perfis de linha em uma Análise de Correspondência simples (matriz **A**). A última linha representa o centroide de linha.

	col1	col2	...	colJ	Total
$\mathbf{r_1}$	n_{11}/n_{1+}	n_{12}/n_{1+}	...	n_{1J}/n_{1+}	1
$\mathbf{r_2}$	n_{21}/n_{2+}	n_{22}/n_{2+}	...	n_{2J}/n_{2+}	1
	...	...	...	...	...
$\mathbf{r_I}$	n_{I1}/n_{I+}	n_{I2}/n_{I+}	...	n_{IJ}/n_{I+}	1
$\mathbf{r_0}$	r_{01}	r_{02}	...	r_{0J}	1

Como são semelhantes (ou seja, suas posições no espaço são iguais ou aproximadamente iguais), a nova linha da tabela de contingência é dada por $2[n_{i1};n_{i2};...;n_{ij}]$ e massa $2n_{i+}/n_{++}$, então a soma total de cada coluna, n_{+j} não se altera, resultando em um mesmo centroide de linha r_{0j} (equação (3)). Assim, as distâncias dos perfis ao centroide de linha, dadas pela equação (11), e entre quaisquer perfis (equação (7)), não se modificam.

Os gráficos utilizando os perfis de linha, em relação às categorias da outra variável, localizadas nas colunas, ensejam uma interpretação geométrica peculiar devido ao formato da Tabela 1 (todas as linhas somam uma unidade e é uma matriz não negativa). Tal formato gera uma figura geométrica regular (ou um *simplex* regular) em (J - 1) dimensões para os perfis de linha [1,12,13,14]. Uma vez que a soma dos elementos de um perfil de linha totaliza a unidade, pode-se imaginar que, se para determinado perfil de linha somente existisse uma única categoria de coluna, a massa deste perfil estaria localizada nesta categoria [ver equação (1)]. Assim, um perfil de linha fictício, com toda sua massa concentrada na J-ésima coluna, teria elementos [0;0;...;1]. Tal ponto seria um dos vértices do simplex [15,16]. Tal sistema de representação é denominado de sistema de coordenadas baricêntricas [13].

O mais comum na disposição gráfica de vetores multivariados é utilizar a origem das próprias variáveis ou determinar uma nova origem por meio do método de mínimos quadrados, no qual a soma dos quadrados das distâncias dos pontos até os eixos é minimizada. O método computacional mais utilizado para tal minimização é o algoritmo de Decomposição por Valores Singulares (DVS), em que a matriz de perfis é fatorada em três matrizes, uma das quais diagonal com os valores singulares [17]. Adicionalmente, a quantidade de valores singulares positivos indica o posto da matriz. Entretanto, na AC, a DVS é aplicada aos perfis de forma ponderada, usando-se os elementos do centroide correspondente de maneira a ressaltar a importância que cada perfil tem, e permitir que o ponto correspondente ao centroide seja a origem dos eixos cartesianos [13]. A forma de padronização na AC pode ser expressa por:

$$\mathbf{S} = \mathbf{D_c^{-0,5}} \times \mathbf{P} \times \mathbf{D_r^{-0,5}} \quad (12)$$

sendo que, fatorando **S** pela DVS, tem-se:

$$\mathbf{S} = \mathbf{U} \times \mathbf{D} \times \mathbf{V^t} \quad (13)$$

em que **U** e **V** são as matrizes de vetores singulares à esquerda e direita, $\mathbf{V^tV} = \mathbf{U^tU} = \mathbf{I_d}$ respectivamente, **D** é a matriz diagonal com os valores singulares em ordem decrescente $\lambda_0 > \lambda_1 \geq ... \geq \lambda_K > \lambda_{K+1} = ... = 0$, em que K é o posto da matriz **S**, $K \leq \min\{I,J\}$ $\mathbf{I_d}$ é a matriz identidade [17]. O valores singulares elevados ao quadrado são os autovalores da matriz quadrada $\mathbf{S^tS}$, que traduzem a dispersão das categorias-ponto (CP) sobre a dimensão correspondente.

Define-se como coordenadas padronizadas das CPs de linha a matriz

$$\Phi = \mathbf{D_c^{-0,5}} \times \mathbf{U} \quad (14)$$

e para as CPs de coluna a matriz

$$\Gamma = \mathbf{D_r^{-0,5}} \times \mathbf{V} \quad (15)$$

Entretanto, a apresentação mais comum na AC é utilizar as coordenadas principais das CPs de linhas e colunas no mesmo gráfico, dadas por:

$$\mathbf{F} = \Phi \times \mathbf{D} \quad (16)$$

para as linhas e

$$\mathbf{G} = \Gamma \times \mathbf{D} \quad (17)$$

para as colunas.

A padronização de **S** conforme a equação (12) resulta que o primeiro valor singular (λ_0) seja unitário (e máximo) e seus correspondentes vetores singulares tenham todos os elementos positivos ou negativos [4,18]. Representam, portanto, a solução trivial, que é descartada da análise, resultan-

do na desconsideração da primeira coluna das matrizes **F** e **G**. Como resultado, a dimensionalidade da matriz **S** é dada por $K = \min\{(I\text{-}1,(J\text{-}1))\}$. Assim, a AC é, simplificadamente, a análise de duas nuvens de pontos, a nuvem de CPs de linhas e a nuvem de CPs de colunas.

