# Análise de Correspondência Simples e Múltipla para Dados Amostrais Complexos

Augusto Carvalho Souza[1], Ronaldo Rocha Bastos[2] e
Marcel de Toledo Vieira[2]

[1] CEDEPLAR, Universidade Federal de Minas Gerais

[2] Departamento de Estatística, Universidade Federal de Juiz de Fora

**Resumo**

Neste trabalho é apresentado um estudo sobre o efeito de planos amostrais complexos em alguns resultados de Análise de Correspondência Simples e Múltipla. Resultados indicam que tais métodos de análise exploratória de dados multivariados podem ser seriamente afetados pela estrutura do desenho adotado para a seleção da amostra estudada. A principal contribuição deste trabalho é uma proposta de uma metodologia que permitir a consideração do plano amostral na Análise de Correspondência Múltipla.

**Palavras-chave**: análise de correspondência, análise de correspondência múltipla, dados amostrais complexos.

## 1 Introdução

Amostras da maioria das pesquisas domiciliares são selecionadas a partir de planos amostrais complexos, envolvendo estratificação, conglomeração em vários estágios e seleção com probabilidades desiguais. Os efeitos de desenhos amostrais complexos sobre os resultados da aplicação de toda a classe de técnicas relacionadas à Análise de Correspondência (AC) ainda foram muito pouco abordados na literatura da amostragem. De acordo com Nyfjäll (2002), os resultados encontrados a partir da aplicação de Análise de Correspondência Simples (ACS) podem ser seriamente afetados pela estrutura de planos amostrais complexos.

Apesar de ter havido um aumento considerável de publicações referentes à AC (Beh, 2002) e

o principal congresso internacional da área, CARME (*Correspondence Analysis and Related Methods Conference*), ter crescido bastante nos últimos anos, pouca ou quase nenhuma atenção tem sido dada ao efeito da amostragem nos resultados de AC. Este artigo tem como principal objetivo propor metodologia para a consideração do plano amostral na Análise de Correspondência Múltipla (ACM), sobretudo no que diz respeito ao cálculo das estimativas das projeções fatoriais.

Na Seção 2 é apresentada uma revisão sobre ACS e ACM, enquanto que na Seção 3, algumas questões relacionadas aos planos amostrais complexos são abordadas. Uma breve revisão sobre as principais contribuições encontradas na literatura sobre a consideração de planos amostrais em ACS é apresentada na Seção 4. Na Seção 5, metodologia para a consideração do plano amostral na ACM é proposta, enquanto que considerações finais são apresentadas na Seção 6.

## 2 Análise de Correspondência Simples e Múltipla

ACS pode ser definida como uma técnica de análise multivariada, adequada para dados categóricos, que permite analisar graficamente as relações existentes através da redução de dimensionalidade do conjunto de dados. Tal técnica é aplicada a tabelas de contingência com o objetivo de determinar o grau de associação global entre suas linhas e as colunas, indicando como as variáveis estão relacionadas. Este método tem como base a decomposição do valor singular de uma matriz retangular (tabela de contingência adaptada) e é utilizado para representar graficamente as linhas e as colunas desta tabela como pontos em espaços vetoriais de pequena dimensão. Com os gráficos produzidos podemos avaliar visualmente se as variáveis de interesse se afastam do pressuposto de independência, sugerindo possíveis associações e ainda perceber como se dá esta associação. Os níveis das variáveis de linha e de coluna assumem posições nos gráficos de acordo com a associação ou similaridade entre elas. (Greenacre, 1984; Greenacre, 2007; Benzécri, 1992).

A ACM, por sua vez, tem como base uma adaptação da estrutura dos dados para que se tenham casos nas linhas e variáveis nas colunas, gerando uma tabela de código binário 0 e 1 (matriz

indicadora, que estaremos denotando por $\mathbf{M_{ind}}$) que fornecerá os mesmos resultados que a ACS se apenas duas variáveis forem analisada. Entretanto, esta estrutura permite que mais de duas variáveis sejam analisada ao mesmo tempo; situação na qual a aplicação e interpretação da ACS se torna bastante complexa. Além desta diferença, a ACM permite que se estabeleçam os perfis de cada unidade observada, possibilitando a avaliação de relações entre estes e a variáveis analisadas.

