Presidente da República
Luíz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Paulo Bernardo Silva

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor Executivo
Sérgio da Costa Côrtes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Wasmália Socorro Barata Bivar

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Luiz Fernando Pinto Mariano

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Sérgio da Costa Côrtes (interino)

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 67 número 226 janeiro/junho 2006

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 67, n. 226, p. 1-119, jan./jun. 2006

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Francisco Louzada-Neto
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Helem Ortega da Silva - Coordenação de Métodos e Qualidade - DPE/COMEQ/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2004.

Capa
Renato J. Aguiar – Coordenação de *Marketing*/CDDI/IBGE

Ilustração da Capa
Marcos Balster – Coordenação de *Marketing*/CDDI/IBGE

Revista brasileira de estatística/IBGE, - v.1, n.1 (jan./mar.1940), - Rio de Janeiro:IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística.

Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).

Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev.98) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor

Este é o primeiro volume da RBEs do ano de 2006, o qual é composto de quatro artigos.

O primeiro artigo propõe um método de otimização para calibração de pessoas em amostras. O segundo artigo apresenta os fatores que influenciaram na aprovação de candidatos de diferentes grupos socioeconômicos no Vestibular de 2004 da Universidade Federal de Minas Gerais. O terceiro artigo propõe a utilização da técnica *bootstrap* em delineamentos por conjuntos imperfeitamente ordenados. Finalmente, o quarto artigo aplica procedimentos multivariados de análise de correspondência múltipla e análise de agrupamento na redução de dimensionalidade de indicadores de eventos de vida.

Aproveito a oportunidade para agradecer aos Editores Associados do periódico, bem como aos revisores dos artigos submetidos e à equipe do IBGE.

Uma excelente leitura.

Francisco Louzada-Neto
Editor Responsável

Um método de otimização para calibração de pesos em amostras

José André de Moura Brito[*]
Adilson Elias Xavier[**]

Resumo

Este trabalho propõe uma nova metodologia de resolução para o problema de calibração dos pesos utilizados em expansão de amostras (Silva, 2004). Tal metodologia é baseada em uma formulação matemática que encapsula uma função distância quadrática e dois conjuntos de restrições lineares: igualdade e desigualdade. Para a resolução desta formulação, e a conseqüente obtenção dos pesos utilizados para ponderar cada unidade amostral (Cochran, 1977), utilizamos um método de otimização não linear proposto por Santos (1998). Com a finalidade de demonstrar as potencialidades e a aplicabilidade da nova metodologia, consideramos a problemática de cálculo de pesos (ajustados) no Censo Demográfico 2000, utilizando os dados do universo e da amostra. São apresentados os principais resultados computacionais e também um *software* com interface amigável desenvolvido para fácil manuseio.

Palavras-chave: Calibração, amostra, penalização.

[*] Endereço para correspondência: Instituto Brasileiro de Geografia e Estatística - IBGE/DPE/COMEQ. Av. República do Chile, nº 500, 10º Andar, Cep.: 21031-170 - Rio de Janeiro - RJ - e-mail: britom@ibge.gov.br
[**] Programa de Engenharia de Sistemas e Computação - COPPE/UFRJ. e-mail: adilson@cos.ufrj.br

1. Introdução

O presente trabalho propõe uma nova metodologia para a resolução do problema de calibração com pesos não-negativos, utilizando, como exemplo, os dados do Censo Demográfico 2000.

Como em muitos outros países, na coleta de informações do Censo Demográfico do Brasil (Metodologia do Censo Demográfico 2000, 2003) são utilizados dois modelos de questionários:

- Um questionário básico aplicado às unidades não-selecionadas para a amostra e contendo perguntas referentes às características que foram investigadas para 100% da população (universo); e
- Um segundo questionário (amostra) aplicado somente aos domicílios selecionados para a amostra contendo, além das perguntas que também constam do questionário básico, outras perguntas mais detalhadas sobre características do domicílio e de seus moradores, referentes aos temas religião, cor ou raça, deficiência, migração, escolaridade, fecundidade, nupcialidade, trabalho e rendimento, etc.

Para fins de modelagem e definição do problema (Pessoa, 2000), associa-se as informações comuns aos dois questionários às variáveis $X_1,...,X_p$, que são definidas como variáveis auxiliares ou explicativas e as informações contidas apenas no questionário do amostra são associadas às variáveis $Y_1,...,Y_q$.

Como exemplo de variáveis explicativas X, no Censo Demográfico 2000 tem-se o total de pessoas, o total de homens, o total de homens por faixa etária, etc. Como exemplo de variáveis Y, estão disponíveis informações de domicílios e pessoas: existência de televisores, cor ou raça, anos de estudo, etc.

Tendo em vista que as informações associadas às variáveis explicativas são conhecidas para toda a população, os totais populacionais $T_{X_1},...,T_{X_p}$ também são

conhecidos, embora seja possível calcular estimativas amostrais correspondentes $\hat{T}_{X_1},\ldots,\hat{T}_{X_p}$ [1] considerando somente a amostra, onde cada estimativa amostral é da forma

$$\hat{T}_{X_i} = \sum_{k=1}^{n} d_k x_{ki}, \qquad i = 1,\ldots \tag{1.1}$$

sendo d_k o peso inicial (de desenho amostral) para cada unidade k da amostra, geralmente igual ao inverso da respectiva probabilidade de seleção (Cochran, 1977). E x_{ki} está associada ao valor i-ésima variável auxiliar para a k-ésima unidade da amostra.

No caso deste trabalho, os valores conhecidos d_k e x_{ki} estão associados a cada um dos domicílios visitados no Censo .

Para as variáveis Y_j, são calculáveis somente estimativas amostrais, ou seja:

$$\hat{T}_{Y_j} = \sum_{k=1}^{n} d_k y_{kj}, \quad j = 1,\ldots \tag{1.2}$$

Normalmente, as estimativas amostrais $\hat{T}_{X_1},\ldots,\hat{T}_{X_p}$ diferem dos valores conhecidos $T_{X_1},\ldots,T_{X_p}$. Todavia, estas estimativas podem ser ajustadas de forma que fiquem iguais aos valores conhecidos dos totais populacionais, através da modificação dos pesos usados para ponderar cada unidade amostral. A esse procedimento denomina-se calibração dos pesos amostrais.

---

[1] Tanto os totais T_{X_i}, quanto as estimativas de totais $\hat{T}_{X_i}$, $\hat{T}_{Y_j}$ foram calculadas para cada área de ponderação definida no Censo Demográfico (Censo Demográfico 2000).

Observa-se, do ponto de vista prático, que ao calibrar os pesos utilizados para o cálculo destas estimativas, também são ajustadas as estimativas amostrais associadas às variáveis Y_j.

Com base nestas considerações, propõe-se então neste trabalho, uma nova metodologia para efetuar o ajuste dos pesos utilizados para ponderar unidades amostrais. Tal metodologia é baseada na adaptação de um método de otimização não-linear desenvolvido por Santos (1998).

O presente trabalho está dividido em cinco seções: Na seção 2 apresenta-se o problema de calibração em sua forma geral, propondo uma nova formulação para a sua resolução. Na seção 3 discute-se o método aplicado na resolução da nova formulação. Na seção 4 é apresentado um conjunto de resultados computacionais obtidos a partir do ajuste dos pesos da amostra do Censo . Para o ajuste destes pesos, foi utilizado um programa desenvolvido em linguagem Delphi (que implementa a nova metodologia) e o *software* GES (Estevao et al., 1995). Concluindo a exposição, no Anexo 1 é apresentado o programa desenvolvido com base na nova metodologia.

2. Definição do problema de calibração e apresentação da metodologia

O problema clássico de calibração é definido da seguinte forma: selecionar um conjunto "ótimo" de pesos $\{w_k = d_k.g_k, k = 1,...,n\}$ (n = número de unidades amostrais), entre todos os conjuntos de pesos aceitáveis, tais que sejam satisfeitas as equações (restrições) de calibração abaixo:

$$\sum_{k=1}^{n} w_k x_{ki} = T_{X_i}, \quad \forall i = 1,...,p \tag{2.1}$$

Quando temos $p = 1$, ou seja, apenas uma variável explicativa X, a solução do problema de calibração é trivial e consiste em calcular $w_k = (T_X / \hat{T}_X) d_k$. No caso de duas ou mais variáveis, a definição de novos pesos torna-se complicada, pois deve-se ter um único conjunto de pesos para todas as variáveis X_i.

A racionalidade do processo de calibração, isto é, satisfazer as equações do tipo (2.1), está associada aos seguintes fatos (Silva, 2004):

- Ao se efetuar o cálculo dos pesos w_k de forma a satisfazer as equações do tipo (2.1), produz-se, em conseqüência, estimativas dos totais das variáveis Y_j que deverão levar a estimadores "melhores" que os baseados nos pesos de desenho iniciais (d_k), ou seja, estimadores com melhor precisão;
- Os estimadores de calibração são lineares. Isto significa que o registro de cada questionário pode ser associado a um peso simples que será utilizado para estimação de todas as variáveis deste questionário;
- Flexibilidade de incorporar a informação auxiliar associada à variáveis do tipo contínuo, discreto ou mistas; e
- Os estimadores de calibração também podem oferecer alguma proteção contra o vício de não-resposta.

Tendo em vista que os pesos d_k são inicialmente fornecidos, a determinação dos pesos w_k corresponde, então, à escolha dos ajustes g_k, considerando a definição de algum critério de otimalidade. Um critério normalmente adotado é de que os pesos finais calibrados fiquem "próximos" dos pesos iniciais d_k, de forma a reduzir o vício do desenho (Silva, 2004).

Uma possível solução para o problema é obtida minimizando-se alguma função de distância entre os pesos iniciais d_k e os pesos finais ajustados w_k $(d_k \cdot g_k)$, sujeitos às equações de calibração (2.1). A função de uso mais comum é:

$$f(g) = \sum_{k=1}^{n} d_k (1 - g_k)^2, \quad g \in \Re^n \tag{2.2}$$

Desta forma, os ajustes g_k e conseqüentemente os pesos finais w_k, podem ser obtidos, inicialmente, através da resolução de uma formulação que encapsula a função objetivo quadrática apresentada na equação (2.2) e as restrições lineares de igualdade (2.1), ou seja, a seguinte formulação de programação não-linear:

Minimizar

$$f(g) = \sum_{k=1}^{n} d_k (1 - g_k)^2 \tag{2.3}$$

Sujeito a

$$\sum_{k=1}^{n} d_k g_k x_{ki} = T_{X_i}, \quad \forall i = 1,\ldots,p \tag{2.4}$$

Pela equação (2.3), pode-se observar que quanto mais próximos de 1 estiverem os ajustes g_k , menor será o valor da função objetivo, ou seja, mais próximos de d_k estarão os pesos w_k.

Um fato bem conhecido é que ao resolver um problema definido pelo uso da função do tipo 2.3 em conjunção com as restrições lineares do tipo 2.4, pode-se obter pesos negativos ou muito grandes. O que não seria aceitável do ponto de vista da aplicação real.

Com a finalidade de contornar esta deficiência, há na literatura um conjunto de metodologias alternativas para a resolução do problema de calibração.

- Pessoa (2005) fez um estudo detalhado do problema de calibração e implementou um conjunto de rotinas em R. Através destas rotinas é possível efetuar a calibração, escolhendo entre quatro tipos de funções (associadas com a distância entre os pesos d_k e os pesos finais ajustados w_k) e calcular os erros para estimativas oriundas dos pesos calibrados;

- Rao (1999) utilizou um método iterativo de regressão para resolver o problema de calibração;
- Bankier (1995) utilizou o método dos mínimos quadrados para efetuar a calibração dos pesos no Censo do Canadá. Durante a aplicação deste método, são descartadas as restrições "pequenas" e "linearmente dependentes", associadas a algumas variáveis auxiliares; e
- Deville e Särndal (1992) consideraram algumas funções especiais para calibração e estudaram as propriedades destas funções.

No caso da metodologia proposta neste trabalho, com a finalidade de contornar esta deficiência, agrega-se à formulação definida por (2.3-2.4) o seguinte conjunto de restrições:

$$a \leq w_k = d_k . g_k \leq b \quad , k = 1,\ldots,n \tag{2.5}$$

Estas restrições indicam que os ajustes g_k calculados e multiplicados pelos seus respectivos pesos iniciais do desenho amostral, estarão compreendidos entre os valores $a > 0$ e $b > 0$.

Com base na consideração dessas novas restrições, podemos então escrever a seguinte formulação:

Minimizar

$$f(g) = \sum_{k=1}^{n} d_k (1 - g_k)^2 \tag{2.6}$$

Sujeito a

$$\sum_{k=1}^{n} d_k g_k x_{ki} = T_{X_i}, \quad \forall i = 1,\ldots,p \tag{2.7}$$

$$a \leq d_k . g_k \leq b \ , \quad k = 1,\ldots,n \tag{2.8}$$

$$g \in \Re^n$$

Para a resolução de (2.6-2.8), será utilizado o método de penalização hiperbólica proposto por Santos (1998). Ao contrário dos métodos ortodoxos de penalização (Bazaraa et al., 1993), este método possibilita resolver problemas que agreguem restrições de desigualdade e igualdade, como as restrições do tipo (2.7) e (2.8) que aparecem na formulação acima e que são o diferencial para o cálculo das estimativas dos totais.

De forma a facilitar o entendimento do método de penalização hiperbólica, pode-se reescrever a formulação proposta na forma padrão de programação não-linear:

Minimizar

$$f(g) \tag{2.9}$$

Sujeito a

$$h_i(g) = 0\,, \qquad \forall\, i = 1,\ldots,p \tag{2.10}$$

$$m_k(g) \geq 0, \qquad k = 1,\ldots,n \tag{2.11}$$

$$s_k(g) \geq 0, \qquad k = 1,\ldots,n \tag{2.12}$$

$$g \in \Re^n,$$

sendo,

$$h_i(g) = \sum_{k=1}^{n} d_k g_k x_{ki} - T_{X_i}, \quad \forall\, i = 1,\ldots,p \tag{2.13}$$

$$m_k(g) = b \; - d_k g_k, \quad k = 1,\ldots,n \tag{2.14}$$

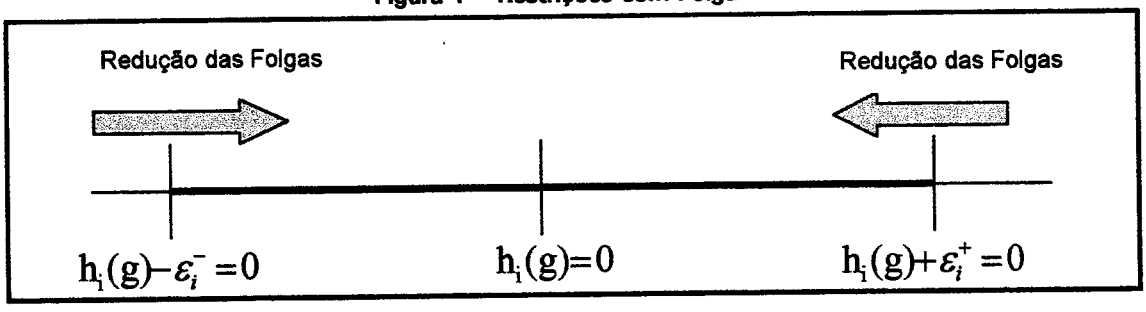
$$s_k(g) = d_k g_k - a\,, \qquad k = 1,\ldots,n \tag{2.15}$$

De uma forma resumida, o método de penalização hiperbólica transforma as restrições de igualdade (2.10), em restrições de desigualdade, considerando uma folga (tolerância) ε_i, $i = 1,\ldots,p$, que será definida *a priori* para cada restrição do tipo (2.10).

Uma vez definida esta folga inicial, em cada iteração do processo de minização (resolução da formulação acima), a folga ε_i será reduzida para cada uma das restrições (Figura 1), com o objetivo de que ao final de um certo número de iterações k para

cada um dos ε_i tenha-se $\varepsilon_i \cong 0$, ou seja, $h_i(g) = 0$.

Figura 1 – Restrições com Folga



A seguir, apresenta-se a formulação modificada considerando uma folga (tolerância) para cada restrição do tipo (2.10).

Minimizar

$$f(g) \tag{2.16}$$

Sujeito a

$$h_i(g) \geq \varepsilon_i^- , \ \forall\, i = 1,\ldots,p \tag{2.17}$$

$$h_i(g) \leq \varepsilon_i^+ , \ \forall\, i = 1,\ldots,p \tag{2.18}$$

$$m_k(g) \geq 0, \quad k = 1,\ldots,n \tag{2.19}$$

$$s_k(g) \geq 0, \quad k = 1,\ldots,n \tag{2.20}$$

$$\varepsilon_i^+ \geq 0 \quad \text{e} \quad \varepsilon_i^- \leq 0$$

Similarmente a outros métodos de penalização (Bazaraa et al., 1993), transforma-se o problema (2.16-2.20), inicialmente com restrições, em um problema de programação não-linear irrestrito. A seguir, é descrito este processo.

Através da utilização da função de penalização hiperbólica $P(y,\alpha,\tau)$ (Santos,

1998) agrega-se todas as restrições de (2.17-2.20) e a função objetivo original $f(g)$, obtendo uma nova função objetivo $F(g,\alpha,\tau)$, ou seja, resolvendo o seguinte problema irrestrito:

Minimizar

$$F(g,\alpha,\tau) = f(g) + \sum_{i=1}^{2p} P(H_i(g),\alpha^i,\tau) + \sum_{k=1}^{n} P(m_k(g),\alpha^{2p+k},\tau) + \sum_{k=1}^{n} P(s_k(g),\alpha^{2p+k},\tau) \quad (2.21)$$

sendo,

$$H_i(g) = h_i(g) - \varepsilon_i^- \geq 0, \quad i = 1,\ldots, p \quad (2.22)$$

$$H_{p+i}(g) = -h_i(g) + \varepsilon_i^+ \geq 0, \quad i = 1,\ldots, p \quad (2.23)$$

$$P(y,\alpha,\tau) = \frac{-1}{2} tg(\alpha).y + \sqrt{(\frac{1}{2} tg(\alpha).y)^2 + \tau^2}$$

Nesta função, o parâmetro y está associado à restrição que se deseja penalizar, o parâmetro α está associado ao tradicional efeito da penalização externa, enquanto o parâmetro τ está mais vinculado ao tradicional efeito da penalização interna (Bazaraa et al., 1993). O algoritmo de penalização proposto para resolver o problema de calibração ou qualquer outro problema de programação não-linear restrito é composto por duas fases:

Inicialmente, aumenta-se o ângulo α da função penalidade, até que se obtenha uma solução viável. No caso deste trabalho, até que se obtenha um conjunto de ajustes g_k tais que todas as restrições do problema (2.10–2.12) sejam satisfeitas. Uma vez atingida a viabilidade, mantém-se α constante e diminui-se seqüencialmente o valor de τ.