Análise dual

Desenvolvendo a equação (16) que rege as coordenadas principais das CPs de linhas, e utilizando (13) e (14), obtém-se:

$$\mathbf{F} = \Phi \times \mathbf{D} = \mathbf{D}_c^{-0,5} \times \mathbf{U} \times \mathbf{D} = \mathbf{D}_c^{-0,5} \times \mathbf{S} \times \mathbf{V} \quad (18)$$

e, por consequência, sabendo-se ainda que $\mathbf{G} = \Gamma \times \mathbf{D} = \mathbf{D}_r^{-0,5} \times \mathbf{V} \times \mathbf{D}$, e substituindo-se **V** por $\mathbf{D}_r^{-0,5} \times \mathbf{G} \times \mathbf{D}^{-1}$ (equações (15) e (17)), tem-se:

$$\mathbf{F} = \mathbf{D}_c^{-0,5} \times \mathbf{S} \times \mathbf{D}_r^{0,5} \times \mathbf{G} \times \mathbf{D}^{-1} \quad (19)$$

em que $\mathbf{S} = \mathbf{D}_c^{-0,5} \times \mathbf{P} \times \mathbf{D}_r^{-0,5}$ e, portanto:

$$\mathbf{F} = \mathbf{D}_c^{-1} \times \mathbf{P} \times \mathbf{G} \times \mathbf{D}^{-1} \quad (20)$$

Similarmente,

$$\mathbf{G} = \mathbf{D}_c^{-1} \times \mathbf{P}^t \times \mathbf{F} \times \mathbf{D}^{-1} \quad (21)$$

As equações (20) e (21) são denominadas de equações de transição e permitem a transição do espaço Euclidiano das CPs de linha para o das CPs de coluna, e vice-versa. Além disso, é possível verificar que as coordenadas principais das CPs de linha possuem relação com as coordenadas padronizadas das CPs de coluna, pois, substituindo (17) em (20), tem-se:

$$\mathbf{F} = \mathbf{D}_c^{-1} \times \mathbf{P} \times \Gamma \quad (22)$$

Uma vez que,

$$\mathbf{D}_c^{-1} \times \mathbf{P} = diag\,(1/\mathbf{c}_0) \times \mathbf{N}/n_{++} =$$

$$= diag\,(n_{++}/\mathbf{r}) \times \mathbf{N}/n_{++} = diag\,(1/\mathbf{r}) \times \mathbf{N} = \mathbf{A}, \text{ então:}$$

$$\mathbf{F} = \mathbf{A} \times \Gamma \quad (23)$$

Similarmente, para as colunas tem-se:

$$\mathbf{G} = \mathbf{B}^t \times \Phi \quad (24)$$

As equações (23) e (24) mostram que as coordenadas principais de linhas (colunas) são o resultado da média ponderada das coordenadas padronizadas de colunas (linhas).

As equações de transição possibilitam a inclusão de CPs suplementares, ou ilustrativas, em coordenadas principais. As CPs suplementares podem ser incluídas na análise sempre que seu uso enriquecer a interpretação dos dados ou por razões de influência, como por exemplo, nos casos em que a massa de determinada CP é muito pequena, esta pode ser considerada um *outlier*. As CPs que foram utilizadas para gerar a solução são chamadas de CPs ativas. Para linhas suplementares, tem-se:

$$\mathbf{f}_{sup} = \mathbf{a}_{sup} \times \Gamma \quad (25)$$

em que $\mathbf{f}_{sup}$ é o vetor de coordenadas principais e $\mathbf{a}_{sup}$ é o vetor de perfil de linha suplementar, calculado segundo equação (1). Similarmente, para colunas suplementares:

$$\mathbf{g}_{sup} = \mathbf{b}_{sup}^t \times \Phi \quad (26)$$

Diversos gráficos utilizando as coordenadas dadas pelas matrizes Φ, Γ, **F** e **G** podem ser usados, sendo os mais comuns o mapa simétrico e os mapas assimétricos. O mapa simétrico apresenta as CPs de linhas e colunas em coordenadas principais e, neste caso, as equações (16) e (17) são utilizadas. Os mapas assimétricos podem mostrar as CPs de coluna em coordenadas principais e as CPs de linha em coordenadas padronizadas, ou vice-versa.

Na AC, as distâncias euclidianas entre CPs dispostas em coordenadas principais somente podem ser matematicamente determinadas dentre os seus respectivos espaços Euclidianos. Entretanto, Gabriel [19] e Greenacre [2] sugerem que a utilização dos mapas simétricos representa razoavelmente a relação entre as CPs de linha e coluna.

Estatísticas

A estatística de teste mais comum para inferir sobre a hipótese de independência (ou homogeneidade) de duas variáveis categóricas, dispostas em uma tabela de contingência é a qui-quadrado:

$$\chi^2 = \sum_{i,j=1}^{IJ} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad (27)$$

em que n_{ij} e E_{ij} são os valores observados e esperados de cada célula, respectivamente. O valor esperado, pressupondo a independência entre as variáveis, é o produto da probabilidade de ocorrência de cada uma delas na população,

$$E_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}}$$

Dividindo-se a estatística qui-quadrado pelo total da tabela de contingência, tem-se:

$$\Lambda^2 = \frac{\chi^2}{n_{++}} \quad (28)$$

em que Λ^2 é o coeficiente de contingência de Pearson. Quanto maior for Λ^2, maior a soma das diferenças entre n_{ij} e E_{ij}, ou seja, maior a dispersão entre os valores esperados e observados por célula, e menos confiança se tem na inde-

pendência entre as variáveis. A partir de (27), tem-se:

$$\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i,j=1}^{I,J} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n_{++}}\right)^2}{\frac{n_{i+}n_{+j}}{n_{++}}} \quad (29)$$

Dividindo-se o numerador e denominador de cada um dos termos pelos correspondentes quadrados dos totais de linha, obtém-se os perfis de linha observados e esperados,

$$\frac{n_{ij}}{n_{i+}} \text{ e } \frac{n_{i+}n_{+j}}{n_{++}},$$

respectivamente:

$$\chi^2 = \sum_{i,j=1}^{I,J} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n_{++}}\right)^2 \cdot \frac{1}{n_{i+}^2}}{\frac{n_{i+}n_{+j}}{n_{++}} \cdot \frac{1}{n_{i+}^2}} =$$

$$= \sum_{i,j=1}^{I,J} \left[n_{i+} \cdot \frac{\left(\frac{n_{ij}}{n_{++}} - \frac{n_{+j}}{n_{i+}}\right)^2}{\frac{n_{+j}}{n_{++}}} \right] \quad (30)$$

Lembrando que em (3) definiu-se

$$\frac{n_{+j}}{n_{++}} = r_{0j},$$

então, dividindo-se ambos os lados por n_{++}, e utilizando (1), (4) e (28) obtém-se:

$$\frac{\chi^2}{n_{++}} = \Lambda^2 = \sum_{i,j=1}^{I,J} \left[c_{0i} \cdot \frac{(r_{ij} - r_{0j})}{r_{0j}} \right] = \sum_{i=1}^{I} c_{0i} \cdot D_{qui,i}^2 \quad (31)$$