## 3 Planos Amostrais Complexos

Planos amostrais complexos normalmente envolvem a seleção de unidades de uma população finita com probabilidades distintas, estratificação, e conglomeração das unidades em múltiplos estágios. Quando os dados a serem analisados foram coletados a partir de desenhos complexos as técnicas estatísticas adotadas devem levar em consideração as características do procedimento amostral (Pessoa e Silva, 1998; e Lehtonen e Pahkinen, 1996).

## 4 ACS e Planos Amostrais Complexos

Nyfjäll (2002), através de simulação, testou a adequação dos gráficos gerados a partir de dados amostrais com gráficos obtidos a partir de dados populacionais. Os planos amostrais considerados foram: amostragem de Poisson e amostragem estratificada com alocações uniforme, proporcional e alocação proporcional a x-total (alocação da amostra de acordo com características de variável auxiliar x, correlacionada com a variável de estudo y). Neste trabalho, as principais conclusões obtidas foram: (i) se as proporções da tabela de correspondência (ou dos perfis de linha ou coluna) são somente as proporções amostrais, isto é, não são estimadores não viciados da verdadeira proporção populacional (tabela expandida), então a análise de correspondência pode gerar gráficos que apresentam estruturas completamente inversas comparadas às apresentadas na população; (ii) para os planos amostrais investigados, concluiu-se que, sempre que possível, é importante utilizar um estimador não viciado da proporção populacional ao se aplicar a AC, o que permite que a

estrutura de relação entre as linhas e colunas revelada pela aplicação da AC aos dados amostrais tenda a se manter como na população.

Além disso, para investigar o comportamento da inércia, ou seja, se o cálculo "tradicional" da inércia é ou não afetado pelo plano amostral, Nyfjäll (2002) propõe, analiticamente, um estimador da inércia de uma população finita que leve em consideração o plano amostral. O vício deste estimador é investigado através de simulação. Neste caso, os planos amostrais considerados foram: amostragem aleatória simples com e sem reposição, amostragem de Bernoulli, amostragem de Poisson (com três diferentes probabilidades de inclusão) e amostragem estratificada simples com alocações uniforme e proporcional. O mesmo autor chegou às seguintes conclusões: (iii) o estimador tradicional da inércia é um estimador ruim para a inércia populacional quando se está sob os planos amostrais estudados (em todos os planos amostrais o vício foi considerado alto); (iv) o estimador proposto, que é determinado sob cada plano amostral, apresentou-se praticamente não viciado em quase todos os planos se o tamanho da amostra é grande o suficiente (acima de 500); (v) considerando o estimador da inércia populacional proposto, Nyfjäll (2002) afirma que quanto maior o tamanho da tabela de contingência, para um dado tamanho da amostra e inércia populacional, pior o desempenho de tal estimador.

Resumidamente, podemos esperar que, ao considerar os pesos amostrais no cálculo das proporções de cada célula da tabela analisada (tabela expandida), o gráfico resultante manterá a mesma estrutura que na população.

## 5 ACM e Planos Amostrais Complexos

Dada a relação entre a ACM e a ACS, espera-se que os resultados gerais encontrados por Nyfjäll (2002) sejam mantidos para a ACM. Intuitivamente, é possível perceber que a melhor estimativa pontual das projeções fatoriais só pode ser obtida a partir de tabelas de contingência "expandidas" com os pesos amostrais, já que estas se constituem estimativas pontuais dos totais de cada célula de

uma tabela. Como as técnicas relacionadas à AC são, em essência, descritivas, é de se esperar, portanto, que a melhor estimava da localização dos pontos (perfis) em um gráfico de correspondências seja aquela dada por perfis obtidos a partir de uma tabela de contingência "expandida".