A seguir, apresenta-se, na Figura 2 o pseudocódigo simplificado do algoritmo proposto para resolver o problema de calibração.

Figura 2 - Algoritmo para calibração

(1) Defina Número de Iterações do Algoritmo (Tot_Iter)

(2) Defina $\alpha = \frac{\pi}{3}$, $\tau = 100,\ k = 0$

(3) Defina $g^0 = g$ (valor inicial para os ajustes)

(4) Defina as folgas ε_i^- , $\varepsilon_i^+, i = 1,\ldots, p$

(5) Enquanto (k < Tot_Iter) Faça

(6) Faça k = k + 1

(7) Resolva o problema {*Minimizar* F(g,α,τ) } a partir de g^{k-1} obtendo g^k

(8) Se g^k satisfaz todas as restrições (viável) Então

(9) Atualize τ , Atualize as folgas ε_i^- , ε_i^+ das restrições (redução das folgas)

(10) Senão

(11) Aumente o ângulo α das restrições violadas

(12) Fim-se

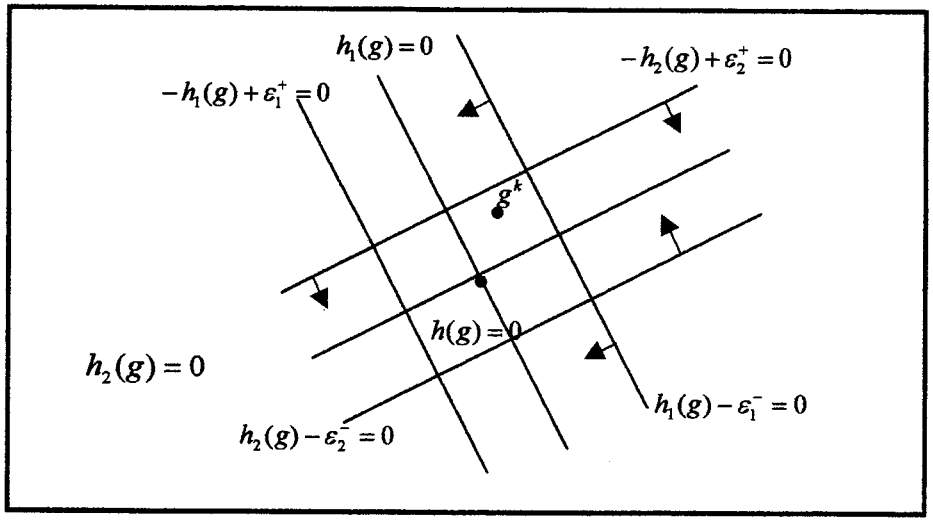
(13) Fim-Enquanto

3. Discussão do algoritmo

A estrutura do algoritmo é constituída de tal forma que, para cada restrição, o intervalo entre as folgas ε_i^+ e $-\varepsilon_i^-$ seja reduzido monotonamente. Este procedimento tem como perspectiva forçar uma aproximação gradativa dos pontos de mínimos intermediários g^k (obtidos em cada iteração do algoritmo) ao ponto ótimo g^*.

O valor g^* possibilita o menor valor para $f(g) = \sum_{k=1}^{n} d_k (1 - g_k)^2$ e satisfaz todas as restrições do tipo $\sum_{k=1}^{n} d_k g_k x_{ki} = T_{X_i}$, $\forall i = 1,\ldots, p$ e $a \le d_k . g_k \le b$, $k = 1,\ldots, n$.

Figura 3 – Redução do intervalo entre as folgas



A Figura 3 dá uma idéia de como a região viável associada às restrições de igualdade, relaxadas pelas folgas ε vai, gradativamente, se aproximando do espaço viável definido por $S=\{g\in\Re^n / h(g)=0,\ i=1,...,p\}$, quando estas folgas são diminuídas, provocando, dessa maneira, os deslocamentos indicados pelas setas.

Outra componente fundamental do algoritmo é o processo de aumento do ângulo α (quando necessário) com o objetivo de se obter pontos de ótimo intermediários g^k viáveis.

A última componente do algoritmo, também fundamental, é a diminuição do parâmetro τ, com o objetivo de se aproximar o problema penalizado (definido pelas equações 2.16-2.20) gradativamente do problema original (definido pelas equações 2.6-2.8).

4. Resultados computacionais e conclusões

De forma a avaliar a nova metodologia proposta neste trabalho, é apresentado, nesta seção, um conjunto de resultados computacionais obtidos a partir da utilização dos dados do Censo Demográfico 2000.

Para tal análise, foram escolhidas 15 dentre as 27 Unidades da Federação, listadas

a seguir: Acre, Alagoas, Bahia, Ceará, Espírito Santo, Goiás, Mato Grosso do Sul, Minas Gerais, Paraná, Pernambuco, Piauí, Rio de Janeiro, Rio Grande do Sul, São Paulo, Santa Catarina.

Em seguida, foram selecionadas para cada uma destas unidades cinco áreas de ponderação, que foi o domínio de calibração considerado no Censo . Tal seleção foi efetuada a partir da utilização de um programa desenvolvido em linguagem SAS.

As áreas de ponderação são formadas por setores censitários, que por sua vez, são agregados de domicílios.

O algoritmo proposto neste trabalho, foi então aplicado em cada uma das áreas de ponderação, considerando como unidade amostral os domícilios presentes em cada área, ou seja, foram calculados os pesos w_k de cada um dos domicílios associados às áreas de ponderação (domínio de calibração).

O valor do peso do desenho d_k, foi determinado por setor, efetuando-se o cálculo da razão $\frac{D_{Univ}}{D_{Amos}}$, sendo D_{Univ} o total de domicílios do universo para o setor em questão e D_{Amos} o total de domicílios da amostra para o mesmo setor (todos os domicílios associados a um determinado setor j têm um mesmo peso d_k).

Os valores de a e b, que correspondem, respectivamente, ao limite mínimo e ao limite máximo para os pesos finais w_k (calibrados), foram definidos para cada área de ponderação, levando-se em conta a fração amostral planejada dos municípios.

O limite mínimo utilizado foi 1, de maneira que um domicílio representasse pelo menos o próprio. O limite máximo foi definido como cinco vezes o peso médio esperado, ou seja, 25 no caso das áreas de ponderação associadas a municípios grandes (mais de 15 000 habitantes) com fração amostral planejada de 20% (caso em que o peso médio esperado era de 5) e 50 no caso das áreas de ponderação associadas aos municípios pequenos (número de habitantes menor ou igual a 15 000) com fração amostral planejada de 10% (caso em que o peso médio esperado era de 10). Sem a utilização desses limites haveria a possibilidade de se gerar pesos negativos ou muito grandes, o que não teria sentido prático.

Para o ajuste dos pesos em cada área de ponderação, foi utilizado um conjunto de 38 variáveis explicativas X_i que corresponderam ao número total de pessoas, número total de unidades domiciliares, número de pessoas do sexo masculino, número de pessoas do sexo masculino por faixa etária, número de pessoas do sexo feminino, número de pessoas do sexo feminino por faixa etária (Censo Demográfico 2000).

Além da utilização das variáveis auxiliares descritas acima, foi escolhido *a priori* um conjunto variáveis Y_j (descritas abaixo) para efetuar o cálculo de estimativas, considerando os pesos obtidos após o processo de calibração.

- V0213: Existência de iluminação elétrica, proveniente ou não de uma rede geral, com ou sem medidor ou relógio que registre o consumo exclusivo do domicílio.

- V0219: Existência de linha telefônica convencional, ainda que alugada, extensão ou ramal de centrais telefônicas no domicílio.

- V0221: Número total de televisores existentes no domicílio, tanto em cores como em preto e branco, desde que em condições de uso, expresso em classes de valores.

4.1 Resultados cmputacionais

Apresenta-se, nesta seção, um conjunto de tabelas com resultados obtidos a partir da utilização do programa PCALIBRA (desenvolvido em *Delphi* e que implementa a nova metodologia) e do *software* GES. Este *software* efetua o ajuste dos pesos (processo de calibração) considerando a teoria baseada no estimador de regressão generalizado (*generalized regression estimator*). As principais funções deste *software* são: cálculo dos pesos do desenho da amostra, cálculo dos ajustes g_k, cálculo de estimativas considerando os pesos calibrados, estimativas de totais, médias, razões, etc. Informações detalhadas sobre a metodologia utilizada no *software* GES são apresentadas no trabalho de Estevao et al. (1995).

Os dois *softwares* foram executados num computador *Pentium Celeron* 600 Mhz, com 196 Mb de memória.

Para cada uma das UFs foi construída uma tabela com as seguintes informações:

- Número de domicílios (observações) presentes em cada área de ponderação (na amostra) selecionada.
- Valor do erro médio relativo, para estimar os totais da população.

$$M_1 = \frac{1}{p}\sum_{i=1}^{p} | \hat{T}_{X_i} - T_{X_i} | / T_{X_i}$$

- Coeficiente de variação para estimar a distribuição dos ajustes g_k.

$$M_2 = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(g_k - \bar{g})^2} / \bar{g}$$

- Valor final da função objetivo descrita na seção 2, considerando os ajustes g_k, obtidos a partir do algoritmo de penalização e do GES. Tal função foi utilizada para estimar a distância entre o peso inicial do desenho d_k e os pesos finais ajustados.

$$M_3 = \frac{1}{n}\sum_{k=1}^{n} d_k (1 - g_k)^2$$

- Coeficientes de variação dos totais estimados para as variáveis Y.

$$M_4 = CV(\hat{T}_{Y_C})$$

As quatro medidas são apresentadas no trabalho de Silva (2004) e possibilitam avaliar a qualidade do processo de calibração derivado do método proposto neste trabalho e implementado no programa PCALIBRA e do método implementado no *software* GES.

Pelas tabelas apresentadas a seguir, pode-se fazer as seguintes considerações:

- Em 76% das áreas de ponderação, para as quais foi aplicado o processo de calibração, o valor da medida M_3, obtida a partir do programa PCALIBRA foi menor que o valor da medida M_3 utilizando o *software* GES. Isto indica, que os pesos w_k calibrados fornecidos a partir da nova metodologia estão mais próximos dos pesos iniciais do desenho amostral, o que é desejável (Silva, 2004).
- Os valores de M_1 obtidos a partir dos dois métodos foram razoavelmente pequenos, o que indica que os pesos w produzidos pelo PCALIBRA e pelo GES forneceram totais estimados $\hat{T}_{X_i}$ iguais ou bem próximos dos totais conhecidos T_{X_i}.
- Em 84% das áreas de ponderação, observou-se que a medida M_2 derivada do programa PCALIBRA foi menor que a medida M_2 do GES, o que implica uma distribuição melhor dos ajustes g_k calculados pelo método proposto neste trabalho.

Observamos, ainda, que há a possibilidade de se melhorar as estimativas dos totais e, por conseqüência os valores de M_1 e M_2, através da retirada de restrições pequenas (com poucas obsevações x_{ki} diferentes de zero na população), seguindo os critérios definidos no trabalho de Bankier (1995).

Este tipo de restrição tem um efeito de desbalanceamento indesejável nas restrições de igualdade $\sum_{k=1}^{n} d_k g_k x_{ki} = T_{X_i}, \forall i = 1,\ldots,p$, produzindo totais estimados diferentes dos totais conhecidos, ou seja, $|T_{X_i} - \hat{T}_{X_i}| \neq 0$.

- O algoritmo proposto mostrou robustez e estabilidade, tendo em vista o número de variáveis (observações da amostra) consideradas no processo de calibração de cada área de ponderação.
- Todos os pesos w_k calculados ficaram dentro das faixas definidas por a e b, o que não ocorreu em alguns casos com a utilização do GES, onde foi observado o indesejável aparecimento de alguns pesos negativos.

Tabelas com resultados dos programas PCALIBRA e do *software* GES

Tabela 1 - Acre - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	4,26E-04	2,04E-05	6,78E-04	1,44E-04	1,01E-04
	PCALIBRA	2,45E-04	1,60E-05	7,21E-04	8,59E-04	4,51E-04
M_2	GES	28,20	21,75	42,55	36,97	22,38
	PCALIBRA	27,92	21,82	39,55	36,98	22,52
M_3	GES	462,43	133,71	939,61	167,48	259,29
	PCALIBRA	453,81	134,89	814,14	166,97	263,16
M_4 - *V0213*	GES	2,55%	1,49%	1,75%	3,03%	1,63%
	PCALIBRA	2,56%	1,48%	1,75%	3,04%	1,62%
M_4 - *V0219*	GES	20,39%	6,91%	8,18%	9,61%	7,56%
	PCALIBRA	20,25%	6,91%	7,98%	9,54%	7,58%
M_4 - *V0221*	GES	3,45%	2,56%	3,08%	4,80%	3,29%
	PCALIBRA	3,46%	2,57%	2,94%	4,80%	3,29%
Observações na amostra		589	561	550	236	520

Tabela 2 - Alagoas - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	6,76E-05	3,41E-04	1,07E-03	5,06E-03	7,59E-04
	PCALIBRA	2,46E-05	1,29E-04	3,97E-03	5,78E-03	1,94E-04
M_2	GES	23,21	19,91	18,46	26,97	50,67
	PCALIBRA	23,19	19,92	17,97	35,59	47,32
M_3	GES	259,33	282,56	248,31	342,99	1274,28
	PCALIBRA	266,20	282,60	241,10	595,86	1106,13
M_4 - *V0213*	GES	1,50%	0,00%	0,31%	0,00%	1,98%
	PCALIBRA	1,50%	0,00%	0,28%	0,00%	1,86%
M_4 - *V0219*	GES	26,34%	4,36%	3,64%	2,02%	15,70%
	PCALIBRA	26,61%	4,44%	3,51%	2,19%	15,82%
M_4 - *V0221*	GES	2,76%	1,24%	1,13%	0,71%	3,88%
	PCALIBRA	2,76%	1,34%	1,12%	0,79%	3,71%
Observações na amostra		506	714	739	466	588

Tabela 3 - Bahia - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	1,95E-03	1,25E-03	3,98E-03	2,45E-04	3,54E-05
	PCALIBRA	4,36E-03	3,41E-03	7,44E-03	2,08E-04	8,07E-05
M_2	GES	24,52	32,92	27,07	22,64	25,52
	PCALIBRA	26,85	32,14	26,31	22,03	25,23
M_3	GES	190,01	445,85	404,94	181,20	308,53
	PCALIBRA	228,26	419,73	382,70	171,82	305,24
M_4 - *V0213*	GES	0,82%	0,00%	0,10%	1,18%	3,07%
	PCALIBRA	0,80%	0,00%	0,09%	1,18%	3,06%
M_4 - *V0219*	GES	13,28%	1,21%	3,62%	21,15%	21,40%
	PCALIBRA	13,44%	1,20%	3,63%	21,11%	21,48%
M_4 - *V0221*	GES	2,05%	0,55%	1,39%	2,11%	3,61%
	PCALIBRA	2,04%	0,54%	1,39%	2,11%	3,61%
Observações na amostra		599	419	570	733	531

Tabela 4 -Ceará - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	3,26E-04	2,03E-04	6,83E-05	2,14E-04	2,78E-05
	PCALIBRA	1,44E-04	8,18E-04	1,96E-05	9,99E-04	2,99E-05
M_2	GES	28,19	22,90	11,91	19,71	25,18
	PCALIBRA	27,32	22,83	11,74	19,35	23,15
M_3	GES	301,14	316,30	278,86	135,41	276,04
	PCALIBRA	284,96	317,15	272,49	130,43	234,04
M_4 - *V0213*	GES	1,65%	1,48%	0,16%	1,71%	2,10%
	PCALIBRA	1,65%	1,49%	0,16%	1,71%	2,07%
M_4 - *V0219*	GES	11,78%	11,97%	2,67%	10,08%	15,90%
	PCALIBRA	11,78%	11,99%	2,67%	10,09%	15,89%
M_4 - *V0221*	GES	2,25%	2,15%	0,63%	2,33%	3,95%
	PCALIBRA	2,25%	2,15%	0,63%	2,33%	3,90%
Observações na amostra		421	650	1983	718	496

Tabela 5 - Espírito Santo - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	3,94E-04	1,06E-04	5,02E-04	5,66E-04	5,02E-05
	PCALIBRA	3,37E-04	3,51E-04	1,92E-04	7,37E-04	8,25E-05
M_2	GES	22,53	32,11	23,94	27,77	22,74
	PCALIBRA	21,64	31,03	21,53	25,15	22,50
M_3	GES	233,76	521,68	169,57	377,32	240,18
	PCALIBRA	216,49	503,35	137,68	308,70	236,28
M_4 - *V0213*	GES	0,38%	0,09%	0,37%	0,62%	0,23%
	PCALIBRA	0,37%	0,08%	0,33%	0,62%	0,23%
M_4 - *V0219*	GES	27,77%	5,85%	6,54%	10,92%	5,10%
	PCALIBRA	27,85%	5,87%	6,51%	10,80%	5,10%
M_4 - *V0221*	GES	1,90%	1,04%	1,23%	1,76%	1,10%
	PCALIBRA	1,90%	1,04%	1,20%	1,74%	1,09%
Observações na amostra		477	535	599	493	465