Assim, o coeficiente de contingência de Pearson é a soma do produto da massa de cada perfil de linha multiplicado pelo quadrado da sua distância qui-quadrado ao centroide. Por analogia à Física Clássica, essa grandeza é denominada na AC de inércia total. Como já citado, a distância qui-quadrado entre perfis é a distância Euclidiana com novas coordenadas, então a inércia para as CP da linha *i*, relativa às coordenadas principais, pode ser reescrita como:

$$I_T = \sum_{i=1}^{I} c_{0i} \cdot D_{euc,i}^2 = \Lambda^2 \quad (32)$$

em referência à origem. Cada CP contribui para a inércia total de acordo com a sua massa e a sua distância à origem, dada pela equação (32). A análise dos perfis de coluna leva à mesma conclusão e, portanto, o desenvolvimento algébrico a seguir somente levará em consideração as CPs de linha. Assim, a inércia total é proporcional ao quadrado de uma métrica de distância em relação ao centroide e também à estatística qui-quadrado, representando, portanto, a variabilidade dos dados da tabela de contingência. Uma vez que o centroide está localizado na origem dos eixos, relativos às coordenadas principais, a equação (32) torna-se:

$$I_T = \sum_{i=1}^{I} \sum_{k=1}^{I} c_{0i} \cdot f_{ik}^2 \quad (33)$$

em que f_{ik} é a projeção da CP de linha *i* (em coordenadas principais) relativa à dimensão *k*.

Por sua vez, a inércia total pode ser decomposta por cada uma das dimensões, então

$$I_T = \sum_{k=1}^{K} I_k$$

e a contribuição de cada dimensão à inércia total é:

$$Ctr_k = \frac{I_k}{I_T} \quad (34)$$

em que Ctr_k é também um indicador de importância da dimensão. A definição empírica do número de dimensões consideradas importantes na solução (k^*) é similar à definição na ACP e compreende, por exemplo, ou o gráfico de Scree [20] ou reter os autovalores maiores que a inércia média $\frac{I_T}{K}$, similarmente ao método de Kaiser [2,6], ou, ainda, considerar as dimensões que, somadas, possuam um porcentual mínimo determinado de inércia. O gráfico de Scree considera o último autovalor como "ruído", e os autovalores que possuem pequena diferença entre estes e aquele são desconsiderados.

A inércia de cada dimensão é igual a:

$$I_k = \sum_{i=1}^{I} c_{0i} \cdot f_{ik}^2 = \lambda_k^2 \quad (35)$$

Portanto, a inércia total é dada também por

$$I_T = \sum_{k=1}^{K} \lambda_k^2 \quad (36)$$

A contribuição absoluta da CP da linha *i* (CP_i) para todas as dimensões é:

$$Cta_i = \sum_{k=1}^{K} c_{0i} \cdot f_{ik}^2 \quad (37)$$

A Cta_i representa a dispersão que determinada CP possui em relação ao centroide, ponderada pela sua massa. É uma medida da soma dos afastamentos dos valores observados das células de uma determina categoria da tabela de contingência em relação aos valores esperados correspondentes [2]. De fato, utilizando a equação (30), a contribuição de cada célula à estatística qui-quadrado é o resíduo (quadrático) absoluto

relativo a esta célula da tabela de contingência, e é dado por [1,2]:

$$\text{Res}_{ij}^2 = \frac{(n_{ij} - E_{ij})}{E_{ij}} = \frac{\left(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n_{++}}\right)^2}{\frac{n_{i+} \cdot n_{+j}}{n_{++}}} =$$

$$= n_{i+} \cdot \frac{\left(\frac{n_{ij}}{n_{i+}} - \frac{n_{+j}}{n_{++}}\right)^2}{\frac{n_{+j}}{n_{++}}} = n_{i+} \cdot \frac{(r_{ij} - r_{0j})^2}{r_{0j}} =$$

$$= n_{++} \cdot c_{0i} \cdot \frac{(r_{ij} - r_{0j})^2}{r_{0j}} \quad (38)$$

A contribuição à estatística qui-quadrado da *i-ésima* categoria (de linha) é, portanto:

$$\text{Res}_{i+}^2 = \sum_{j=1}^{J} n_{++} \cdot c_{0i} \cdot \frac{(r_{ij} - r_{0j})^2}{r_{0j}} =$$

$$= n_{++} \cdot c_{0i} \sum_{j=1}^{J} \frac{(r_{ij} - r_{0j})^2}{r_{0j}} \quad (39)$$

Como a distância qui-quadrado entre um perfil de linha e o centroide de linha é a distância euclidiana do CP à origem em coordenadas principais, então a equação (39) torna-se:

$$\text{Res}_{i+}^2 = n_{++} \left(c_{0i} \cdot \sum_{k=1}^{K} f_{ik}^2\right) = n_{++} \cdot Cta_i \quad (40)$$

Cada CP_i contribui para a inércia da dimensão *k* por meio de:

$$Cta_{i,k} = c_{0i} \cdot f_{ik}^2 \quad (41)$$

que é chamada de contribuição absoluta da CP_i relativa à dimensão k.

A contribuição relativa para a mesma dimensão é então:

$$Ctr_{i,k} = \frac{Cta_{i,k}}{I_k} \quad (42)$$

E, consequentemente, a contribuição relativa total da CP_i para as dimensões consideradas é:

$$Ctr_i = \sum_{k=1}^{K} \frac{c_{0i} \cdot f_{ik}^2}{I_k} = Cta_i \sum_{k=1}^{K} \frac{1}{I_k} \quad (43)$$

As contribuições das CP's à inércia total são úteis, portanto, para avaliar os resíduos da tabela de contingência. A Ctr_i é importante na determinação de qual CP contribui mais para a estatística qui-quadrado, enquanto a Cta_i auxilia na composição absoluta da inércia de cada dimensão, além de ser a base do cálculo da Ctr_i.