Assim, de forma análoga, espera-se que na ACM a simples substituição dos valores "1" de cada caso na $\mathbf{M}_{\mathbf{ind}}$ pelo respectivo peso amostral seja suficiente para gerar estimativas pontuais dos perfis. Para verificar a validade desta proposta, consideramos a matriz de Burt, denotada aqui por $\mathbf{M}_{\mathbf{Burt}}$, que é uma transformação da $\mathbf{M}_{\mathbf{ind}}$ que gera o conjunto de todas as combinações de variáveis duas a duas. Portanto, já que a $\mathbf{M}_{\mathbf{Burt}}$ é somente uma estrutura alternativa à $\mathbf{M}_{\mathbf{ind}}$, levando a resultados iguais, e como a primeira pode ser obtida a partir da segunda, então as informações contidas em uma devem ser iguais às contidas na outra.

Logo, se for montada uma $\mathbf{M}_{\mathbf{Burt}}$ com as tabelas de contingência "expandidas", a $\mathbf{M}_{\mathbf{ind}}$ associada a esta $\mathbf{M}_{\mathbf{Burt}}$ também deve apresentar os mesmos resultados e também as mesmas propriedades em relação à consideração do plano amostral. Desta forma, uma vez que não é possível derivar algebricamente a $\mathbf{M}_{\mathbf{ind}}$ a partir da $\mathbf{M}_{\mathbf{Burt}}$ e dada a equivalência entre estas, assumimos que a argumentação acima é válida e apenas comparamos os resultados obtidos pelos dois caminhos.

Nesta comparação constatou-se que a $\mathbf{M}_{\mathbf{ind}}$ "expandida" gera corretamente a $\mathbf{M}_{\mathbf{Burt}}$ esperada e os resultados algébricos também são idênticos para ambas. Assim, como a $\mathbf{M}_{\mathbf{Burt}}$ "ponderada" produz estimativas dos totais na população, então, a $\mathbf{M}_{\mathbf{ind}}$ ajustada com os pesos amostrais também produz uma estimativa da $\mathbf{M}_{\mathbf{ind}}$ populacional. Com isso, estamos propondo que a incorporação das características do plano amostral em ACM seja realizada a partir da multiplicação de cada linha da $\mathbf{M}_{\mathbf{ind}}$ original pelo respectivo peso amostral.

## 6 Considerações Finais

As questões aqui discutidas demonstram que este tema é de grande relevância uma vez que a não

consideração do plano amostral na AC pode resultar em resultados de qualidade questionável. Entretanto, ainda há muito a ser avançado nesta área de investigação. Nossos próximos passos envolveram o estabelecimento formal de uma forma analítica para os resultados encontrados para a ACM que foram apresentados acima de maneira intuitiva. Outro ponto a ser abordado, que vale para todas as técnicas derivadas da AC, será encontrar um estimador para a variância amostral das projeções. No que se refere ao que não foi abordado aqui, vale destacar que há ainda várias outras técnicas estatísticas derivadas ou relacionadas à AC que também merecem atenção, tais como, análise de conjuntos de tabelas, modelos log-lineares, análise de componentes principais, escalonamento multidimensional, entre outros.

## 7 Referências


Benzécri, J.-P. (1992) *Correspondence analysis handbook*. New York: Marcell Dekker

Beh, E.J. (2002) *Correspondence Analysis in the Statistical Literature*. Research Report No. QMMS2004.15. School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Australia.

Greenacre, M. J. (1984) *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. J. (2007) *Correspondence Analysis in Practice*. 2 ed. Chapman & Hall/CRC.

Lehtonen, R.; Pahkinen, E. J. (1996) *Practical Methods for design and analysis of complex surveys*. John Wiley & Sons.

Nyfjäll, M. (2002) *On Correspondence Analysis under Complex Surveys Sampling Designs*. Thesis for the Degree of Licentiate of Philoso (Fil. Lic.) in Statistics, Uppsala University.

Pessoa, D G. C. e Silva, P. L. N. *Análise de Dados Amostrais Complexos*. 13º SINAPE, Associação Brasileira de Estatística.