Tabela 6 - Goiás - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	1,13E-05	1,77E-03	2,15E-06	7,17E-03	1,96E-03
	PCALIBRA	4,75E-05	6,39E-03	1,65E-06	9,17E-03	8,52E-03
M_2	GES	27,76	24,67	18,67	38,77	59,77
	PCALIBRA	27,72	21,97	18,63	42,25	39,64
M_3	GES	294,07	331,27	227,91	133,54	2446,62
	PCALIBRA	298,25	262,55	227,33	160,51	1102,96
M_4 - *V0213*	GES	0,67%	0,00%	0,00%	0,87%	0,72%
	PCALIBRA	0,67%	0,00%	0,00%	0,84%	0,50%
M_4 - *V0219*	GES	11,25%	2,25%	0,94%	9,02%	5,28%
	PCALIBRA	11,26%	2,23%	0,94%	8,81%	5,07%
M_4 - *V0221*	GES	1,77%	1,02%	0,32%	2,62%	1,42%
	PCALIBRA	1,77%	1,01%	0,32%	2,43%	1,23%
Observações na amostra		438	533	645	177	828

Tabela 7 - Mato Grosso do Sul - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	1,03E-04	1,29E-03	4,11E-03	9,82E-04	1,85E-05
	PCALIBRA	5,62E-04	5,62E-04	5,62E-04	5,62E-04	5,62E-04
M_2	GES	29,94	32,35	22,50	27,31	16,21
	PCALIBRA	28,46	30,63	21,96	24,12	15,74
M_3	GES	401,33	443,36	389,32	332,23	82,37
	PCALIBRA	362,44	397,79	371,93	260,10	78,31
M_4 - *V0213*	GES	0,93%	0,21%	0,00%	1,21%	0,61%
	PCALIBRA	0,94%	0,22%	0,00%	1,19%	0,61%
M_4 - *V0219*	GES	6,97%	2,69%	0,87%	8,09%	6,13%
	PCALIBRA	6,98%	2,65%	0,83%	8,01%	6,12%
M_4 - *V0221*	GES	1,80%	1,13%	0,33%	2,08%	1,64%
	PCALIBRA	1,81%	1,09%	0,29%	2,05%	1,64%
Observações na amostra		461	427	787	452	657

Tabela 8 - Minas Gerais - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	6,14E-03	8,10E-04	1,73E-04	1,24E-04	1,26E-03
	PCALIBRA	1,11E-03	1,75E-04	8,34E-04	1,02E-04	1,30E-03
M_2	GES	73,25	49,19	21,29	26,71	30,98
	PCALIBRA	36,75	44,08	21,16	26,71	30,02
M_3	GES	202,51	220,80	278,03	201,06	173,17
	PCALIBRA	52,37	176,30	274,31	202,79	162,80
M_4 - *V0213*	GES	1,48%	2,36%	0,00%	0,00%	0,44%
	PCALIBRA	1,44%	2,35%	0,00%	0,00%	0,44%
M_4 - *V0219*	GES	16,35%	28,79%	2,10%	5,52%	8,49%
	PCALIBRA	16,12%	27,95%	2,10%	5,54%	8,34%
M_4 - *V0221*	GES	3,39%	4,27%	0,62%	0,82%	1,41%
	PCALIBRA	3,40%	4,28%	0,62%	0,82%	1,44%
Observações na amostra		83	168	596	584	385

Tabela 9 - Paraná - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	7,90E-05	3,82E-04	1,73E-05	3,36E-03	1,93E-03
	PCALIBRA	3,44E-05	3,34E-04	3,43E-05	7,44E-03	5,85E-03
M_2	GES	19,40	18,31	22,59	37,99	29,55
	PCALIBRA	19,04	17,55	22,41	29,32	13,62
M_3	GES	283,73	272,43	173,94	81,28	368,61
	PCALIBRA	276,32	251,51	171,48	46,89	78,42
M_4 - *V0213*	GES	0,66%	0,33%	1,07%	0,63%	0,20%
	PCALIBRA	0,66%	0,33%	1,06%	0,63%	0,22%
M_4 - *V0219*	GES	5,71%	5,69%	9,01%	16,72%	7,26%
	PCALIBRA	5,71%	5,70%	9,01%	16,70%	7,03%
M_4 - *V0221*	GES	1,59%	1,18%	1,81%	2,60%	1,31%
	PCALIBRA	1,59%	1,17%	1,81%	2,59%	1,34%
Observações na amostra		778	829	703	125	430

Tabela 10 - Pernambuco - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	2,70E-05	2,23E-05	2,15E-03	5,78E-07	5,04E-04
	PCALIBRA	1,61E-05	1,41E-05	8,16E-03	3,05E-07	8,83E-04
M_2	GES	22,35	24,96	28,76	18,06	24,55
	PCALIBRA	22,32	24,54	20,24	18,11	23,50
M_3	GES	329,60	277,47	406,37	203,88	374,10
	PCALIBRA	329,22	268,31	201,69	207,18	346,09
M_4 - *V0213*	GES	0,46%	0,00%	0,00%	0,18%	0,00%
	PCALIBRA	0,47%	0,00%	0,00%	0,18%	0,00%
M_4 - *V0219*	GES	12,85%	6,68%	1,84%	6,08%	1,82%
	PCALIBRA	12,84%	6,67%	1,77%	6,07%	1,83%
M_4 - *V0221*	GES	1,49%	1,02%	0,64%	1,05%	0,56%
	PCALIBRA	1,49%	1,02%	0,66%	1,05%	0,57%
Observações na amostra		682	455	485	618	609

Tabela 11 - Piauí - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	1,69E-03	2,85E-04	1,15E-04	2,04E-03	8,85E-04
	PCALIBRA	6,15E-03	9,22E-04	8,38E-04	5,55E-04	6,54E-04
M_2	GES	41,16	18,93	21,20	37,97	76,49
	PCALIBRA	35,80	19,11	20,59	38,28	68,71
M_3	GES	168,50	163,22	277,36	226,89	439,43
	PCALIBRA	127,37	171,04	262,94	229,97	355,47
M_4 - *V0213*	GES	5,29%	1,31%	2,68%	3,32%	8,52%
	PCALIBRA	5,24%	1,32%	2,68%	3,31%	8,22%
M_4 - *V0219*	GES	21,48%	12,24%	23,03%	21,53%	127,68%
	PCALIBRA	20,88%	12,34%	22,96%	21,64%	70,18%
M_4 - *V0221*	GES	7,73%	2,28%	4,55%	6,15%	14,27%
	PCALIBRA	7,48%	2,29%	4,54%	6,14%	14,20%
Observações na amostra		200	486	624	335	129

Tabela 12 - Rio de Janeiro - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	2,60E-04	6,23E-05	7,50E-04	2,35E-05	3,25E-04
	PCALIBRA	9,39E-04	1,41E-05	5,19E-04	3,05E-05	3,71E-04
M_2	GES	19,35	21,33	28,65	18,81	26,49
	PCALIBRA	19,01	21,26	26,75	18,59	24,17
M_3	GES	352,87	320,49	531,83	290,21	296,79
	PCALIBRA	333,45	320,42	450,09	283,91	249,17
M_4 - *V0213*	GES	0,00%	0,00%	0,99%	0,00%	0,13%
	PCALIBRA	0,00%	0,00%	1,00%	0,00%	0,14%
M_4 - *V0219*	GES	5,61%	1,83%	8,26%	4,53%	2,23%
	PCALIBRA	5,61%	1,82%	8,15%	4,53%	2,21%
M_4 - *V0221*	GES	0,45%	0,32%	1,68%	0,54%	0,60%
	PCALIBRA	0,44%	0,32%	1,64%	0,54%	0,59%
Observações na amostra		990	706	724	833	424

Tabela 13 - Rio Grande do Sul - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	5,43E-03	6,80E-05	3,04E-03	6,12E-04	9,49E-03
	PCALIBRA	3,63E-03	7,80E-05	1,54E-03	1,10E-04	2,39E-03
M_2	GES	88,77	19,86	31,32	22,14	68,92
	PCALIBRA	58,70	18,11	28,15	21,44	49,95
M_3	GES	382,37	287,14	179,42	244,20	400,95
	PCALIBRA	160,67	239,13	160,33	229,52	216,32
M_4 - *V0213*	GES	0,00%	0,71%	1,36%	0,22%	1,50%
	PCALIBRA	0,00%	0,71%	1,45%	0,23%	1,55%
M_4 - *V0219*	GES	4,11%	3,12%	6,73%	3,47%	9,77%
	PCALIBRA	4,35%	3,12%	6,70%	3,47%	10,04%
M_4 - *V0221*	GES	2,05%	1,01%	2,32%	0,75%	4,39%
	PCALIBRA	1,70%	1,00%	2,38%	0,76%	4,23%
Observações na amostra		101	730	354	499	172

Tabela 14 - São Paulo - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	4,65E-05	1,82E-04	8,08E-05	4,73E-05	3,60E-04
	PCALIBRA	1,89E-05	2,09E-04	1,29E-05	5,44E-05	1,43E-04
M_2	GES	23,36	24,50	25,08	19,83	35,28
	PCALIBRA	23,41	23,97	23,71	19,56	35,12
M_3	GES	413,45	363,35	269,58	202,61	411,31
	PCALIBRA	415,05	348,33	243,02	197,45	417,34
M_4 - *V0213*	GES	0,11%	0,00%	0,00%	0,19%	0,69%
	PCALIBRA	0,11%	0,00%	0,00%	0,20%	0,68%
M_4 - *V0219*	GES	1,49%	1,16%	1,22%	3,24%	11,01%
	PCALIBRA	1,49%	1,16%	1,20%	3,24%	11,01%
M_4 - *V0221*	GES	0,46%	0,49%	0,44%	1,02%	1,49%
	PCALIBRA	0,46%	0,49%	0,45%	1,03%	1,47%
Observações na amostra		750	621	422	510	429

Tabela 15 - Santa Catarina - Medidas de qualidade da calibração para áreas de ponderação e variáveis selecionadas, utilizando os programas GES e PCALIBRA

Medida de qualidade da calibração	Programa utilizado	Áreas de ponderação				
		1	2	3	4	5
M_1	GES	1,98E-04	2,66E-04	7,24E-04	8,21E-04	2,58E-04
	PCALIBRA	3,36E-02	7,41E-04	2,86E-04	1,98E-04	1,10E-04
M_2	GES	32,74	27,90	20,29	27,86	58,32
	PCALIBRA	37,03	27,72	21,00	28,81	43,99
M_3	GES	183,30	169,13	317,63	192,52	167,91
	PCALIBRA	243,23	166,89	347,00	207,55	97,37
M_4 - *V0213*	GES	0,62%	1,21%	0,00%	0,39%	1,93%
	PCALIBRA	0,58%	1,22%	0,00%	0,38%	2,01%
M_4 - *V0219*	GES	6,25%	7,24%	1,16%	5,82%	32,61%
	PCALIBRA	6,36%	7,22%	1,15%	5,79%	30,17%
M_4 - *V0221*	GES	1,73%	1,95%	0,22%	1,01%	4,94%
	PCALIBRA	1,79%	1,95%	0,23%	1,01%	4,93%
Observações na amostra		346	431	778	509	101

Anexo 1: O Programa PCALIBRA

Conforme descrito na seção 4, a nova metodologia proposta neste trabalho foi implementada utilizando a linguagem de programação *Delphi da Borland*. O programa desenvolvido recebeu o nome de PCALIBRA e é composto por três módulos de operação descritos a seguir. Uma cópia gratuita deste programa pode ser solicitada enviando um *e-mail* para o autor principal (britom@ibge.gov.br).

Módulo 1 – Parâmetros

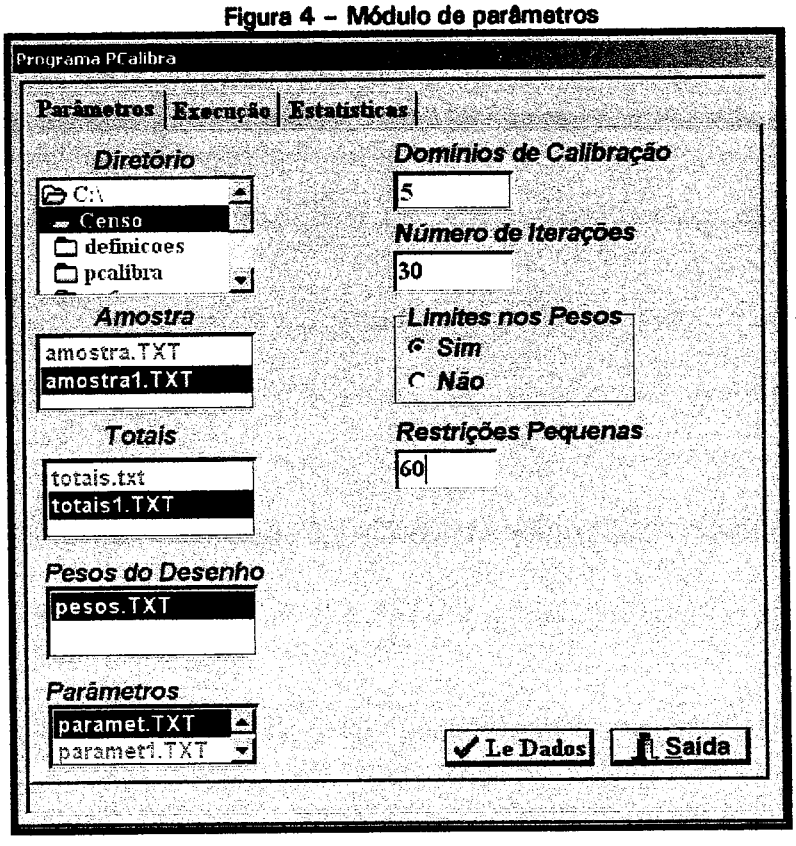
No primeiro módulo, janela representada pela Figura 4, tem-se a seleção dos arquivos de dados e a definição de um conjunto de parâmetros que serão considerados durante o processo de calibração.

Neste módulo é selecionado o diretório onde estão os arquivos com as observações da amostra, dos totais populacionais, dos pesos do desenho (utilizados para cada uma das unidades amostrais) e o arquivo de parâmetros. Mais adiante, será apresentada uma descrição de cada um destes arquivos.

À direita desta tela define-se o número de domínios de calibração, o número de iterações do algoritmo e se serão considerados limites para os pesos w.

Na caixa "restrições pequenas" atribui-se um valor percentual V entre 0 e 100. Este valor indica que as variáveis auxiliares cujo percentual associado ao número de observações não nulas na amostra for menor ou igual a V não serão consideradas no processo de calibração, ou seja, pode-se previamente descartar algumas restrições. O *default* considerado é zero.

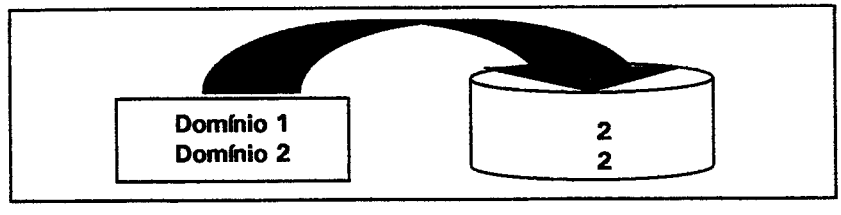
Figura 4 – Módulo de parâmetros



Arquivos do Módulo de Parâmetros

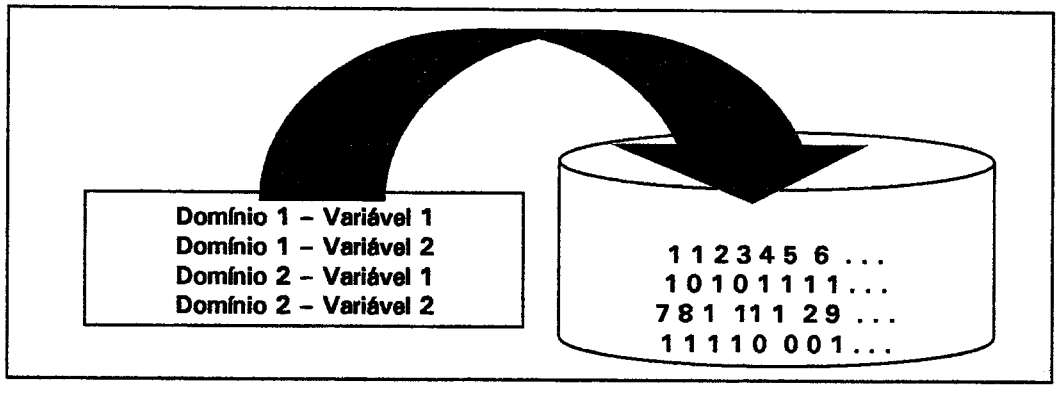
Arquivo de Parâmetros: Um arquivo do tipo texto cujo o número de linhas corresponde ao número de domínios de calibração. Neste arquivo, em cada linha, tem-se o número de variáveis auxiliares que serão consideradas em cada domínio de calibração. Na figura 5 há um exemplo deste arquivo.

Figura 5 - Arquivo de parâmetros

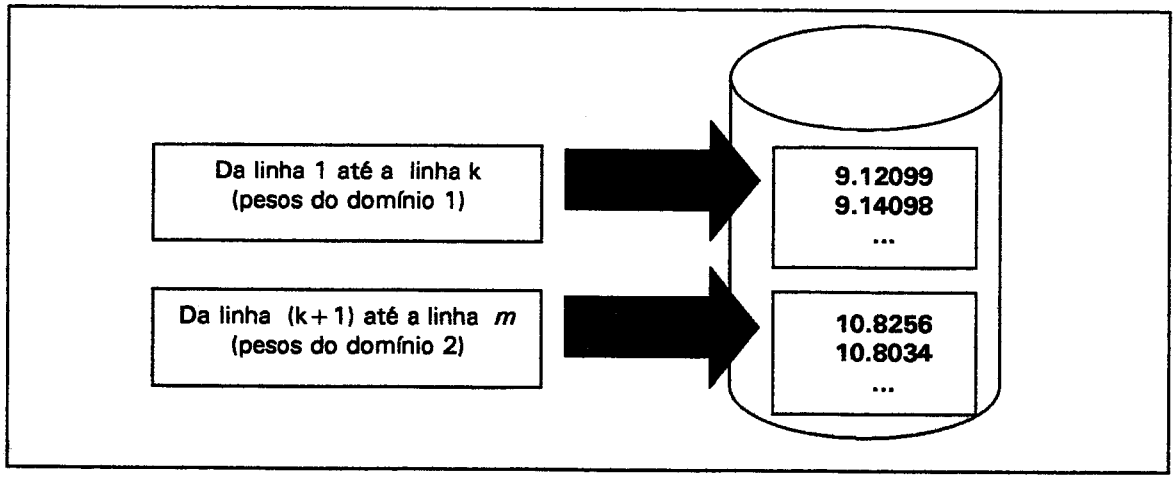


Arquivo da Amostra: Um arquivo do tipo texto que contém para cada um dos domínios de calibração as observações (separadas por um espaço) das variáveis explicativas X que serão utilizadas no processo de ajuste dos pesos. Na Figura 6, há um exemplo deste arquivo, considerando 2 domínios de calibração e 2 variáveis por domínio. Neste arquivo, o número de observações (colunas) em cada linha pode variar de acordo com o domínio.