Outras estatísticas podem ser calculadas, uma das quais a correlação quadrada, que é dada pelo quadrado do cosseno do ângulo formado pela distância da CP_i ao centroide e f_{ij}:

$$COR2_{i,k} = \cos^2(\theta_i) = \frac{f_{ik}^2}{D_{euc,i}^2} =$$

$$= \frac{c_{0i} \cdot f_{ik}^2}{\sum_{k=1}^{K} c_{0i} \cdot f_{ik}^2} = \frac{Cta_{i,k}}{I_k} \quad (44)$$

A estatística $COR2_{i,k}$ traduz a proximidade desta CP com a dimensão e, portanto, a qualidade de representação desta CP na mesma. A correlação quadrada total, referente a cada CP de linha no subespaço considerado, é dada por:

$$COR2_i = \sum_{k=1}^{K^*} COR2_{i,k} \quad (45)$$

Comparação com a ACP

Um dos métodos de análise de dados multivariados mais utilizados é a ACP. O assunto é exaustivamente abordado em textos de análise multivariada [3,6], mas uma breve introdução é oferecida a seguir.

A partir de uma matriz **X** (*n* x *p*), na qual *n* sujeitos são descritos por *p* variáveis, a ACP gera novas variáveis, não correlacionadas entre si, formadas por combinações lineares das variáveis originais, que são os componentes principais (CP's).

Assim como na AC, a fatoração da matriz **X** resulta em:

$$\mathbf{X} = \mathbf{U} \times \mathbf{D} \times \mathbf{V}^t \quad (46)$$

em que **U** e **V** são as matrizes com vetores singulares respectivamente à esquerda e à direita, e **D** é a matriz diagonal com os valores singulares.

A matriz

$$\mathbf{Z} = \mathbf{X} \times \mathbf{V} = \mathbf{U} \times \mathbf{D} \quad (47)$$

é a matriz dos escores dos componentes principais (CP's), e a matriz de coeficientes dos CP's é a matriz **V**.

Na maioria dos casos, as *p* variáveis são quantitativas e cada uma delas representa um atributo distinto. A aplicação da ACP a uma tabela de contingência **N** de dimensão *I* x *J* não é direta, pois a tabela **N** possui 2 variáveis qualitativas. Para contornar essa limitação transforma-se a tabela **N** em uma matriz binária, **X***, em que nas colunas estão dispostas as categorias das duas variáveis e, nas linhas, os sujeitos. Assim, para cada categoria é atribuído o valor unitário caso o sujeito possua este atributo, e zero, caso contrário. Como ca-

da sujeito possui uma categoria por variável, a matriz **X*** tem $n = n_{++}$ linhas e $p = I + J$ colunas. Portanto, toda informação contida na tabela **N** é preservada na matriz **X*** e a ACP pode ser aplicada de modo tradicional. A matriz **X*** é também conhecida como matriz indicadora [2].

Com base nos gráficos da AC e ACP para as dimensões escolhidas pelo método de Kaiser ou pelo gráfico de Scree, as duas soluções podem ser comparadas.

Exemplo de aplicação: AC entre Clínicas Médicas e Indicadores Hospitalares

Para ilustrar o método foram analisados os dados discutidos em Bayer & Souza [21], referentes ao Hospital Universitário de Santa Maria (Rio Grande do Sul), para os atendimentos entre os anos 2000 a 2005. Para os cálculos e gráficos foi utilizado o *software* estatístico R, livremente disponível pela *web* (http://www.r-project.org). A comparação dos resultados foi realizada usando-se a ACP.

Cada Clínica Médica (CMED) foi dividida em 17 clínicas especializadas: Obstetrícia (o); Ginecologia (g); Clínica Cirúrgica Geral (c); Hematologia (m4); Hemodiálise-Diálise Peritoneal (d); Clínica Médica Geral (m5); CTI adultos (ctia); Pediatria (p); UTI Pediátrica (uti1); UTI de Recém-nascidos (uti2); Unidade de Cuidados Intermediários Neonatal (uci); Centro de Tratamento de Crianças com Câncer (cc); Centro de Tratamento de Medula Óssea (co); Pronto Atendimento de Adultos (pa) e Pronto Atendimento Pediátrico (pp); Serviço de Tratamento de Dependentes Químicos (se) e Clínica Paulo Guedes, para pacientes psiquiátricos (pg). Os Indicadores Hospitalares (IND) são Internação (I), Transferências Recebidas (TR), Altas (AL), Transferências Enviadas e Óbitos (O). Para verificar a associação entre essas duas variáveis (CMED e IND), podemos dispor os dados em uma tabela de contingência de dimensão 17x5 (Tabela 2). As categorias da variável de indicadores hospitalares serão representadas no mapa de correspondência em letras maiúsculas e as categorias da variável de clínicas médicas serão representadas por minúsculas.

A coluna referente a Óbitos foi considerada como suplementar para ilustração, tendo em vista a sua participação reduzida na representação dos dados (massa menor que 2,5%).

A inércia total dos dados é 0,399, e as contribuições das inércias por dimensão são 65,7%, 34,1% e 0,2%. É possível verificar que a magnitude da associação entre as variáveis é alta, pois $\chi^2 = 53945$, com 48 graus de liberdade ($p < 0,001$).

Apesar de ser possível a análise dos perfis de linha em relação às categorias das colunas (e vice-versa), é mais comum o uso do agrupamento dos perfis de linha e coluna em um mapa simétrico (Figura 1a). A comparação com as soluções pela ACP pode ser feita usando-se as Figuras 1b e 1c. Pelo gráfico de Scree e pela inércia média verificou-se que $k^* = 2$. Ainda, o primeiro plano principal possui 99,8% da inércia total, mostrando que não há necessidade de analisar mais dimensões. As Tabelas 3 e 4 mostram um resumo das estatísticas.

Análise geral

Analisando-se as Tabelas 3 e 4, verifica-se que todas as CPs têm excelente qualidade de representação nos dois eixos (a menor é a CP suplementar, Óbitos = 75,9%). Nenhuma das CPs contribui exageradamente para a inércia total, não existindo então CPs influentes. Para Clínicas Médicas, aquelas que possuem as maiores Ctr_i são CTI adultos, Clínica Cirúrgica Geral e Pronto Atendimento de Adultos, com 10,1%, 13% e 28,4%, respectivamente, apesar de CTI adultos possuir baixa massa (0,0021). Sua alta contribuição é devida à distância do centroide em relação à segunda dimensão, implicando a maior contribuição absoluta para essa dimensão (0,038). Claramente, as categorias Serviço de Tratamento de Dependentes Químicos e Clínica Paulo Guedes podem ser reunidas em uma única categoria (propriedade distributiva), tanto pela análise gráfica quanto pela análise das estatísticas de interesse, pois a alteração na inércia total com a integração das duas CPs foi desprezível (0,3994696 *vs.* 0,3994699). A menor Ctr_i é dada por Centro de Tratamento de Medula Óssea, seguida por Hemodiálise-Diálise Peritoneal, esta apresentando menor distância ao centroide que aquela. Embora Hemodiálise-Diálise Peritoneal esteja mais próxima do centroide, Centro de Tratamento de Medula Óssea possui a menor massa entre as CPs (0,005), daí a sua menor Ctr_i. Para Indicadores Hospitalares, a maior contribuição é Transferências Recebidas (39,9%), e a menor é Altas (12,5%). A soma da coluna representativa das contribuições absolutas totais de ambas as variáveis é igual à inércia total (0,399), servindo também como verificação dos cálculos de contribuição à inércia pelas CPs ativas, já que a CP suplementar, por definição, não possui massa, não contribuindo, portanto, para a inércia. A análise das duas CP's de maior Ctr_i (pa e c; TR e TE), sugere estar nas células pa-TR, pa-TE, c-TR e c-TE a maior contribuição ao resíduo quadrático da tabela de contingência. De fato, os dois maiores resíduos estão nas células pa-TR e pa-TE (-68,6 e -81,2, respectivamente).

Tabela 2

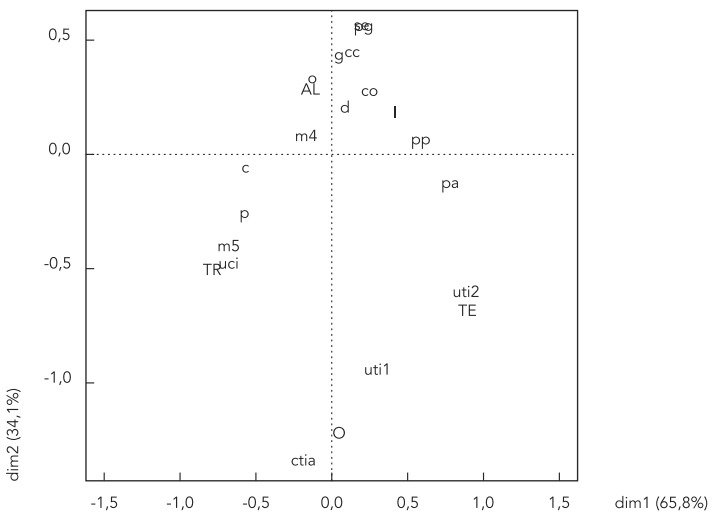
Tabela de Contingência relativa à CMED (Clínica Médica) e IND (Indicadores Hospitalares) em um hospital universitário, 2000-2005 [21].

	I	TR	AL	TE	O	Total
o	8968	3289	11933	228	2	24420
g	2881	464	3187	109	63	6704
c	3033	5468	7919	367	264	17051
m4	2415	1646	3328	380	350	8119
d	1042	277	1061	198	57	2635
m5	522	2496	2250	382	421	6071
ctia	258	1483	77	1016	655	3489
p	952	2112	2200	305	13	5582
uti1	479	613	186	764	149	2191
uti2	1541	280	230	1369	228	3648
uci	169	1355	1214	290	5	3033
cc	2300	256	2357	125	25	5063
co	328	40	305	65	7	745
pp	4049	94	2515	1660	20	8338
pa	16524	45	6791	8986	824	33170
se	2031	41	2024	40	0	4136
pg	1834	41	1818	38	1	3732
Total	49326	20000	49395	16322	3084	138127

Figura 1

Mapa de correspondência, com as categorias de Indicadores Hospitalares em maiúsculas e as categorias de Clínicas Médicas em minúsculas (1a); em (1b) e (1c) os gráficos das duas primeiras categorias ponto (CPs), para Indicadores Hospitalares e Clínicas Médicas.

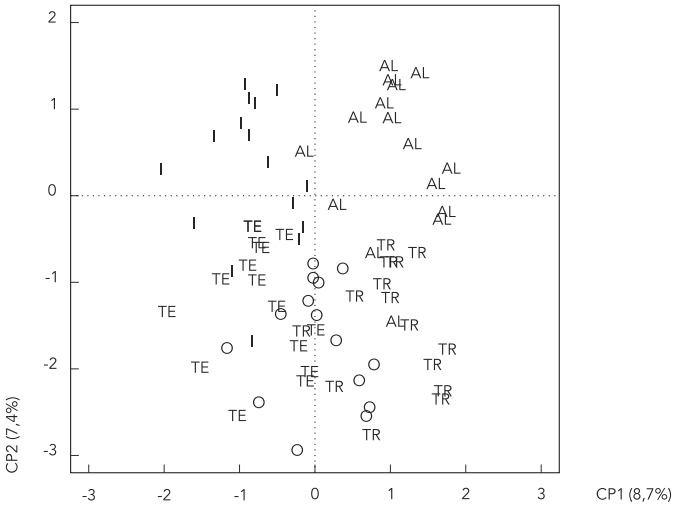
1a)