Figura 6 – Arquivo da Amostra

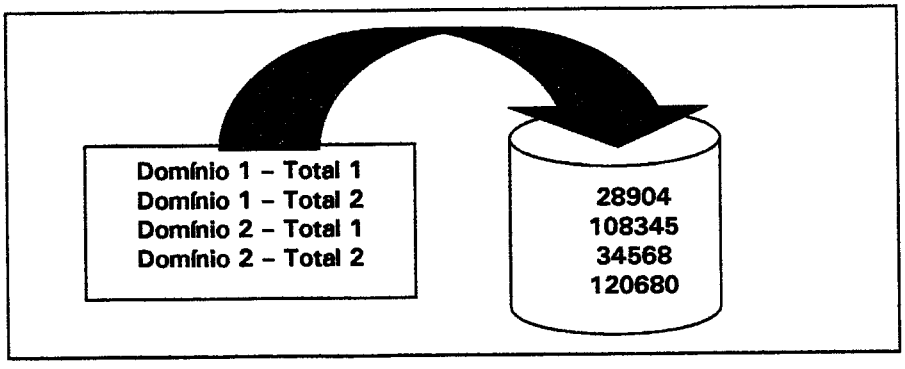


Arquivo de Pesos: Um arquivo do tipo texto que contém os pesos do desenho, associados a cada um dos domínios e a cada uma das unidades amostrais que serão ponderadas (arquivo amostra). Este arquivo tem um total de m linhas, que correspondem ao número de unidades amostrais (observações) associadas a cada um dos domínios definidos no arquivo da amostra, considerando a estrutura exemplificada (2 domínios) abaixo.



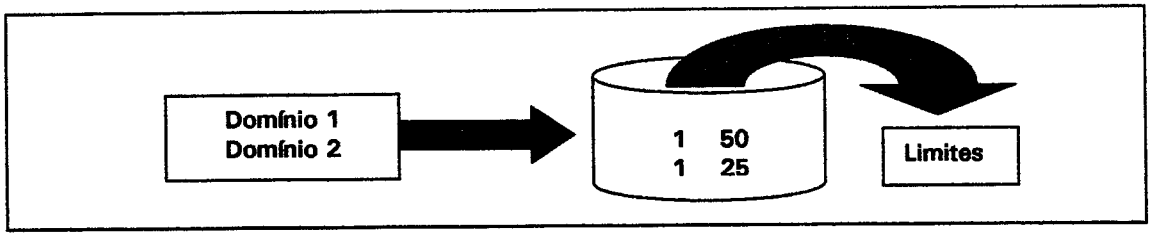
Arquivos de Totais : Um arquivo do tipo texto que contém os totais populacionais associados as variáveis explicativas X (da amostra) que serão utilizadas para calibração de cada um dos domínios. Na Figura 8, há um exemplo deste arquivo, considerando novamente 2 domínios e 2 totais por domínio.

Figura 8 – Arquivo de Totais



Arquivo de Limites: Um arquivo do tipo texto que contém em cada linha (associada a um domínio de calibração) limites inferior e superior (separados por um espaço). Na Figura 9 há um exemplo de um arquivo de limites para três domínios de calibração.

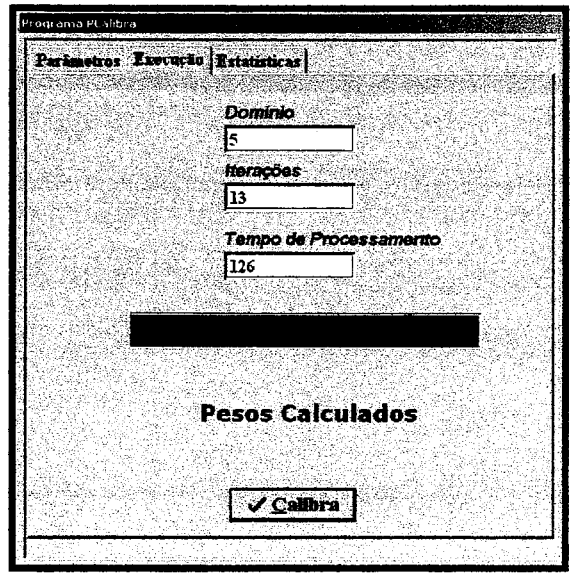
Figura 9 – Arquivo de Limites



Módulo 2 - Execução

Uma vez definidos os arquivos e os parâmetros de entrada, seleciona-se a janela execução (módulo 2) que contém as informações do número de domínios que foram processados (calibrados), a iteração atual do algoritmo e o tempo de processamento em segundos.

Figura 10 – Módulo de execução



Módulo 3 - Estatísticas

Neste módulo apresenta-se um conjunto de informações derivadas do processo de calibração:

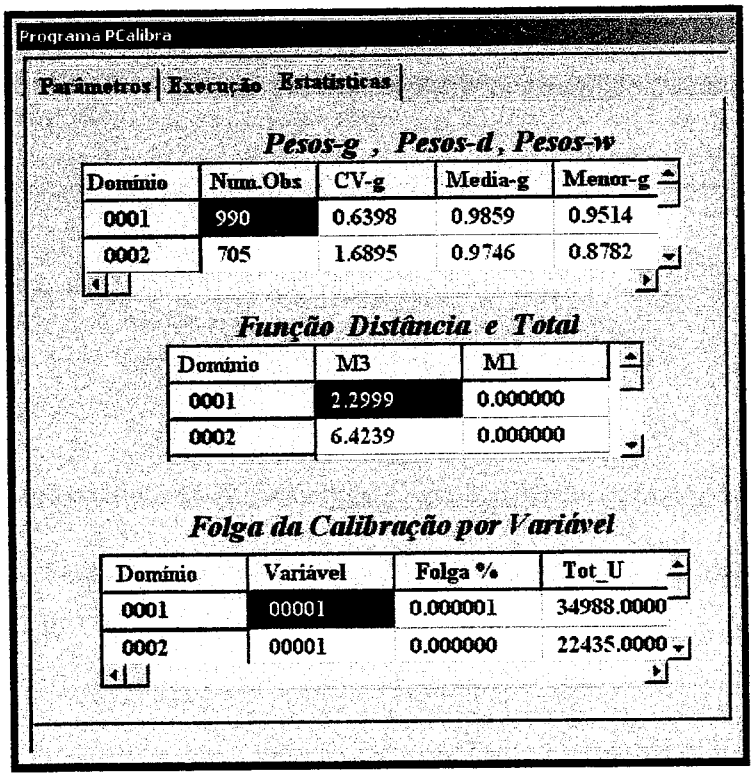
(1) Número de unidades amostrais em cada um dos domínios de calibração;

(2) Coeficiente de variação, média e valores máximo e mínimo para os pesos d do desenho, para os ajustes g e para os pesos finais w;

(3) Medidas M1 e M3 (descritas na seção 4); e

(4) Folgas em valores percentuais entre os totais do universo (conhecidos) e os totais estimados a partir dos pesos w.

Figura 11- Módulo de estatísticas



Referências bibliográficas

BANKIER, M. (1995). Two Step – Generalized Least Squares Estimation. Social Survey Methods. Division Statistics Canadá.

BAZARAA, M.S., SHERALI H.D e SHETTY C.M. (1993). *NonLinear Programming – Theory and Algorithms*. John Wiley & Sons Publishers, New York.

COCHRAN, WILLIAN G. (1977). Sampling Techniques. Third Edition, Wiley.

Censo Demográfico 2000. Primeiros Resultados da Amostra . Parte 1/ IBGE.

DEVILLE, J. C., and SÄRNDAL, C.E (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ESTEVAO, V., HIDIROGLOU M.A. and SÄRNDAL C.E. (1995). "Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics,11, 181-204.

Metodologia do Censo Demográfico 2000 /IBGE. Rio de Janeiro: IBGE, 2003. 574p. (Relatórios metodológicos, ISSN 0101-2843; n.25)

PESSOA, DJALMA GALVÃO CARNEIRO, SILVA, PEDRO LUIS DO NASCIMENTO e PFEFFERMANN, DANNY (2000). Calibration Methods for the Brazilian Census. Relatório IBGE/DPE/DEMET.

PESSOA, DJALMA GALVÃO CARNEIRO (2005). Estimadores de Calibração em Pesquisas Amostrais. Relatório Técnico. Coordenação de Métodos e Qualidade (COMEQ) / DPE.

RAO, J.N.K and SINGH, A.C. (1997). Range Restricted Weight Calibration for Survey Data. Unpublished manuscript.

SILVA, PEDRO LUIS NASCIMENTO (2004). Calibration Estimation: When and Why, How Much and How. Textos para Discussão n. 15, DPE/IBGE.

SANTOS, TURÍBIO J. G. (1998). Um Novo algoritmo de Penalização Hiperbólica para a Resolução do Problema de Programação Não-Linear com Restrições de Igualdade. Tese de Doutorado COPPE/UFRJ.

Agradecimentos

Os autores agradecem a Antonio José Ribeiro Dias, da Coordenação de Métodos e Qualidade do Instituto Brasileiro de Geografia e Estatística - IBGE, pelas idéias e sugestões dadas no trabalho.

Abstract

This paper reports a new methodology to solve the problem of calibration of sample expansion weights (Silva, 2004). This methodology is based on a mathematical formulation that take in consideration all at once a quadratic distance function and two sets of linear restrictions, namely, equality and inequality restrictions. To solve this formulation and, thus, obtaining the weights to be used in the expansion of the sample (Cochran, 1977), we have applied a method of non-linear optimization due to Santos (1998). Aiming to show the potentialities and applicability of the new methodology, we have considered the real problem of adjusting the weights in the Brazilian demographic census of the year 2000, taking in account both the sample and the universe data. Main computational results as well a user friendly interface software developed for easy operation are presented.

Keywords: Calibration, sample, penalization.

Fatores que influenciaram na aprovação de candidatos de diferentes grupos socioeconômicos no Vestibular 2004 da UFMG[1]

Ludmila V. Lage[*]
Rosangela H. Loschi[*]
Glaura C. Franco[*]
Mauro M. Braga[**]

[1] Os dados utilizados neste trabalho foram cedidos pela COPEVE-UFMG. Agradecemos aos Professores Antônio Zumpano P. Santos e Maria do Carmo L. Peixoto por seus comentários e sugestões em versões preliminares deste trabalho e a dois pareceristas anônimos por suas revisões minusciosas que contribuíram significativamente para a melhora deste trabalho. Esta pesquisa foi parcialmente financiada pelo CNPq (projetos no 300325/2003-7, 472066/2004-8) e COPEVE-UFMG.

[*] Endereços para correspondências: Departamento de Estatística, Instituto de Ciências Exatas – UFMG.
[**] Departamento de Química e Pró-Reitoria de Graduação – UFMG.

Resumo

Este trabalho visa a identificar grupos de candidatos que sejam homogêneos segundo suas características socioeconômicas e identificar, em cada grupo, quais as variáveis referentes às características pessoais e à formação acadêmica que mais influenciaram na aprovação do candidato no Vestibular-2004 da UFMG. Consideramos todas as variáveis definidas no questionário socioeconômico e cultural aplicado no ato da inscrição do vestibular. As metodologias utilizadas no desenvolvimento deste trabalho foram: Regressão logística, classificação por árvores de regressão (CART) - utilizada para separar os candidatos em grupos homogêneos do ponto de vista socioeconômico - e a curva ROC para avaliar o desempenho da classificação obtida. Resumidamente, conclui-se que há quatro grupos homogêneos do ponto de vista socioeconômico. Nos grupos menos favorecidos há uma predominância de candidatos que estudaram em escolas públicas, principalmente nas estaduais, e cuja renda familiar é inferior a 10 salários mínimos. Os grupos mais favorecidos são compostos por candidatos cuja maioria estuda em escolas particulares e possui renda superior a 20 salários mínimos. Percebe-se que a aprovação dos candidatos dos dois grupos menos favorecidos do ponto de vista socioeconômico está mais fortemente associada ao fato do candidato ter feito algum vestibular anteriormente (Grupo 1) e ter cursado o ensino médio no turno diurno (Grupo 2). Já para os candidatos dos dois grupos mais favorecidos do ponto de vista socioeconômico, verifica-se que a aprovação está mais fortemente associada ao fato de o candidato ter cursado o ensino médio em escola particular ou pública federal (Grupo 3) e ter feito ensino médio não-profissionalizante (Grupo 4).

Palavras-chave: CART, curva ROC, raça, regressão logística, tipo de escola de ensino médio.

1. Introdução

É cada vez mais atual o tema de democratização do acesso ao ensino público superior no Brasil. O anteprojeto de lei do governo federal que versa sobre a Reforma Universitária propõe a criação de cotas, para as universidades públicas como um mecanismo para a maior inclusão social, as quais devem ser implementadas nos próximos 10 anos (Informativo MEC, novembro de 2004). Segundo o Informativo MEC de julho de 2005 *"A Reforma Universitária deve ajudar a reduzir as desigualdades sociais no País. A proposta do governo federal é reservar 50% das vagas em cada curso – inclusive nos mais concorridos – e em todos os turnos, nas instituições federais de ensino superior, para estudantes egressos de escolas públicas reservado um percentual*

de acordo com a proporção de grupos étnicos historicamente prejudicados nos respectivos estados, segundo o Instituto Brasileiro de Geografia e Estatística."

A proposta de criação de cotas é polêmica e tem gerado discussões na sociedade. Algumas universidades públicas criaram alternativas para enfrentar o problema de maior inclusão social. A Universidade de São Paulo criou curso preparatório gratuito destinado aos estudantes de escolas públicas e abriu uma nova unidade na Zona Leste da cidade de São Paulo. A Universidade Estadual de Campinas concedeu pontos adicionais aos estudantes provenientes de escolas públicas e aos que se declaram indígenas, pretos ou pardos. A Universidade Federal de Pernambuco adotou, no caso das licenciaturas da área de exatas, um processo de seleção que envolve o oferecimento de um curso preparatório gratuito para a segunda etapa do concurso aos candidatos aprovados na primeira etapa. A Universidade Federal do Paraná e Universidade de Brasília criaram cotas para negros, seguindo o exemplo de algumas universidades estaduais do Rio de Janeiro e da Bahia. A Universidade Federal de Minas Gerais - UFMG definiu a expansão no turno da noite como mecanismo prioritário para a inclusão social.

Visando conhecer o perfil do candidato no seu vestibular, a UFMG aplica um questionário socioeconômico com perguntas referentes às condições econômicas, aos antecedentes escolares, raça, etc. do candidato (seção 2 para uma descrição mais detalhada das variáveis). Vários estudos vêm sendo elaborados com base em tais dados. Lopes *et al.* (2005) estudam as características de candidatos provenientes de escolas pública e privada que tentaram ingressar na UFMG, em 2004, e concluem que, embora o conhecimento de língua estrangeira e local de moradia não sejam os únicos fatores associados às altas chances de aprovação, estes são os fatores mais fortemente associados à aprovação de candidatos provenientes de escolas pública e particular, respectivamente. Os autores mostram que candidatos vindos de escolas públicas que lêem inglês, francês ou duas ou mais línguas estrangeiras tendem a ser mais favorecidos, e candidatos de escolas privadas que moram em Belo Horizonte, em geral, têm chances mais altas de aprovação. Percebem, ainda, para candidatos provenientes de escolas públicas, que nos grupos onde há maior chance de aprovação há, em geral, um predomínio de candidatos que cursaram escolas federais. Em outro

estudo sobre o assunto, Dias *et al.* (2005) concluem que o conhecimento de língua estrangeira também está mais fortemente associado à aprovação do candidato, quando se considera a totalidade dos inscritos e dos candidatos aos cursos diurno e noturno. Os autores mostram que, em geral, o conhecimento de inglês, francês ou duas ou mais línguas estrangeiras favorecem o ingresso do candidato na universidade. Para os cursos noturnos, Dias *et al.* (2005) notam que as variáveis que aparecem mais freqüentemente para explicar a aprovação do candidato são variáveis de cunho socioeconômico e, para os cursos diurnos, mais freqüentemente aparecem as variáveis relativas à formação do candidato. Dias *et al.* (2005), também, mostram que o "Fator Socioeconômico" (FSE)[2] médio tende a ser maior para os candidatos aprovados. Os candidatos aos cursos noturnos (inscritos e aprovados) apresentam "FSE" médio menor que os candidatos aos cursos diurnos e, além disso, verificam que os candidatos da raça preta são os que apresentam "FSE" médio mais baixo e os da raça branca são os que apresentam "FSE" médio mais alto, seja para a totalidade dos candidatos inscritos, seja para os candidatos aprovados. Os autores ressaltam ainda que as discrepâncias entre as raças começam a surgir quando analisadas as questões socioeconômicas, quando percebem, por exemplo, que pretos e indígenas são aqueles que, por se encontrarem nas faixas menos favorecidas economicamente, tendem a demorar mais a ingressar na universidade.

Este trabalho pretende identificar grupos de candidatos que sejam homogêneos segundo suas características socioeconômicas e, em cada um destes grupos, identificar as variáveis referentes à formação acadêmica do candidato e às suas características pessoais que mais influenciaram na aprovação no Vestibular 2004 da UFMG. Duas ferramentas estatísticas serão utilizadas: o CART (*Classification and Regression Trees*) para a identificação dos grupos de candidatos socioeconomicamente homogêneos e a Regressão Logística para a identificação dos fatores associados à aprovação em cada grupo. A capacidade de classificação dos modelos construídos é avaliada usando a curva ROC (*Receiver Operating Characteristic*). Neste trabalho utilizamos o *software* estatístico SPSS e seu anexo *Answer Tree*.