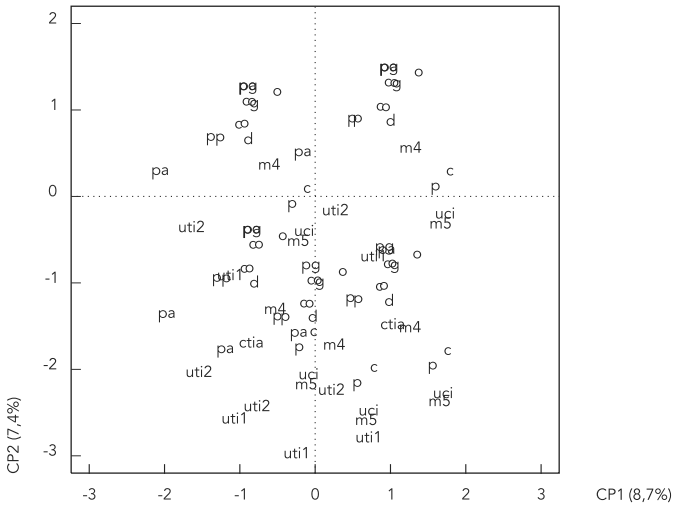
(continua)

Figura 1 (continuação)

1b)



1c)



Análise da dimensão 1

Entre as contribuições relativas das CPs da variável Clínicas Médicas para a inércia da dimensão 1, sobressaem (acima da média 1/17 = 0,059) Clínica Cirúrgica Geral, Clínica Médica, Pediatria e Pronto Atendimento de Adultos. Dessas, Clínica Cirúrgica Geral, Clínica Médica Geral e Pediatria (total de 36,4%) estão em oposição a Pronto Atendimento de Adultos (total de 41,4%). Assim, para a variável Clínicas Médicas, retendo-se as maiores contribuições, a dimensão 1 é regida pelo contraste entre Pronto Atendimento de Adultos (0,414) e Clínica Cirúrgica Geral (0,195), ou seja, contraste entre atendimentos que demandem cirurgia eletiva e Pronto Atendimento. Em relação a Indicadores Hospitalares, há contraste entre Transferências Recebidas (0,454) e Transferências Enviadas (0,284), sugerindo que a dimensão 1 seja referente a transferências entre pacientes, para esta variável.

Tabela 3

Estatísticas para a variável Clínica Médica. As duas colunas à direita são as coordenadas principais das dimensões 1 e 2.

	Massa	Cta_i	$Cta_{i,1}$	$Cta_{i,2}$	$Ctr_{i,1}$	$Ctr_{i,2}$	Ctr_i	$COR2_i$	DIM1	DIM2
	(4)	(36)	(41)	(41)	(42)	(42)	(43)	(45)	(16)	(16)
o	0,181	0,026	0,008	0,019	0,029	0,137	0,066	1,000	-0,206	0,321
g	0,049	0,009	0,000	0,009	0,000	0,064	0,022	0,999	-0,031	0,423
c	**0,124**	0,052	0,051	0,001	0,195	0,004	**0,130**	0,998	-0,642	-0,070
m4	0,058	0,005	0,005	0,000	0,018	0,002	0,013	0,995	-0,290	0,067
d	0,019	0,001	0,000	0,001	0,000	0,005	0,002	0,996	0,008	0,192
m5	0,042	0,034	0,027	0,007	0,102	0,053	0,085	1,000	-0,800	-0,415
ctia	**0,021**	0,040	0,002	**0,038**	0,008	0,282	**0,101**	0,996	-0,317	**-1,352**
p	0,041	0,021	0,018	0,003	0,067	0,022	0,052	0,996	-0,655	-0,268
uti1	0,015	0,014	0,000	0,014	0,002	0,101	0,036	0,999	0,166	-0,955
uti2	0,025	0,024	0,014	0,010	0,054	0,070	0,060	1,000	0,749	-0,615
uci	0,022	0,020	0,014	0,005	0,054	0,040	0,049	0,995	-0,794	-0,491
cc	0,037	0,007	0,000	0,007	0,000	0,052	0,018	0,997	0,036	0,436
co	0,005	0,001	0,000	0,000	0,000	0,003	0,001	1,000	0,146	0,265
pp	0,062	0,014	0,014	0,000	0,053	0,001	0,035	0,998	0,475	0,053
pa	0,240	0,113	0,109	0,005	0,414	0,033	**0,284**	1,000	0,674	-0,138
se	0,031	0,010	0,000	0,009	0,001	0,069	0,024	0,999	0,098	0,553
pg	0,028	0,009	0,000	0,008	0,001	0,061	0,022	0,998	0,098	0,550
Total	**1,000**	**0,399**	**0,263**	**0,136**	**1,000**	**1,000**	**1,000**	*	*	*

* Não aplicável.

Nota: entre parênteses estão as equações correspondentes.

Tabela 4

Estatísticas para a variável Indicadores Hospitalares. As duas colunas à direita são as coordenadas principais das dimensões 1 e 2.

	Massa	Cta_i	$Cta_{i,1}$	$Cta_{i,2}$	$Ctr_{i,1}$	$Ctr_{i,2}$	Ctr_i	$COR2_i$	DIM1	DIM2
	(4)	(36)	(41)	(41)	(42)	(42)	(43)	(45)	(17)	(17)
I	0,365	0,057	0,046	0,011	0,175	0,078	0,142	0,996	0,355	0,170
TR	0,148	0,159	0,119	0,040	0,454	0,293	**0,399**	1,000	-0,897	-0,519
AL	0,366	0,050	0,023	0,027	0,087	0,197	**0,125**	0,996	-0,251	0,271
TE	0,121	0,133	0,075	0,059	0,284	0,432	0,334	0,999	0,785	-0,698
O	*	*	*	*	*	*	*	**0,759**	-0,043	-1,234
Total	**1,000**	**0,399**	**0,263**	**0,136**	**1,000**	**1,000**	**1,000**	*	*	*

* Não aplicável.

Nota: entre parênteses estão as equações correspondentes.

Análise da dimensão 2

Para Clínicas Médicas, verifica-se que são sete as CPs cujas contribuições ultrapassam a média (Clínica Obstétrica, Ginecologia, CTI adultos, UTI Pediátrica, UTI de Recém-nascidos, Serviço de Tratamento de Dependentes Químicos e Clínica Paulo Guedes), totalizando 78,5% da inércia da dimensão. A dimensão apresenta contraste entre um grupo formado pelas CPs CTI adultos, UTI Pediátrica, UTI Recém-nascidos e um grupo composto pelas CPs Clínica Obstétrica, Ginecologia, Serviço de Tratamento de Dependentes Químicos, Clínica Paulo Gue-

des, sugerindo oposição entre internação em unidades e centros de atendimento intensivo e serviços ambulatoriais. Para Indicadores Hospitalares, a dimensão 2 apresenta contraste entre transferências com internação e altas, embora estas não possuam contribuições acima da média (25%).

Análise conjunta

Como todas as CPs são bem representadas no plano (COR2 > 0,995), e levando-se em conta o discutido na seção *Teoria* sobre a análise de espaços Euclidianos distintos, as distâncias entre as CPs podem ser interpretadas quanto à sua proximidade, principalmente quanto aos respectivos quadrantes (Figura 1a). Assim, as CPs Serviço de Tratamento de Dependentes Químicos e Clínica Paulo Guedes podem ser reunidas em uma única CP, enquanto Transferências Recebidas apresenta maior proximidade com Clínica Médica Geral e Unidade de Cuidados Intermediários Neonatal, porém menos com Pediatria e Clínica Cirúrgica Geral (agrupamento 1). O mesmo comportamento existe para Transferências Enviadas, que está mais próximo da UTI de Recém-nascidos e no mesmo quadrante que Pronto Atendimento de Adultos e UTI Pediátrica (agrupamento 2).

Verifica-se que a CP Internações está mais relacionada também a clínicas de Pronto Atendimento Adultos e Pronto Atendimento Pediátrico e, em menor grau de associação, a Centro de Transplantes de Medula Óssea (agrupamento 3). Por outro lado, a CP Altas è mais fortemente relacionada com Obstetrícia, Ginecologia e Hematologia (agrupamento 4). A CP Óbitos está mais relacionada a CTI adultos (agrupamento 5).

Assim, o agrupamento 1 pode ser associado a pacientes transferidos de outras unidades do hospital, como pacientes provenientes de unidades cirúrgicas e pediátricas de tratamento intensivo. O agrupamento 2, por outro lado, representa os pacientes enviados a outras unidades, após pronto atendimento de adultos e pacientes pediátricos transferidos das UTIs de recém-nascidos e pediátricas. Para o agrupamento 3, uma sobreposição de pacientes adultos com o agrupamento 2 pode ser explicada pela internação de doentes mais graves após pronto atendimento e para tratamento de doenças relacionadas à medula óssea. Atendimento ambulatorial e partos sem complicações parecem descrever o agrupamento 4. O agrupamento 5 é caracterizado pela probabilidade de Óbitos maior em CTI, em oposição clara ao agrupamento 4.

Comparação com ACP

O gráfico de escores foi subdividido por cada variável, assim, cada escore foi duplamente representado pelas categorias às quais pertence. Portanto, a análise deve ser realizada com a sobreposição das Figuras 1b e 1c. As variâncias dos dois primeiros CPs (8,7 e 7,4%) são menores do que as das duas primeiras dimensões da AC, tendo em vista a menor dimensionalidade desta solução em relação àquela (3x20). Sem prejuízo da análise, a categoria Óbitos não foi considerada como suplementar, por sua baixa contribuição à inércia. A dimensionalidade da solução pelo gráfico de Scree é 4 e pelo critério de Kaiser é 17.

A análise da Figura 1b sugere contraste em relação ao CP1, entre Altas *vs.* Internação e entre Transferências Enviadas *vs.* Transferências Recebidas, enquanto que em relação ao CP2 o contraste é entre Internação/Altas e Transferências Enviadas/Recebidas. A projeção de Óbitos sobre o primeiro CP está situada próxima ao centroide, entretanto, para o segundo, há sugestão de contraste entre Óbitos *vs.* Internação/Altas.

Para a variável Clínicas Médicas, verifica-se um espalhamento generalizado entre os sujeitos, exceto para pa que situa-se no terceiro quadrante; e dois grupos similares, no primeiro e segundo quadrantes, formados pelas categorias m4-d-co-c-g-o-pg-pp. A análise da sobreposição das Figuras 1b e 1c não sugere associação relevante entre as categorias das duas variáveis, principalmente pela baixa variância associada a estes dois CPs.

Discussão

O objetivo deste trabalho foi mostrar a AC como uma técnica de análise de dados categóricos, dispostos em tabelas de contingência. Embora introduzida mais formalmente na década de 1960, a AC ainda não é muito utilizada na literatura biomédica brasileira e espera-se, com este estudo, que a mesma seja mais conhecida e incorporada ao arsenal de técnicas estatísticas mais difundidas na área da saúde.

Até o momento, não existe consenso quanto ao melhor método para determinar o número de dimensões relevantes na AC, ou na ACP. Assim, as abordagens aqui apresentadas são as mais utilizadas na literatura referenciada. Jackson [22] propôs um método computacionalmente intensivo e o teste *Broken-Stick* para a determinação de *k**, entretanto, estes métodos possuem complexidade maior do que a AC ou ACP *per se*. Logo, o *scree plot* e o teste da inércia média (que é similar ao teste de Kaiser) foram utiliza-

dos. Outro ponto de subjetividade diz respeito à definição de "proximidade" na interpretação gráfica das associações, pois a definição de "próximo" ou "distante" depende da escala do gráfico e da definição formal de um índice de similaridade. No contexto da AC, Carvalho & Struchiner [9] propuseram o uso conjunto de um algoritmo de classificação hierárquica (ACH) ascendente, enquanto que, para a ACM, Costa et al. [23] sugeriram um índice baseado na razão entre as inércias das dimensões. A análise de proximidade entre categorias na AC deve levar em conta esses aspectos, sendo melhor caracterizar a proximidade relativa entre as categorias e entre estas e o centroide. Além de dispensar a pressuposição de distribuição de probabilidade, outra vantagem da AC é que a mesma não é somente aplicável a tabelas de contingência [2].