---

[2] O FSE é um índice que combina aspectos da trajetória escolar do candidato, com o padrão de renda familiar e profissão e escolaridade dos pais. Ver detalhes em Braga *et al.* (2001) e em www.ufmg.br/inclusaosocial/ cursos_noturnos.doc.

Este trabalho está organizado da seguinte forma: na seção 2, descreve-se as variáveis utilizadas no trabalho, e apresenta-se brevemente as metodologias utilizadas; na seção 3, o CART é utilizado para identificar grupos de candidatos mais homogêneos segundo as características socioeconômicas; na seção 4, considera-se os grupos obtidos e identifica-se, em cada grupo, os fatores referentes à formação acadêmica e às características pessoais do candidato que influenciaram na sua aprovação no Vestibular-2004 da UFMG; e a seção 5 finaliza o artigo apresentando as conclusões mais relevantes.

2 . Variáveis e metodologias estatísticas utilizadas

As variáveis utilizadas neste trabalho (Quadros 1 e 2) abordam assuntos relacionados à situação socioeconômica do candidato e questões relativas às características pessoais e à formação acadêmica do candidato, todas oriundas do questionário socioeconômico aplicado pela UFMG no Vestibular de 2004.

A variável resposta é "O candidato foi aprovado?", a qual é uma variável indicadora do candidato ter sido aprovado (1) ou reprovado (0) no Vestibular 2004.

Para a análise, utilizando a Regressão Logística, as variáveis foram dicotomizadas, por simplicidade. Esta dicotomização foi feita seguindo trabalhos anteriores em que o CART foi utilizado para construir o modelo (Dias *et al.*, 2005).

Quadro 1 - Variáveis relativas à situação socioeconômica do candidato, Vestibular 2004, UFMG6

Variável
Ocupação do pai
Ocupação da mãe
Escolaridade do pai
Escolaridade da mãe
Número de pessoas que vivem da renda mensal
Situação familiar
Trabalha em atividade remunerada?
Por quantos anos teve atividade remunerada?
Participa economicamente na família
Renda
FSE
ABIPEME

Estas variáveis foram consideradas no estudo da mesma forma que estão definidas no questionário socioeconômico aplicado pela UFMG no Vestibular de 2004. Suas descrições são longas por isto não a faremos aqui. Este questionário pode ser obtido junto à COPEVE/UFMG (URL: http://www.ufmg.br/copeve/). Dois índices são utilizados para medir as características socioeconômicas do candidato: o FSE, já citado anteriormente, e o critério ABIPEME desenvolvido pela Associação Brasileira de Institutos de Pesquisa de Mercado (www.ufrn.br/sites/ fonaprace/perlanexo3.doc). A variável "Renda" refere-se à renda mensal do grupo familiar medida em número de salários mínimos. Esta variável inclui os rendimentos líquidos de todos os membros ativos do grupo familiar deduzidos todos os encargos.

Quadro 2 - Variáveis relativas às características pessoais e a antecedentes acadêmicos do candidato, Vestibular 2004, UFMG

Variável	Descrição	Valor
Sexo	Feminino	0
	Masculino	1
Estado civil	Solteiro	0
	Casado ou outros	1
Onde mora	BH	0
	Fora de BH	1
Raça_*Dummy*1	Branca	0
	Amarela, indígena ou não deseja declarar	0
	Preta ou parda	1
Raça_*Dummy*2	Branca	0
	Amarela, indígena ou não deseja declarar	1
	Preta ou parda	0
Sabe redigir um texto no computador?	Não	0
	Sim	1
Onde concluiu o ensino médio	BH	0
	Fora de BH	1
Curso do ensino médio	Não profissionalizante	0
	Profissionalizante, supletivo ou outro equivalente	1
Tipo de escola do ensino médio	Pública federal ou particular	0
	Pública municipal ou estadual ou curso livre	1
Tempo de conclusão do ensino médio	3 anos	0
	4 anos ou mais	1
Turno ensino médio	Diurno	0
	Noturno	1
Você cursou ensino médio	Integralmente em escola pública	0
	Parte pública, parte particular ou integralmente na particular	1
Já prestou vestibular?	Não ou sim, como treinante	0
	Sim	1
Há quanto tempo tenta ingressar no ensino superior	1 ano ou menos	0
	Mais de 1 ano	1
Freqüentou cursinho pré-vestibular?	Não	0
	Sim	1
Turno do curso na UFMG	Diurno	0
	Noturno	1
Conhecimento de língua estrangeira	Lê inglês ou 2 ou mais línguas estrangeiras	0
	Não Lê, ou lê espanhol, francês ou outra língua estrangeira	1
Já freqüentou curso de idiomas?	Sim, de inglês	0
	Não ou sim, de uma outra língua	1
Opção_*Dummy*1 (de língua estrangeira no vestibular)	Inglês	0
	Espanhol	0
	Francês	1
Opção_*Dummy*2 (de língua estrangeira no vestibular)	Inglês	1
	Espanhol	0
	Francês	0

Para o tratamento da variável "Raça" optou-se por construir duas variáveis indicadoras, denotadas aqui por "Raça_*Dummy*1" e "Raça_*Dummy*2", separando os candidatos em 3 grupos, a saber, um grupo formado por candidatos que se declararam da raça branca; outro grupo formado por candidatos das raças preta e parda; e um terceiro grupo formado por aqueles candidatos das raças indígena e amarela e os que não desejam declarar a raça. Esta divisão leva em consideração dois fatos. O primeiro é que as cotas para universidades federais são definidas para candidatos das raças preta e parda (Informativos MEC, de novembro de 2004 e julho de 2005) e, por isto, tais candidatos foram considerados em um mesmo grupo. O segundo fato é que a maior parte dos candidatos inscritos no vestibular da UFMG são declaradamente da raça branca e, por isto, estes foram considerados em um grupo separado.

Deve-se ressaltar que o número de candidatos inscritos no Vestibular 2004 foi 60 616, um número relativamente alto. Problemas de dados esparsos, portanto, não ocorreram neste caso, apesar do grande número de variáveis explicativas.

2.1. O CART

Para a identificação dos agrupamentos de candidatos, que sejam homogêneos do ponto de vista socioeconômico, foi utilizado o CART (*Classification and Regression Trees*). O CART pode ser considerado como um modelo de regressão não-paramétrico que tem por objetivo estabelecer uma relação entre um conjunto de variáveis explicativas e uma única variável resposta, que neste caso é "O candidato foi aprovado?". O modelo é ajustado mediante sucessivas divisões binárias no conjunto de dados, de modo a tornar os grupos resultantes cada vez mais homogêneos em relação à variável resposta, a qual pode ser contínua ou categorizada. O uso do CART não pressupõe que a suposição de normalidade seja verificada para os dados. Uma vantagem do CART é que as interações entre as variáveis explicativas são automaticamente captadas pelo modelo. Para maiores detalhes ver, por exemplo, Bell (1996) e Diniz e Louzada-Neto (2000).

A construção de uma árvore, via CART, consiste em determinar, a partir do vetor de variáveis explicativas, aquelas que melhor dividem o conjunto de dados. Para esta escolha, todas as combinações possíveis entre as variáveis são testadas, sendo escolhida aquela que mais reduzir a heterogeneidade dos subconjuntos criados. Essa divisão resulta em dois subconjuntos mais homogêneos que o conjunto original. O ideal é que todos os elementos dos subconjuntos resultantes possuam o mesmo valor para a variável resposta, o que implicaria em grupos (ou nós) completamente homogêneos.

Quando a variável resposta é uma variável dicotômica, que é o caso estudado neste trabalho, para se medir a homogeneidade ou, equivalentemente, a impureza de determinado nó, é utilizado o índice de *Gini*, denotado aqui por $i(t)$, que é definido por:

$$i(t) = 2p(1|t)(1-p(1|t)),$$

em que t é um nó arbitrário e $p(1|t)$ é a probabilidade da variável resposta ser classificada na categoria 1 para o nó t. Calculado o índice de *Gini*, escolhe-se para fazer parte do modelo aquelas variáveis com maior grau de associação, isto é, maior valor para $i(t)$. Quando uma das probabilidades se aproxima de 1, o valor da função $i(t)$ tende a 0, significando a maior homogeneidade. Na prática, as probabilidades são substituídas pela proporção de indivíduos de determinada categoria naquele nó considerado.

O processo de divisão é repetido recursivamente até que alguma das regras de parada seja alcançada. Estas regras de parada são definidas pelo usuário e, em geral, consistem em uma limitação no tamanho da impureza do nó, do número de gerações e do número de nós da árvore.

Para variáveis resposta dicotômicas, além das variáveis explicativas associadas à variável resposta, também se pode obter a probabilidade condicional da variável resposta resultar em sucesso. Conseqüentemente, para o caso em estudo, uma análise via CART fornece a probabilidade de aprovação dos candidatos que possuem as características apresentadas em cada grupo resultante da divisão, além das variáveis

que melhor explicam a aprovação do candidato.

Além do CART, existem outras possibilidades para a construção de árvores de decisão, como o CHAID, o CHAID EXAUSTIVO e o QUEST. Nestes casos, a medida de associação em cada em nó se dá através da estatística Qui-Quadrado.

2.2. Regressão logística

Uma vez identificados os grupos homogêneos do ponto de vista socioeconômico, a Regressão Logística será utilizada para identificar os fatores associados com a aprovação do candidato em cada um destes grupos. A Regressão Logística é uma técnica usada para análise de dados com resposta binária e estabelece uma relação entre a probabilidade de ocorrência de cada um dos resultados da variável resposta associados às variáveis explicativas. Esta técnica pode ser usada de forma descritiva ou preditiva. Na prática, é bastante empregada nas áreas de Saúde, Bioestatística, Epidemiologia, Econometria, entre outras. Brevemente, pode-se descrever o modelo logístico como segue.

Seja Y a variável resposta dicotômica em que $Y=1$ denota um sucesso e $Y=0$ denota um fracasso. Considere variáveis explicativas x as quais podem ser tanto contínuas quanto categorizadas.

A probabilidade de ocorrência de cada um dos eventos definidos sobre Y é dada por:

$$\pi(x)=\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}},$$

onde β_0 é o intercepto e β_i é o coeficiente associado à variável explicativa x_i, $i=1,\ldots p$.

A estimação dos parâmetros do modelo logístico pode ser feita através do método de máxima verossimilhança e, neste caso, é utilizado o algoritmo de Newton Raphson.

A interpretação dos coeficientes se dá através da razão das chances (*Odds Ratio*).

Esta mede o quanto é mais provável que a variável resposta assuma um valor positivo ($Y=1$) quando à variável explicativa é acrescida uma unidade. Matematicamente, a razão das chances para determinada variável explicativa, x_i, é expressa da seguinte forma:

$$\frac{\pi(x_i+1)/[1-\pi(x_i+1)]}{\pi(x_i)/[1-\pi(x_i)]}=e^{\beta_i}.$$

Para a seleção das variáveis que permanecerão no modelo logístico, será utilizado o teste da razão de verossimilhanças e o ajuste do modelo será avaliado através do teste de Hosmer-Lemeshow. Para maiores detalhes ver Hosmer and Lemeshow (1989).

2.3. A curva ROC

A curva ROC (*Receiver Operating Characteristic*) é uma ferramenta útil para avaliar a performance de modelos de classificação de objetos que podem ser classificados em duas categorias.

Para uso da curva ROC é necessário que as variáveis provindas dos modelos de classificação (variáveis teste) sejam quantitativas. Elas são freqüentemente probabilidades resultantes de análises discriminantes, regressões logísticas ou escores que indicam a "força de convicção" que um objeto pertença a uma categoria ou outra. A variável de estado pode ser de qualquer tipo e indica a categoria verdadeira à qual o objeto pertence. É assumido que, à medida que aumenta o valor da variável teste, aumenta a crença de que um objeto pertença a uma categoria, enquanto uma diminuição no valor da variável teste aumenta a crença de que o objeto pertença à outra categoria. Também é assumido que a verdadeira categoria, à qual o objeto pertence, é conhecida.

Quando os modelos de classificação produzem uma resposta sob a forma de uma variável contínua, emprega-se uma regra de decisão, baseada na busca de um ponto de corte que resume tal quantidade em uma resposta dicotômica, de forma que um objeto com mensurações menores ou iguais ao ponto de corte é classificado como pertencente

a uma categoria e um objeto com uma resposta ao teste maior que o ponto de corte é classificado como pertencente à outra categoria.

A Curva ROC é, então, uma curva contínua obtida a partir das probabilidades de aprovação de cada candidato, fornecidas pelo modelo da seguinte forma: para todo ponto de corte no intervalo (0,1), determina-se a sensibilidade, Se, e a especificidade, Es, do modelo, que no caso apresentado aqui são dadas por:

$$Se = P(\textit{classificar como aprovado} \mid \textit{realmente foi aprovado}),$$

$$Es = P(\textit{classificar como não aprovado} \mid \textit{realmente não foi aprovado}).$$

A curva ROC é formada pelos pontos $(Se,\ 1-Es)$ obtidos para cada ponto de corte.

Uma maneira global conveniente de quantificar a precisão de um modelo de classificação é expressar sua performance por um único número. A medida mais comum é a área sob a curva ROC. Por convenção, essa área é sempre maior ou igual a 0,5. Os valores variam entre 1 (perfeita separação dos valores resultantes do modelo de classificação das duas categorias) e 0,5 (sem diferença aparente entre as distribuições dos valores resultantes dos modelos de classificação das duas categorias). A área é uma expressão quantitativa e descritiva da proximidade entre a curva ROC associada ao modelo de interesse e uma curva ROC ótima (área = 1). Uma área de 0,9, por exemplo, indica que um objeto selecionado aleatoriamente de uma categoria tem um valor resultante do modelo de classificação maior do que para um objeto escolhido aleatoriamente da outra categoria, em 90% das vezes. Usando métodos não-paramétricos pode ser realizado um teste de hipóteses do tipo

$$\begin{cases} H_0 : \textit{Área} = 0{,}5 \\ H_a : \textit{Área} \neq 0{,}5 \end{cases}$$

e, assim, saber se tal área é significativamente diferente de 0,5. Um modelo ideal é aquele que apresenta tanto a especificidade quanto a sensibilidade iguais a 1 e, conseqüentemente, cuja área sob a curva ROC seja igual a 1. Um modelo é considerado como tendo muito boa capacidade de classificação se a área sob a curva ROC é superior a 0,8. Para maiores detalhes sobre a Curva ROC ver Hanley and McNeil (1982), Zweig and Campbell (1993) e Martinez *et al.* (2003).

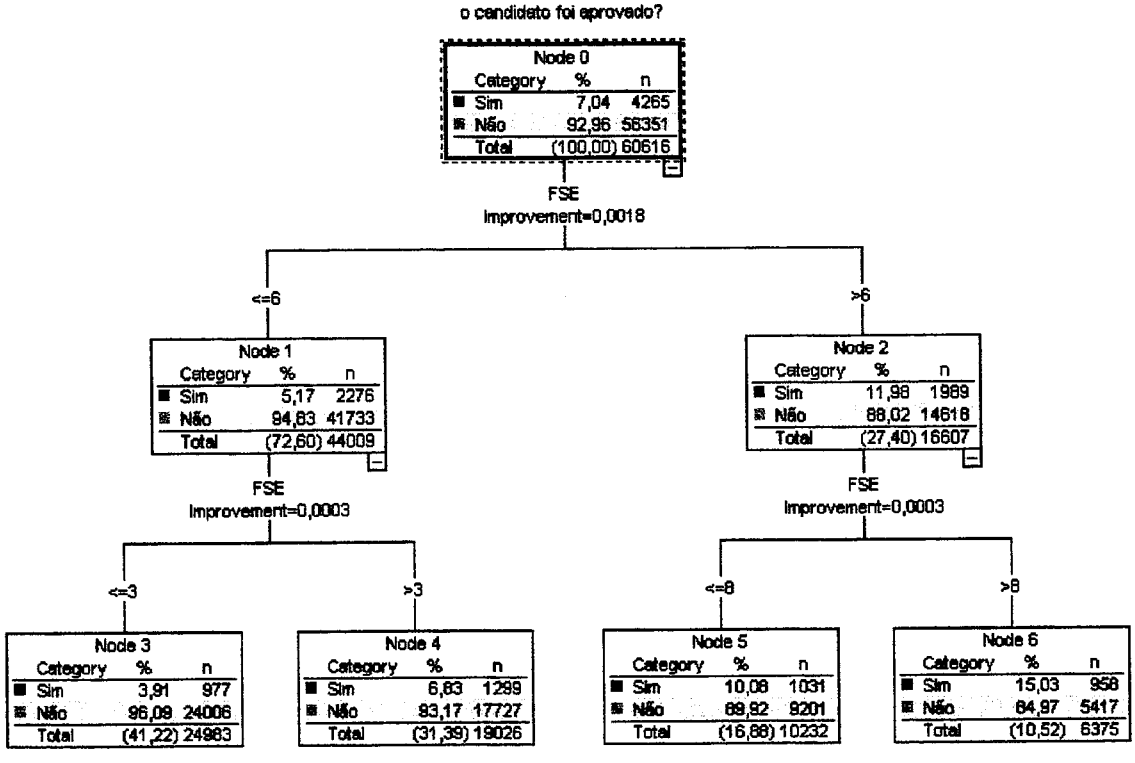
3. Construindo agrupamentos homogêneos do ponto de vista socioeconômico via CART

Nesta seção, utiliza-se o CART para encontrar grupos de candidatos, ao Vestibular 2004 da UFMG, que sejam o mais homogêneos possíveis segundo suas características socioeconômicas que são mensuradas através das variáveis apresentadas no Quadro 1. Para complementar o estudo, também se faz uma análise descritiva dos dados relativos a cada grupo, com o intuito de verificar como eles se distribuem segundo as seguintes variáveis: Tipo de escola do ensino médio, Renda e Raça. Estas são variáveis importantes para o estudo, uma vez que trazem informações relevantes sobre temas importantes no que tange à discussão sobre a democratização do acesso ao ensino público superior e à questão da definição de cotas para as universidades públicas federais.