Como mencionado, a AC encontra sua principal aplicação em tabelas de contingência com grande número de categorias, de forma a permitir uma visualização da "atração" ou "repulsão" entre as categorias das variáveis. O exemplo de aplicação da AC utilizado neste trabalho mostrou que a análise do mapa simétrico pode ser enriquecida com a utilização das estatísticas descritas no texto, e que a inclusão de uma categoria com baixa ocorrência pode ser considerada na análise como categoria suplementar devido à sua baixa contribuição à inércia. Ainda, as CPs suplementares têm aplicação quando se deseja verificar proximidade/relação com as CPs "explicativas" [3] ou na análise de influência [24]. As estatísticas quantitativas utilizadas diminuem a subjetividade na análise, pois mostram, de maneira inequívoca, as relações entre as CPs e os eixos, a relação de cada CP com a inércia, sugerindo ainda a relação entre as estatísticas de contribuições absolutas e relativas com a análise de resíduos [1,2], o que não pôde ser visualizado pelo mapa de correspondência. Para a ACP foi necessário construir uma matriz indicadora, de modo a comparar o resultado desta com o da AC. Essa comparação mostra a vantagem da AC na análise das relações entre as categorias em uma tabela de contingência, principalmente devido à diferença na informação contida nas dimensões correspondentes.

Resumen

En el campo biomédico, los datos categóricos son frecuentemente utilizados y los métodos de análisis específicos son empleados para revelar patrones intrínsecamente existentes en los mismos. El Análisis de Correspondencias es una de estas técnicas, siendo útil en el análisis de tablas de contingencia con un gran número de clases. A pesar de que muchos artículos han explorado esta técnica, la mayoría de trabajos en revistas brasileñas se centra sólo en su interpretación gráfica. El objetivo de este trabajo es incluir estadísticas que permitan la interpretación cuantitativa de la técnica. Como ejemplo, tenemos el análisis de un mapa simétrico enriquecido con el uso de estadísticas, en el cual la inclusión de una clase de baja ocurrencia puede ser considerada como una categoría suplementaria, debido a su baja contribución a la inercia de datos. Por lo tanto, disminuye la subjetividad en el análisis, siendo posible ahora revelar la relación entre las categorías con el análisis residual, lo que no es fácil observar en los gráficos. La comparación con el análisis de componentes principales mostró sus ventajas.

Análisis Multivariante; Indicadores de Calidad de la Atención de Salud; Interpretación Estadística de Datos

Colaboradores

A. F. C. Infantosi, J. C. G. D. Costa e R. M. V. R. Almeida participaram da pesquisa bibliográfica, análise e interpretação dos dados, redação e revisão do texto.

Agradecimentos

Os autores agradecem ao CNPq, Capes e Faperj pelo apoio financeiro, a Fábio Mariano Bayer e Adriano Mendonça Souza, autores do artigo utilizado no exemplo de aplicação da Análise de Correspondência, por permitirem a utilização dos dados, e aos revisores pelas valiosas sugestões e comentários que contribuíram para a melhoria do manuscrito.

Referências

1. Pereira JCR. Análise de dados qualitativos – estratégias metodológicas para as ciências da saúde, humanas e sociais. 3ª Ed. São Paulo: Edusp; 2004.
2. Greenacre M. Correspondence analysis in practice. 2nd Ed. Barcelona: Chapman & Hall/CRC Press; 2007.
3. Lebart L, Morineau A, Warwick KM. Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Hoboken: John Wiley and Sons; 1984.
4. Fisher RA. The precision of discriminant functions. Ann Eugen 1940; 10:422-9.
5. Gower JC, Hand DJ. Biplots. London: Chapman & Hall; 1996.
6. Le Roux B, Rouanet H. Geometric data analysis: from correspondence analysis to structured data analysis. Dordrecht: Kluwer Academic Publishers; 2004.
7. Benzécri JP. Histoire et préhistoire de l'analyse des données. Partie V: l'analyse des correspondances. Cah Anal Donnees 1977; 2:9-40.
8. Michailidis G, de Leeuw J. The GIFI system of descriptive multivariate analysis. Stat Sci 1998; 13:307-36.
9. Carvalho MS, Struchiner CJ. Análise de Correspondência: uma aplicação do método à avaliação de serviços de vacinação. Cad Saúde Pública 1992; 8:287-301.
10. Hair Jr. JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Análise multivariada de dados. 6ª Ed. Porto Alegre: Bookman; 2009.
11. Benzécri JP. El análisis de correspondencias. Cah Anal Données 1977; 2:125-42.
12. Fienberg SE. The geometry of an rxc contingency table. Annals of Mathematics Statistics 1968; 39:1186-90.
13. Greenacre M, Hastie T. The geometric interpretation of correspondence analysis. J Am Stat Assoc 1987; 82:437-47.
14. Heiser WJ. Geometric representation of association between categories. Psychometrika 2004; 69:513-45.
15. Coxeter HSM. Regular polytopes. 2nd Ed. New York: Dover Publications; 1973.
16. McMullen C. The visual guide to extra dimensions: visualizing the fourth dimension, higher-dimensional polytopes, and curved hypersurfaces. v. 1. Lexington: Custom Books; 2008.
17. Golub GH, Van Loan CF. Matrix computations. 3rd Ed. Baltimore: John Hopkins University Press; 1996.
18. van de Velden M, Neudecker H. On an eigenvalue property relevant in correspondence analysis and related methods. Linear Algebra and its Applications 2000; 321:347-64.
19. Gabriel KR. Goodness of fit of biplots and correspondence analysis. Biometrika 2002; 89:423-36.
20. Cattell RB. The scree test for the number of factors. Multivariate Behav Res 1966; 1:245-76.
21. Bayer FM, Souza AM. Análise de Correspondência entre clínicas médicas e indicadores hospitalares: o caso do Hospital Universitário de Santa Maria. In: IV Simpósio de Excelência em Gestão e Tecnologia. http://www.aedb.br/seget/historico_seget2007.html (acessado em 10/Jun/2011).
22. Jackson DA. Stopping rules in Principal Components Analysis: a comparison of heuristical and statistical approaches. Ecology 1993; 74:2204-14.
23. Costa JCGD, Almeida RMVR, Infantosi AFC. A heuristic index for selecting similar categories in multiple correspondence analysis applied to living donor kidney transplantation. Comput Methods Programs Biomed 2008; 90:217-29.
24. Pack P, Jolliffe IT. Influence in correspondence analysis. Appl Stat 1992; 41:365-80.

Recebido em 14/Jul/2013
Versão final reapresentada em 16/Out/2013
Aprovado em 06/Nov/2013