Da Figura 1, pode-se dizer que há quatro grupos homogêneos do ponto de vista socioeconômico, identificados nos nós terminais 3, 4, 5 e 6. Nota-se que dos 60 616 candidatos ao Vestibular 2004 da UFMG, considerados para a análise, 41,2% deles possuem FSE de 0 a 3 (Grupo 1); 31,4% possuem FSEs de 4 a 6 (Grupo 2); 16,9% possuem FSE igual a 7 ou 8 (Grupo 3); e 10,5% dos candidatos possuem FSE igual a 9 ou 10 (Grupo 4). O Grupo 1 corresponde ao grupo socioeconomicamente menos favorecido. Note que é o grupo que tem maior número de inscritos no vestibular e a menor proporção de aprovados (dos 24 983 candidatos deste grupo inscritos no vestibular, apenas 3,9% foram aprovados). Contrariamente, observa-se que a proporção de candidatos do grupo socioeconomicamente mais favorecido (Grupo 4) que foram

aprovados no vestibular é de 15,0%, apesar de ser o grupo com menor proporção de inscritos no Vestibular-2004 da UFMG. É relevante ressaltar ainda que os candidatos aprovados se distribuem da seguinte forma entre os grupos: 22,9% pertencem ao Grupo 1; 30,5% pertencem ao Grupo 2; 24,2% pertencem ao Grupo 3; e, finalmente, 22,5% pertencem ao Grupo 4.

Figura 1 - Agrupamento dos candidatos segundo os fatores socioeconômicos, Vestibular 2004 da UFMG



Para um estudo mais detalhado destes grupos, seguem nas Tabelas 3, 4 e 5 as distribuições dos candidatos inscrito e aprovado de cada grupo segundo a Renda, Tipo de escola de ensino médio e Raça, respectivamente.

Tabela 3 - Percentual dos candidatos inscrito e aprovado em cada grupo segundo classes renda, Vestibular 2004, UFMG

Classe de renda em salários mínimos	Grupos e situação do candidato							
	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
	Inscritos	Aprovados	Inscritos	Aprovados	Inscritos	Aprovados	Inscritos	Aprovados
Menos de 1	2,1	1,0	1,1	0,4	0,6	0,7	0,0	0,0
De 1 a 2	22,8	14,7	4,8	3,1	0,7	0,6	0,0	0,0
De 2 a 5	52,4	54,2	31,7	27,9	8,1	6,5	0,0	0,0
De 5 a 10	19,9	25,7	41,7	44,1	25,8	24,1	0,0	0,0
De 10 a 15	2,1	3,1	13,7	14,4	31,2	30,7	20,3	17,1
De 15 a 20	0,6	1,1	5,4	8,0	20,1	21,8	16,9	18,6
De 20 a 40	0,1	0,1	1,4	2,0	10,0	11,7	42,1	46,2
De 40 a 60	0,0	0,0	0,2	0,1	2,2	2,5	13,9	13,5
Acima de 60	0,0	0,0	0,1	0,1	1,3	1,4	6,7	4,6
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabela 4 - Percentual dos candidatos inscrito e aprovado em cada grupo segundo o tipo de escola do ensino médio, Vestibular 2004, UFMG

Tipo de escola de ensino médio	Grupos e situação do candidato							
	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
	Inscritos	Aprovados	Inscritos	Aprovados	Inscritos	Aprovados	Inscritos	Aprovados
Público municipal	14,7	15,8	5,8	4,8	1,3	1,4	1,3	1,4
Público estadual	71,1	52,3	31,6	22,6	6,1	3,9	6,1	3,9
Público federal	5,9	22,3	5,0	14,8	3,2	6,5	3,2	6,5
Particular	7,7	9,4	57,2	57,5	89,2	88,3	89,2	88,3
Curso livre	0,6	0,2	0,3	0,3	0,1	0,0	0,1	0,0
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabela 5 - Percentual dos candidatos inscrito e aprovado em cada grupo segundo a raça, Vestibular 2004, UFMG

Raça	Grupos e situação do candidato							
	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
	Inscritos	Aprovados	Inscritos	Aprovados	Inscritos	Aprovados	Inscritos	Aprovados
Branca	46,4	51,7	64,9	65,5	73,8	74,5	80,0	79,8
Preta	11,1	7,8	4,2	3,9	2,4	1,5	1,2	0,5
Parda	33,3	31,9	23,1	22,8	16,9	16,4	12,7	13,8
Amarela	3,6	2,4	2,6	1,6	1,8	1,4	1,4	0,8
Indígenas	0,7	0,7	0,5	0,3	0,4	0,1	0,2	0,3
Não declarou	5,0	5,5	4,7	5,9	4,7	6,1	4,4	4,7
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Da Tabela 3, determina-se que as rendas medianas para os candidatos inscritos dos Grupos 1, 2, 3 e 4 são, respectivamente, 3,45, 6,49, 12,37 e 26,10 salários mínimos. Nota-se que a renda mediana dos candidatos inscritos do Grupo 4 é aproximadamente oito vezes maior do que a observada para os candidatos inscritos do Grupo 1. Se comparamos as rendas medianas para os candidatos aprovados percebemos que a diferença entre os Grupos 1 e 4 diminuem um pouco (medianas dos Grupos 1 e 4 são 4,13 e 26,19, respectivamente). Para o Grupo 1, pode-se dizer que, aproximadamente, 95% dos candidatos inscrito e aprovado possuem renda entre 1 e 10 salários mínimos, e que quase não há candidatos com renda superior a 20 salários mínimos. Mais de 80% dos candidatos inscrito e aprovado do Grupo 2 possuem renda inferior a 15 salários mínimos. No Grupo 3, percebe-se que a maior porcentagem de candidatos inscrito e aprovado recebem de 5 a 20 salários mínimos. Observa-se que, no Grupo 4, não há candidatos inscritos e nem aprovados com renda inferior a 10 salários mínimos e que mais de 50% deles possuem renda superior a 20 salários mínimos. Nota-se também que, nos Grupos 2, 3 e 4, há uma concentração maior de candidatos provenientes de escolas particulares com o percentual aumentando à medida que o "FSE" aumenta, enquanto para o Grupo 1, que é o grupo com FSE mais baixo, nota-se uma predominância de candidatos vindos de escolas públicas, principalmente de escolas públicas estaduais. Cabe ressaltar que o percentual de candidatos de escolas públicas federais e escolas particulares aprovados é sempre (exceto para particular, Grupo 3) maior ou igual que o percentual de candidatos inscritos vindos destas escolas, em todos os quatro grupos. Quanto à "Raça" nota-se que o Grupo 1 é o único grupo em que a percentagem de candidatos das raças preta e parda (44,4%) é bastante próxima do percentual de candidatos da raça branca. Nos demais grupos a maioria dos candidatos inscritos é da raça branca. Entre os candidatos aprovados, o que se percebe é que a maioria é da raça branca, em todos os quatro grupos.

Em resumo, o Grupo 1 é o grupo que apresenta menor percentual de aprovação (3,91%) e é constituído por candidatos com renda mais baixa e que fizeram o ensino médio em escolas públicas estaduais. Já o Grupo 4 é essencialmente formado por candidatos mais ricos, que em sua grande maioria declararam-se da raça branca e que,

em quase a sua totalidade, fizeram o ensino médio em escolas particulares. O Grupo 4 é o grupo em que se observa o maior percentual de aprovação (15,03%).

Como esperado, a chance de aprovação de um candidato no vestibular parece estar diretamente associada à sua condição socioeconômica. No entanto, a condição socioeconômica parece não ser o único fator associado à maior chance de aprovação do candidato. Da Tabela 4, percebe-se que o percentual de candidatos aprovados vindos das escolas públicas federais é substancialmente maior do que o percentual de inscritos vindos destas escolas. Este efeito positivo das escolas públicas federais não é observado para os demais tipos de escola, nem mesmo para as escolas particulares. Este efeito positivo poderia ser explicado, entre outras coisas, pelo rigoroso processo de seleção utilizado por tais escolas.

4 . Identificando os fatores que levam à aprovação em cada grupo: uma análise via Regressão Logística

No Quadro 6, a seguir, exibe-se os modelos encontrados para cada um dos quatro grupos (Grupos 1, 2, 3 e 4) de candidatos descritos na seção 3.

Quadro 6 - Fatores que influenciaram na aprovação para os Grupos 1, 2, 3 e 4, Vestibular 2004, UFMG

PREDITORA	Grupo 1			Grupo 2			Grupo 3			Grupo 4		
	COEF	*ODDS*	P*	COEF	*ODDS*	P	COEF	*ODDS*	P	COEF	*ODDS*	P
Constante	-2,856	0,058	-----	-2,008	0,134	-----	-0,764	0,466	-----	-1,152	0,316	-----
Curso do ensino médio	-----	-----	-----	-----	-----	-----	-0,432	0,649	0,001	-0,894	0,409	0,001
Estado Civil	-0,498	0,608	0,000	-----	-----	-----	-----	-----	-----	-----	-----	-----
Freqüentou cursinho pré-vestibular?	0,440	1,553	0,000	0,159	1,172	0,027	-----	-----	-----	-----	-----	-----
Há quanto tempo tenta ingressar no ensino superior	0,163	1,177	0,035	-----	-----	-----	-----	-----	-----	-----	-----	-----
Já freqüentou curso de idiomas	-0,335	0,716	0,000	-0,167	0,847	0,011	-----	-----	-----	-----	-----	-----
Já prestou vestibular?	1,040	2,830	0,000	0,844	2,326	0,000	0,385	1,470	0,000	0,370	1,447	0,000
Onde concluiu o ensino médio	-0,300	0,741	0,000	-0,163	0,850	0,022	-0,440	0,644	0,000	-0,770	0,463	0,000
Onde mora	-----	-----	-----	-0,541	0,582	0,000	-0,714	0,490	0,000	-0,708	0,493	0,000
Opção_*Dummy*1	-----	-----	-----	-0,426	0,653	0,000	-----	-----	-----	-0,369	0,692	0,039
Opção_*Dummy*2	-----	-----	-----	-0,034	0,967	0,000	-----	-----	-----	0,527	1,694	0,039
Raça_*Dummy*1	0,095	1,100	0,000	-----	-----	-----	-----	-----	-----	-----	-----	-----
Raça_*Dummy*2	-0,001	0,999	0,000	-----	-----	-----	-----	-----	-----	-----	-----	-----
Sabe redigir um texto no computador?	0,246	1,279	0,023	-----	-----	-----	-----	-----	-----	-----	-----	-----
Sexo	0,381	1,464	0,000	0,293	1,340	0,000	0,230	1,259	0,001	0,221	1,247	0,005
Conhecimento de língua estrangeira	-0,455	0,635	0,000	-0,348	0,706	0,000	-0,810	0,445	0,000	-0,791	0,463	0,000
Tempo de conclusão do ensino médio	-0,542	0,581	0,000	-0,631	0,532	0,000	-0,729	0,482	0,000	-0,485	0,616	0,004
Tipo de escola do ensino médio	-0,791	0,454	0,000	-0,454	0,635	0,000	-0,926	0,396	0,000	-----	-----	-----
Turno do curso na UFMG	-0,347	0,707	0,000	-0,530	0,589	0,000	-0,397	0,672	0,000	-0,323	0,724	0,019
Turno ensino médio	-0,649	0,522	0,000	-0,950	0,387	0,000	-----	-----	-----	-----	-----	0,000
Você cursou ensino médio	-----	-----	-----	-----	-----	-----	-0,728	0,483	0,000	-----	-----	-----

*** *P* denota o P-valor da variável, usando o teste da razão de verossimilhança e *ODDS* denota a razão das chances.**

Destaca-se, em cada caso, os perfis dos candidatos com maior e menor probabilidade de aprovação e as variáveis que mais influenciam a aprovação do candidato. Considera-se nível de 1% de significância para o teste de ajuste do modelo e nível de 5% de significância para a seleção de variáveis.

O candidato do Grupo 1 com maior probabilidade de aprovação (37,97%) já prestou vestibular, é do sexo masculino, solteiro, das raças preta ou parda, optou por um curso diurno na UFMG, concluiu o ensino médio em 3 anos, no turno diurno em Belo Horizonte, em uma escola pública federal ou em escola particular, lê inglês ou duas ou mais línguas estrangeiras, tenta ingressar em um curso superior há mais de um ano, freqüentou cursinho pré-vestibular, já freqüentou curso de inglês e sabe redigir um texto no computador. Um candidato que declarou ser das raças amarela ou indígena ou não declarou sua raça e tem o perfil oposto das demais variáveis citadas acima, teve 0,11% de probabilidade de ser aprovado no vestibular da UFMG de 2004. Para o Grupo 1, as variáveis mais associadas à aprovação são Já prestou vestibular, Tipo de escola de ensino médio e Turno do ensino médio, nesta ordem, sendo favorecidos candidatos que já prestaram algum vestibular, que fizeram o ensino médio em escolas públicas federais ou particulares e no turno diurno. Considerando a razão das chances ou *odds*, percebe-se , por exemplo, que um candidato que cursou ensino médio em escolas pública estadual ou municipal ou fez curso livre tem sua chance de aprovação diminuída em aproximadamente 55%, se seu ensino médio foi feito no turno noturno sua chance de aprovação é reduzida em 47,8% e se não lê inglês ou duas ou mais línguas estrangeiras sua chance de aprovação é diminuída em 36,5%. Observa-se, também, que, se comparados a candidatos da raça branca, candidatos das raças preta ou parda têm sua chance de aprovação aumentada em 10,0% e os candidatos das raças amarela e indígena ou que não declararam sua raça têm sua chance de aprovação diminuída em 0,1%.

O candidato do Grupo 2 com maior probabilidade de aprovação (32,92%) é do sexo masculino, mora em Belo Horizonte, optou por um curso diurno na UFMG, concluiu o ensino médio em 3 anos, em Belo Horizonte, no turno diurno, em uma escola pública federal ou em escola particular, lê inglês ou duas ou mais línguas estrangeiras, já

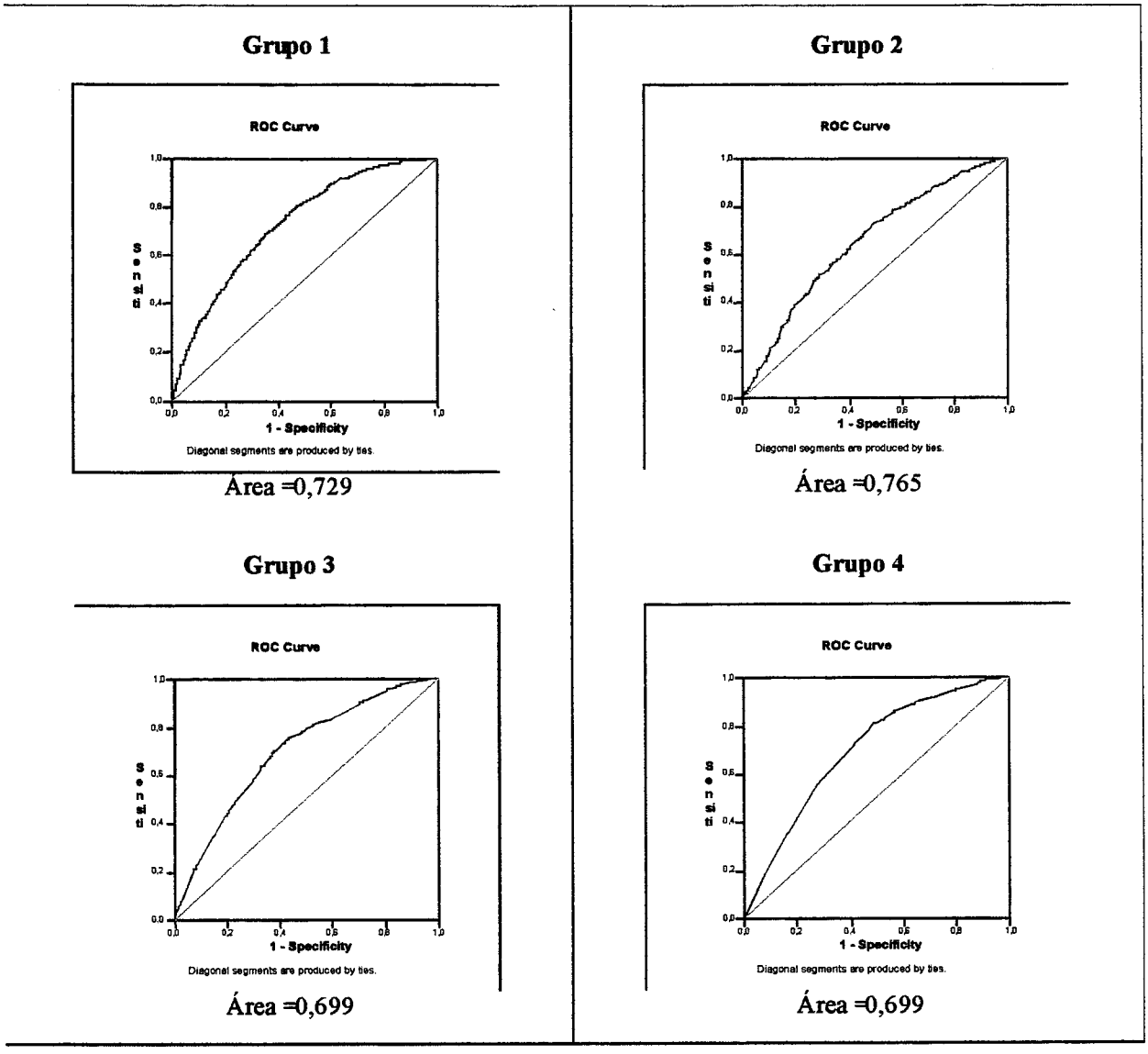
prestou vestibular, freqüentou cursinho pré-vestibular, optou pela prova de espanhol no vestibular e já freqüentou curso de inglês. Um candidato que optou pela prova de espanhol no vestibular e tem o perfil oposto das demais variáveis citadas acima, teve probabilidade 0,19% de ser aprovado no vestibular da UFMG de 2004. Para o Grupo 2, as variáveis mais associadas à aprovação são Turno de ensino médio, Já prestou vestibular e Tempo de conclusão do ensino médio, nesta ordem, sendo favorecidos candidatos que fizeram o ensino médio no turno diurno, já prestaram algum vestibular e que concluíram o ensino médio em 3 anos. Percebe-se, ainda, para este grupo, que um candidato que fez o ensino médio no turno noturno tem sua chance de aprovação reduzida em 61,3%, se não lê o inglês ou duas ou mais línguas estrangeiras sua chance de aprovação é diminuída em 29,4% e se cursou ensino médio em escolas pública estadual ou municipal ou fez curso livre sua chance de aprovação é diminuída em 36,5%.

Para o Grupo 3, nota-se que o candidato com maior probabilidade de aprovação (46,28%) é do sexo masculino, mora em Belo Horizonte, optou por um curso diurno na UFMG, concluiu o ensino médio não-profissionalizante em 3 anos, em Belo Horizonte, em uma escola pública federal, lê inglês ou duas ou mais línguas estrangeiras e já prestou vestibular. Um candidato cujo perfil é oposto ao descrito acima teve probabilidade de 0,26% de ser aprovado no vestibular da UFMG de 2004. Neste grupo, nota-se que as variáveis mais associadas à aprovação são Tipo de escola de ensino médio, Conhecimento de língua estrangeira e Tempo de conclusão do ensino médio. Como para os outros dois grupos, é favorecido um candidato que fez o ensino médio em escola pública federal ou particular e em 3 anos. Além disto, o candidato deve ler o inglês ou duas ou mais línguas estrangeiras. Para este grupo, nota-se que um candidato que fez o ensino médio em escolas pública estadual ou municipal ou fez curso livre tem sua chance de aprovação reduzida em 60,9%, se não lê o inglês ou duas ou mais línguas estrangeiras sua chance de aprovação é diminuída em 55,5% e se tenta ingressar na universidade a mais de um ano esta chance é reduzida em 51,2%.

No Grupo 4, que é formado por candidatos mais favorecidos do ponto de vista socioeconômico, as variáveis mais associadas à aprovação são Curso de ensino

médio, Conhecimento de língua estrangeira e Onde concluiu o ensino médio, sendo favorecidos candidatos que fizeram ensino médio não-profissionalizante, lêem o inglês ou duas ou mais línguas estrangeiras e moram em Belo Horizonte. Neste grupo, um candidato com maior probabilidade de aprovação (49,15%) é do sexo masculino, mora em Belo Horizonte, optou por um curso diurno na UFMG, concluiu o ensino médio não-profissionalizante em 3 anos, em Belo Horizonte, lê inglês ou duas ou mais línguas estrangeiras, já prestou vestibular e optou pela prova de inglês no vestibular. Um candidato que optou pela prova de francês no vestibular e tem o perfil oposto das demais variáveis citadas acima, teve probabilidade 0,11% de ser aprovado no vestibular da UFMG de 2004. Neste grupo, percebe-se que um candidato que não lê inglês ou duas ou mais línguas estrangeiras ou que não concluiu o ensino médio em Belo Horizonte tem sua chance de aprovação diminuída em 53,7%. Se o candidato fez curso do ensino médio profissionalizante, supletivo ou outro equivalente sua chance de aprovação é diminuída em 59,1%.

Figura 2 - Curva ROC para os modelos logísticos de cada grupo de candidatos ao Vestibular 2004, UFMG



Outro ponto interessante a destacar é que um candidato ter prestado o vestibular antes do concurso Vestibular 2005 aumenta a sua chance de aprovação, nos quatro grupos. Este aumento é superior a 44,0% nos quatro grupos, sendo maior que 100,0% nos Grupos 1 e 2.

Devemos ressaltar ainda que as variáveis Raça, Há quanto tempo tenta ingressar no ensino superior, Estado civil e Sabe redigir um texto no computador só se mostraram relevantes para explicar a aprovação dos candidatos do Grupo 1, que é o grupo de candidatos, do ponto de vista socioeconômico, menos favorecidos. A variável Freqüentou cursinho pré-vestibular, Já freqüentou curso de idiomas e Turno de ensino médio só aparecem como variáveis relevantes nos dois grupos menos favorecidos do ponto de vista socioeconômico (Grupos 1 e 2). A variável "Curso do ensino médio" só é relevante para explicar a aprovação nos Grupos 3 e 4. Em diferentes ordens de relevância, as variáveis que aparecem como explicativas para a aprovação ou não do candidato nos quatro grupos são: Já prestou vestibular, Onde concluiu o ensino médio, Sexo, Conhecimento de língua estrangeira, Tempo de conclusão do ensino médio e Turno do curso na UFMG. Entre os fatores mais fortemente associados à aprovação dos candidatos dos dois grupos menos favorecidos do ponto de vista socioeconômico encontram-se Já prestou vestibular e Turno de ensino médio, enquanto para os candidatos dos dois grupos mais favorecidos do ponto de vista socioeconômico (Grupos 3 e 4) encontra-se a variável Conhecimento de língua estrangeira.

A Figura 2 mostra as curvas ROC para os modelos logísticos construídos para cada um dos quatro grupos de candidatos. Em todos os casos, a área sob a curva ROC foi significativamente diferente de 0,5, indicando que os modelos construídos têm alguma capacidade de classificação. No entanto, notamos que o modelo referente ao Grupo 2 tem melhor capacidade de classificação que os modelos construídos para os demais grupos e que nenhum dos modelos pode ser considerado um excelente modelo (áreas inferiores a 0,8), ou seja, para melhorar a capacidade de classificação de cada modelo devemos considerar outros fatores além dos fatores considerados neste trabalho.

5. Conclusões e discussões

A democratização do acesso ao ensino superior é uma questão que vem sendo amplamente discutida por parte da sociedade, nestes últimos tempos, motivada pela reforma universitária que está sendo implantada pelo governo federal. Parte deste

debate, no âmbito da UFMG, pode ser encontrado em Peixoto (2004) que reúne artigos e opiniões de pesquisadores sobre a educação no Brasil e sobre meios para a ampliação do acesso ao ensino superior. As ampliações de cursos noturnos e do ensino a distância são apontadas como alternativas às cotas, neste debate.

"O propósito de uma universidade publica é criar oportunidades de progresso para todas as pessoas, especialmente àquelas que não têm acesso a universidades particulares, por causa de seu nível de pobreza e da baixa preparação para a vida acadêmica." (Hammond (2004), página 119). Segundo Moraes (2004) o acesso ao ensino superior melhora as possibilidades de ascensão social do indivíduo e de sua capacidade de receber mais renda, mas este não é o fruto mais importante e mensurável do ensino superior no Brasil. O maior impacto do ensino superior é aquele produzido no desenvolvimento da sociedade. Moraes (2004), página 77, acrescenta ainda "... *desenvolver o ensino superior é uma estratégia essencial ao interesse público mais elevado, e não somente algo que se defina no âmbito do interesse privado, da ascensão social individual."*

É praticamente um consenso, entre os cidadãos brasileiros, que as universidades federais são elitistas, uma vez que, a maioria dos alunos que ingressam em seus cursos possuem boas condições socioeconômicas. No entanto, como mostrado por Moraes (2004), o percentual de "pobres" nas universidades públicas desenha uma pirâmide similar à observada para as escolas particulares em algumas faixas de renda que é bem menos aguda do que a observada para a totalidade da população, ou seja, segundo Moraes (2004), se a educação superior não elimina a perversidade da concentração de renda, está longe de aprofundá-la. Peixoto e Braga (2004) concluem que a seletividade social no vestibular da UFMG, medida pela diferença entre os FSEs médios de candidatos inscrito e aprovado, é uma realidade, mas tal seletividade é mais acentuada em alguns cursos que em outros. O curso de Medicina, por exemplo, é o curso que apresenta menor seletividade social, ou seja, é um curso que é, raramente, procurado por candidatos menos favorecidos do ponto de vista socioeconômico. Peixoto e Braga (2004) também concluem que a chance de aprovação no vestibular da UFMG é fortemente associada a aspectos econômico e social dos candidatos, sendo

particularmente relevante o conhecimento de língua estrangeira, o acesso a cursinho pré-vestibular, o tipo de escola de ensino médio e o local de moradia. Concluem que um candidato que apresenta habilidade de leitura em alguma língua estrangeira, que freqüentou cursinho pré-vestibular, mora em Belo Horizonte e cursou escola particular ou pública federal tem maior chance de aprovação. Soares (2004) sugere dois caminhos para se aumentar o número de egressos de escolas públicas na UFMG: o primeiro é a criação ou transformação das escolas públicas de ensino médio em ambientes favoráveis ao desempenho acadêmico; e o outro é implementar uma política de cotas em que a universidade a cada ano defina um mínimo aceitável de desempenho para cada curso. A idéia é escolher candidatos vindos de escolas públicas, que tenham mérito, e foram prejudicados pelas condições encontradas no ensino médio. Soares (2004) salienta que, se em 2000 a reserva de 50% das vagas para candidatos de escolas públicas estivesse em vigor, haveria uma queda considerável nas notas mínimas de ingressos na USP, podendo levar a situações preocupantes com a classificação de um candidato do curso de Medicina que tira zero em uma prova específica.

Neste trabalho, identificou-se a presença de quatro grupos de candidatos ao Vestibular 2004 da UFMG que podem ser considerados homogêneos do ponto de vista socioeconômico, identificando-se os fatores associados com a aprovação do candidato em cada um destes grupos. Utilizou-se, para estes fins, o CART e a Regressão Logística. O objetivo subjacente é conhecer melhor os diferentes perfis dos candidatos que ingressam na UFMG e fornecer ferramentas para a discussão sobre a democratização do acesso ao ensino superior.

Os grupos foram determinados exclusivamente pela variável "FSE". Nota-se que a renda varia de acordo com os grupos, havendo maior concentração de candidatos em valores maiores de renda para o Grupo 4, que é o grupo com maior FSE. Neste grupo, observa-se que mais de 50% dos candidatos inscrito e aprovado possuem renda superior a 20 salários mínimos, são predominantemente da raça branca e, em sua quase totalidade, concluíram o ensino médio em escolas particulares. Já para o Grupo 1, que é constituído por candidatos cujos valores para o "FSE" são menores, nota-se que mais de 90% dos candidatos inscrito e aprovado possuem renda entre 1 e 10 salários mínimos

e são candidatos que em sua maioria concluíram o ensino médio em escola pública estadual. Ainda no Grupo 1 observa-se, entre os inscritos, que o percentual de candidatos que se declaram das raças preta ou parda (44,4%) é bem próximo do percentual de candidatos da raça branca. No entanto, entre os candidatos aprovados do Grupo 1, a maioria declarou-se da raça branca.

Conclui-se, portanto, que existem perfis diversos nos diferentes grupos socioeconômicos. Nota-se que para grupos socioeconomicamente menos favorecidos, a realização de atividades que complementam a formação básica do candidato como, por exemplo, freqüentar cursos pré-vestibular e de idiomas e saber usar computador tendem a aumentar a chance de ingresso na universidade. Também observa-se que ter concluído o ensino médio em Belo Horizonte, em três anos e em turno diurno, ler inglês ou duas ou mais línguas estrangeiras e ter prestado vestibular alguma vez (não como treinante) tendem a favorecer a aprovação do candidato em todos os grupos, ou seja, diferentemente do observado por Peixoto e Braga (2004), candidatos de diferentes grupos socioeconômicos têm sua chance de aprovação influenciada por diferentes fatores. No entanto, similar ao observado em Peixoto e Braga (2004), também aqui, observa-se que o tipo de escola de ensino médio e o conhecimento de língua estrangeira são fatores fortemente associados à aprovação em todos os grupos, exceto no Grupo 4 onde os candidatos vêm quase em sua totalidade de escolas particulares. Percebe-se que candidatos que fizerem o ensino médio em escola pública federal ou particular e que lêem inglês ou duas ou mais línguas estrangeiras têm sua chance de aprovação aumentada.

Observa-se que as conclusões obtidas neste trabalho corroboram a proposta feita por Soares (2004), ou seja, uma maneira de ampliarmos o acesso ao ensino superior por candidatos menos favorecidos do ponto de vista socioeconômico, à longo prazo, poderia ser a criação ou transformação das escolas públicas de ensino médio em ambientes favoráveis ao desempenho acadêmico, onde os alunos recebessem estimulo para aprender, tivessem mais acesso à informação através de bibliotecas mais equipadas e computadores com acesso à Internet, e a cursos de idiomas que fornecessem a ele a formação mínima para ler e compreender um texto técnico. Este parece ter sido o

caminho seguido por algumas escolas federais de ensino médio, que, em geral, apresentam um alto índice de aprovação do vestibular, superando muitas vezes os percentuais observados para as escolas particulares.

Cabe ainda um comentário sobre a variável "Raça" que é uma variável importante na atual discussão sobre cotas para as Instituições Federais de Ensino Superior. Nota-se que esta variável só se mostrou relevante para explicar a aprovação dos candidatos do Grupo 1, que é o grupo de candidatos menos favorecidos do ponto de vista socioeconômico. Ressalta-se, no entanto, que no questionário socioeconômico aplicado pela UFMG, tal variável é informada pelo candidato, o que pode acarretar problemas na classificação. Neste trabalho, o eventual erro de classificação não foi levado em consideração na modelagem dos dados (ver Spiegelman *et al.* (2000) para o modelo logístico com erro de classificação nas covariáveis).

Para finalizar, deixa-se uma frase encontrada em Moraes (2004), página 77; *"Por isso, também a educação - como aliás outras políticas sociais – não deve ser entendida apenas como política social compensatória, que eventualmente corrija distorções e disparidades produzidas pelo funcionamento do sistema econômico. Não deve ser vista apenas como prática que em uma palavra, redistribua a riqueza já existente, segundo juízos éticos ou de conveniência ditada pela correlação das forças sociais em confronto. Ela é também, e talvez mais fortemente, parte integrante de uma política de desenvolvimento, garantindo a manutenção, aperfeiçoamento e integração dos cidadãos no processo econômico e social, de modo a viabilizar a ampliação da riqueza e o desenvolvimento das forças produtivas."*, ou seja, muito há que se fazer para chegar-se a uma situação ótima e justa para todos.

Referências bibliográficas

BRAGA, M.M., PEIXOTO, M.C.L. E BOGUTCHI,T.F. (2001). Tendências da demanda pelo ensino superior: estudo de caso da UFMG. *Cadernos de Pesquisa*, 113, 129-152, julho.

BELL, J.F. (1996). Application of Classification Trees to the Habit Preference of Upland Birds. *Journal of Applied Statistics*, 23 (2 e 3), 349-359.

DIAS,T.F.S, LAGE, L.V., RIBEIRO, R.L.F., ROCHA, G.H.M.A., RODRIGUES, J.G., SANTOS, T.R., FRANCO, G.C., LOSCHI, R.H. E BRAGA, M.M. (2005). *Identificação dos fatores que levaram à aprovação no Vestibular-2004 da UFMG via CART: uma comparação dos cursos diurnos e noturnos*. Manuscrito não publicado. Departamento de Estatística, Universidade Federal de Minas Gerais.

DINIZ, C.A.R. E LOUZADA-NETO, F. (2000). *Data Mining: Uma Introdução*. Associação Brasileira de Estatística (ABE). 14°. SINAPE.

HAMMOND, J. L. A política da educação superior da administração Bush. *Educação e Linguagem*, 7 (10), 105-123.

HANLEY, J.A. AND MCNEIL, B.J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143 (1), 29-36.

HOSMER, D.W AND LEMESHOW, S. (1989). *Applied Logistic Regression*. Segunda Edição, New York: J. Wiley.

LOPES, C.B., RIBEIRO, R.L.F., FRANCO, G.C. E LOSCHI, R.H. (2005). *Escola Pública versus Escola Privada: Fatores que levaram candidatos à aprovação no Vestibular-2004 da UFMG.* Manuscrito não publicado. Departamento de Estatística, Universidade Federal de Minas Gerais.

MARTINEZ, E. Z., LOUZADA-NETO, F. E PEREIRA, B.B. (2003). A curva ROC para testes diagnósticos. *Caderno de Saúde Coletiva*, 11(1), 7-31.

MORAES, R. C. (2004). Ensino superior no Brasil-Balanços e perspectivas a partir de 2003. *Educação e Linguagem*, 7 (10), 68-104.

PEIXOTO, M.C.L. E BRAGA, M.M. (2004). Demanda pelo ensino superior no Brasil: o caso da UFMG. *Educação e Linguagem*, 7 (10), 124-149.

PEIXOTO, M.C.L. (organizadora) (2004). *Universidade e Democracia: Experiências e alternativas para a ampliação do acesso à Universidade pública brasileira.* Editora UFMG, Belo Horizonte, Minas Gerais.

SOARES, J. F. (2004). Implementação de cotas na UFMG para alunos egressos de escolas públicas. Em *Universidade e Democracia: Experiências e alternativas para a ampliação do acesso à Universidade pública brasileira.* (M.C.L.Peixoto, organizadora),153-172.

SPIEGELMAN, D, ROSNER, B. AND LOGAN, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449), 51-61.

ZWEIG, M. H. AND CAMPBELL, G. (1993). Receiver Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*, 39 (4), 561-577.

Abstract

This paper aims at obtaining groups of candidates that are homogeneous from the socioeconomic point of view and identifying the factors related to the educational background that led candidates to the approval in the Vestibular-2004 of UFMG in each of these groups. All variables considered in the analysis are defined in the socioeconomic and cultural questionnaire answered by the candidates. To analyze data the following methodologies were considered: Logistic regression, classification and regression tree (CART) - which was considered for clustering analysis - and the ROC curve which was employed as a goodness of fit measure. In short, it can be concluded that there are four groups of candidates, which can be considered homogeneous from the socioeconomic point of view. It is noticeable that, the two groups with low socioeconomic conditions have a large percentage of candidates that attended high school in public establishments, mainly in state schools, and have income up to 10 minimum salaries. In the two groups in which it was observed candidates with better socioeconomic conditions, there are large percentages of candidates that attended high school in private schools and have income higher than 20 minimum salaries. It can be noticed that for the two groups with low socioeconomic conditions, the main factors associated with the approval of the candidates are if they have already done the "vestibular" exam (Group 1) and if they attended high school during daytime (Group 2). Concerning the most privileged socioeconomic groups, the approval is strongly associated with the fact that the candidate attended high school in private or federal institutions (Group 3) and attended high school courses that do not provided special training for a specific profession (Group 4).

Key words: CART, ROC curve, race, logistic regression, type of high schools.

Reamostragem *bootstrap* em delineamentos por conjuntos imperfeitamente ordenados

Maria Cecília Mendes Barreto[*]
Cesar Augusto Taconeli[**]

Resumo

A impossibilidade da extração de uma amostra numerosa requer a utilização de delineamentos amostrais capazes de produzir estimativas mais precisas. Neste contexto, a amostragem por conjuntos ordenados é uma opção indicada quando se tem uma variável de interesse de difícil mensuração, havendo, no entanto, a possibilidade de se ordenar amostras aleatórias de maneira simples, barata e eficiente de acordo com o possível valor da variável de interesse, antes de sua efetiva mensuração. Este trabalho generaliza o procedimento de *reamostragem bootstrap* em conjuntos ordenados apresentado em Taconeli e Barreto (2005), estendendo a metodologia para situações em que a ordenação das amostras é passível de falhas. Na avaliação do impacto de incorreções no processo de ordenação na estimação intervalar *bootstrap* da média populacional, verificou-se que os intervalos percentil, normal, básico, Bca e studentizados gerados pelo procedimento de *reamostragem bootstrap*, aqui proposto, apresentam maior precisão em relação aos intervalos de confiança *bootstrap* análogos baseados em amostragem aleatória simples, independendo da distribuição, tamanho de amostra, grau de imperfeição no procedimento de ordenação e tipo de intervalo de confiança *bootstrap*. Este ganho em precisão decresce conforme aumenta a imprecisão do critério adotado para ordenação, sendo, entretanto, significativo para as situações mais críticas com erros de maior magnitude. Os

[*] Endereços para correspondências: Deptº de Estatística – Universidade Federal de São Carlos – UFSCar, São Carlos, SP – e-mail: cbarreto@power.ufscar.br.
[**] Deptº de Bioestatística – IB – UNESP – Botucatu, SP – e-mail:ctaconeli@yahoo.com.br.

Intervalos de confiança *bootstrap* do tipo studentizado são os mais acurados dentre os considerados. Os procedimentos, aqui apresentados, são utilizados na estimação da altura média de árvores cereja preta, tendo como critério de ordenação imperfeita o diâmetro do tronco à altura do peito.

Palavras-chave: Amostragem por conjuntos ordenados, ordenação imperfeita, intervalos de confiança para a média populacional, reamostragem *bootstrap*.

1. Introdução

A amostragem por conjuntos ordenados tem se mostrado de notável aplicabilidade em situações em que não se pode medir um grande número de unidades amostrais, sendo possível, no entanto, ordená-las de modo eficiente sem de fato efetuar tais mensurações. Esta ordenação pode se basear em algum critério subjetivo, invariavelmente através de julgamento visual, ou de acordo com uma variável concomitante que apresente uma grande correlação com a variável de interesse.

Ao selecionar uma amostra de domicílios para aplicação de um questionário aos moradores, por exemplo, pode-se considerar setores de casas e buscar ordenar as famílias de acordo com suas condições socioeconômicas, comparando-as através de algumas características das residências, como tamanho e estado de conservação.

Uma maneira de se estimar o volume total de madeira em uma determinada região é através da altura de suas árvores, que em área de grande densidade, a inspeção visual ou o uso de escadas e cordas torna-se impraticável, dispendioso e de difícil manejo. Nessas situações, a variável auxiliar utilizada é diâmetro na altura do peito e para o delineamento amostral por conjuntos ordenados é usada na ordenação das árvores em relação a sua altura, antes de se fazer sua efetiva mensuração usando cordas e escadas.

McIntyre (1952) introduz a amostragem por conjuntos ordenados (*ranked set sampling – RSS*) com aplicação na estimação da produção média de pastagens e propõe a média amostral ($\overline{\overline{X}}$) como estimador não-viciado da média populacional. Takahasi e Wakimoto (1968) comprovam a menor variância do estimador da média via RSS sob ordenação perfeita em relação a $\overline{X}$, a média de uma amostra aleatória simples (*single random sample - SRS*) de mesmo tamanho. Dell e Clutter (1972) consideram amostras

ordenadas imperfeitamente e verificam através de simulação, considerando diferentes distribuições e graus de imperfeição, a maior precisão do estimador $\overline{\overline{X}}$ em relação a $\overline{X}$.

Nahhas et al (2002) apresentam a variância de $\overline{\overline{X}}$, supondo ordenação imperfeita de uma variável com distribuição normal com variância conhecida, possibilitando a construção de intervalos de confiança para a média. Sob outras distribuições, a indeterminação de uma forma analítica para a precisão do estimador média amostral via RSS dificulta a determinação de tais estimativas.

Cesário e Barreto (2003) propõem um esquema paramétrico de reamostragem *bootstrap* via RSS com ordenação perfeita e avaliam diferentes intervalos de confiança. Através de simulação, verificam que os intervalos *bootstrap* possuem probabilidades de cobertura bastante próximas aos níveis de confiança desejados. Na situação de desconhecimento do valor do parâmetro de escala (σ), verificam melhor desempenho para intervalos do tipo studentizado.

Também para o delineamento por conjuntos ordenados, Taconeli e Barreto (2003) verificam via simulação que a distribuição de $\overline{\overline{X}}$ para variáveis com distribuição normal é aproximadamente normal, e, para variáveis com distribuição exponencial, aproximadamente gama. A utilização destas distribuições na construção de intervalos de confiança para a média mostra-se eficaz no caso da distribuição normal, principalmente quando menores graus de imperfeição considerados. Brandão (2003) estende este estudo para dados com distribuição uniforme e lognormal, diagnosticando, apenas para a primeira distribuição, a simetria da distribuição do estimador $\overline{\overline{X}}$ para diversos tamanhos de amostras, além da adequabilidade dos intervalos assintóticos normais na estimação da média via RSS.

Taconeli e Barreto (2005) propõem um procedimento original não-paramétrico de reamostragem *bootstrap* baseado em amostras por conjuntos ordenados perfeitamente, com aplicação na obtenção de intervalos de confiança para a média. Verificam, através de simulação, que as estimativas concebidas mediante a metodologia proposta apresentam um ganho de precisão, sem comprometimento quanto à acurácia, em relação às estimativas originadas através do *bootstrap* convencional que é baseado em SRS. Esta conclusão se estende para diferentes distribuições e tamanhos de amostra.

O presente trabalho tem como motivação dar prosseguimento aos estudos realizados em Taconeli e Barreto (2005), através da incorporação de erros ao processo de ordenação das amostras. Tem-se como objetivo apresentar o esquema original de reamostragem *bootstrap* em conjuntos ordenados imperfeitamente, aplicá-lo na obtenção de estimativas intervalares para a média populacional e avaliá-lo comparando tais resultados aos obtidos via SRS e RSS com ordenação perfeita. Usando um conjunto de dados disponível na literatura estatística, o procedimento de reamostragem *bootstrap* em conjuntos ordenados é utilizado para se obter intervalos de confiança para a altura média de árvores.

2. Amostragem por conjuntos ordenados: seleção da amostra, alguns estimadores e modelo com erros de ordenação

A extração de uma amostra por conjuntos ordenados parte da seleção de n amostras aleatórias simples de tamanho n da população de interesse. Em cada uma destas amostras, é realizada a ordenação dos elementos em ordem crescente de seu possível valor da variável de interesse, baseado em algum critério subjetivo, que pode ser um julgamento visual, ou através de alguma variável concomitante fortemente correlacionada. Na r-ésima amostra mensura-se a observação julgada como tendo a r-ésima menor medida da variável de interesse, resultando em uma amostra por conjuntos ordenados de tamanho n. Este procedimento pode ser replicado m vezes, originando uma amostra de tamanho $N = mn$, que pode ser expressa como

$$\left\{x_{[r]i}; r = 1,2,...,n; i = 1,2,...,m\right\}. \tag{2.1}$$

Com o objetivo de ilustrar o processo de seleção de uma amostra por conjuntos ordenados, considere os dados relativos à altura (medida em pés) e ao diâmetro (medido em polegadas) de 31 pés de cerejas pretas (Ryan et al, 1976). Supondo interesse em inferir a respeito da altura das árvores, utilizou-se como critério para ordenação das amostras o diâmetro dos troncos das mesmas, tomado a uma altura de 4 a 6 pés acima

do solo. Nesta aplicação, o tamanho das amostras (n) é igual a 4 e o número de réplicas (m) é igual a 2, resultando numa amostra por conjuntos ordenados de tamanho $N=8$. A Tabela 1 detalha o delineamento amostral, apresentando em negrito as árvores que de fato constituem a amostra por conjuntos ordenados.

O estimador não-viciado proposto por McIntyre (1952) para μ_X, a média populacional, é a média da amostra por conjuntos ordenados

$$\overline{\overline{X}} = \frac{1}{mn}\sum_{r=1}^{n}\sum_{i=1}^{m} X_{[r]i}\,, \tag{2.2}$$

com variância dada por

$$Var(\overline{\overline{X}}) = \frac{1}{mn^2}\sum_{r=1}^{n}\sigma_r^2\,. \tag{2.3}$$

Para dados com distribuições simétricas (uniforme e normal) e assimétrica (exponencial), Dell e Clutter (1972) verificam a maior precisão do estimador $\overline{\overline{X}}$ em relação a $\overline{X}$, mesmo para a situação mais crítica, em que se considera erros de ordenação de maior magnitude. Este ganho, no entanto, decresce à medida que aumenta o grau de imperfeição na ordenação das amostras.

Tabela 1 – Seleção de uma amostra de pés de cerejas pretas por conjuntos ordenados.

Réplica 1											
Amostra 1			Amostra 2			Amostra 3			Amostra 4		
Diâm.	Alt.	*Rank* (diâm.)	Diâm.	Alt.	*Rank* (diâm.)	Diâm.	Alt.	*Rank* (diâm.)	Diâm.	Alt.	*Rank* (diâm.)
11,1	80	2	17,5	82	3	14,5	74	2	13,8	64	1
14,0	78	4	20,6	87	4	20,6	87	4	**18,0**	**80**	**4**
13,7	71	3	**12,9**	**74**	**2**	10,7	81	1	17,5	82	3
8,3	**70**	**1**	11,0	75	1	**17,9**	**80**	**3**	14,0	78	2
Réplica 2											
Amostra 1			Amostra 2			Amostra 3			Amostra 4		
Diâm.	Alt.	*Rank* (diâm.)	Diâm.	Alt.	*Rank* (diâm.)	Diâm.	Alt.	*Rank* (diâm.)	Diâm.	Alt.	*Rank* (diâm.)
8,8	**63**	**1**	18,0	80	4	**13,3**	**86**	**3**	18,0	80	3
13,8	64	4	13,3	86	1	10,8	83	1	17,5	82	2
12,9	85	2	17,9	80	3	11,4	76	2	**20,6**	**87**	**4**
13,7	71	3	**13,7**	**71**	**2**	16,3	77	4	11,4	76	1

Nas situações em que o critério utilizado na ordenação das amostras é algum julgamento subjetivo (geralmente visual), o modelo com erros de ordenação (Dell e Clutter, 1972) pode ser escrito da seguinte maneira

$$Y_i = X_i + e_i, \quad e_i \sim N(0, \sigma_e^2), \tag{2.4}$$

onde Y representa o valor da variável X (que é de interesse), acrescido de um erro aleatório devido a falhas de ordenação cuja variância é σ_e^2.

Nahhas et al (2002) verificam que, usando o modelo de erros de ordenação (2.4) e supondo que a variável X é normalmente distribuída com variância σ^2 e Y tem variância $\sigma^2 + \sigma_e^2$, a precisão relativa (RP) de $\overline{X}$ em relação a $\overline{\overline{X}}$ pode ser escrita como

$$RP = \frac{\text{var}(\overline{X})}{\text{var}(\overline{\overline{X}})} = \left\{ 1 - \rho_{xy}^2 \sum_{r=1}^{n} \alpha_{[r]}^2 / n \right\}^{-1}, \tag{2.5}$$

onde $\alpha_{[r]}$ é a média da r-ésima estatística de ordem de uma amostra de tamanho n com distribuição normal padronizada e $\rho_{xy} = corr(X,Y) = \sigma/(\sigma_\varepsilon^2 + \sigma^2)^{1/2}$.

Nahhas e Wolfe (manuscrito não publicado, citado em Nahhas et al (2002)) mostram que, se a variável de interesse tiver distribuição normal, a probabilidade esperada de ordenar incorretamente um par (Xi, Y_i) é

$$[\{1/2\}] - \{\tan^{-1}(\sigma/\sigma_\varepsilon)\}/\pi. \tag{2.6}$$

tem-se então que

$$(\sigma/\sigma_\varepsilon) = \left[\tan^2\{\pi(1-2p)/2\}\right]^{-1} \tag{2.7}$$

e

$$\rho_{xy}^2 = f(p) = \left[1 + \left[\tan^2\{\pi(1-2p)/2\}\right]^{-1}\right]. \tag{2.8}$$

Assim a precisão relativa pode ser expressa por

$$RP = \left\{1 - \frac{f(p)}{n}\sum_{r=1}^{n}\alpha_{[r]}^2\right\}^{-1} \tag{2.9}$$

ou seja, na presença de erros de ordenação, a variância do estimador $\bar{\bar{X}}$, quando a distribuição da variável de interesse é normal, é dada como

$$Var(\bar{\bar{X}}) = \left\{1 - \frac{f(p)}{n}\sum_{r=1}^{n}\mu_{[r]}^2\right\}^{-1} n/\sigma^2. \tag{2.10}$$

que possibilita a estimação do erro padrão de $\overline{\overline{X}}$ de maneira adequada, servindo como base para a obtenção de intervalos de confiança, sob a condição de que a variável de interesse tenha distribuição normal. A suposição de normalidade, aliada à necessidade de conhecimento do valor de σ, restringe bastante a aplicabilidade do resultado (2.10) na obtenção de estimativas intervalares.

A utilização da reamostragem *bootstrap* não-paramétrica na obtenção de estimativas do erro padrão de $\overline{\overline{X}}$ torna desnecessário qualquer tipo de suposição ou conhecimento prévio acerca da distribuição da variável de interesse. A estimação da função distribuição acumulada (f.d.a.) através da função distribuição empírica (f.d.e.) constitui a essência do *bootstrap* não-paramétrico.

Considere $\{x_{[r]i}; r = 1,2,...,n; i = 1,2,...,m\}$ uma amostra por conjuntos ordenados de tamanho $N = mn$, extraída de uma população com distribuição F. Stokes e Sager (1988) propõem a f.d.e. obtida via RSS como estimador não-viciado de F, da seguinte forma

$$F^*(t) = \frac{1}{mn}\sum_{r=1}^{n}\sum_{i=1}^{m} I_{(-\infty, x_{[r]i})}(t), \qquad (2.11)$$

sendo $I_{(.,.)}(\cdot)$ a função indicadora, e comprovam analiticamente a maior precisão de F^* em relação ao estimador f.d.e. obtido a partir de uma amostra aleatória simples de mesmo tamanho ($\hat{F}$.). Este resultado é válido tanto para ordenação perfeita como imperfeita.

3. Uma proposta original de reamostragem *bootstrap* em amostragem por conjuntos ordenados perfeitamente e intervalos de confiança para a média e sua utilização na amostra de pés de cerejeiras

A reamostragem *bootstrap* (Davison e Hinkley, 1997) consiste na seleção de amostras aleatórias simples com reposição da amostra original (*bootstrap* não-paramétrico) ou da distribuição da variável de interesse, com parâmetros estimados

através da amostra originalmente extraída (*bootstrap*-paramétrico). A amostra assim selecionada é chamada de reamostra *bootstrap*.

Considere $X_1, X_2, ..., X_n$ uma amostra aleatória simples de tamanho n originária de uma população com distribuição F, suponha interesse na estimação de um parâmetro $\theta = t(F)$ e seja $\hat{\theta}$ um estimador qualquer de θ. O *bootstrap*, segundo Canty (2004), tem como princípio a determinação de uma distribuição $\tilde{F}$, denominada função plug-in, de modo que $\hat{\theta} = t(\tilde{F})$. Desta forma, pode-se obter réplicas de $\hat{\theta}$, denotadas por $\hat{\theta}^*$, a partir da distribuição $\tilde{F}$, possibilitando a seguinte aproximação

$$(\hat{\theta} - \theta) \overset{D}{\approx} (\hat{\theta}^* - \hat{\theta}). \tag{3.1}$$

O *bootstrap* não-paramétrico utiliza $\tilde{F} = \hat{F}$, a f.d.e. obtida via SRS. Cientes da maior precisão do estimador F^* em relação a $\hat{F}$ na estimação de F, Taconeli e Barreto (2005) propõem a substituição da função plug-in (de $\hat{F}$ para F^*), através de um procedimento de reamostragem *bootstrap* baseado na extração de amostras e geração de reamostras por conjuntos ordenados perfeitamente.

Propomos, então, uma extensão natural de reamostragem *bootstrap* para conjuntos imperfeitamente ordenados (Taconeli, 2005) descrito no Algoritmo 1.

Algoritmo 1

Passo 1: Seleção de uma amostra por conjuntos ordenados

$$\{X_{[r]i}, r = 1,2,...,n; i = 1,2,...,m\},$$

utilizando algum critério de ordenação passível de falhas;

Passo 2: Ordenação da amostra obtida no passo 1 através dos valores da variável Y, o que na prática equivale a ordená-la segundo o critério de ordenação considerado anteriormente;

Passo 3: Atribuição de *ranks* de 1 a *mn* às unidades amostrais selecionadas no passo 1, de acordo com a ordenação estabelecida no passo 2: $r_1, r_2, ..., r_{mn}$;

Passo 4: Geração de B reamostras $X_1^*, X_2^*, ..., X_B^*$, segundo procedimento de amostragem por conjuntos ordenados, utilizando como critério para ordenação das unidades amostrais selecionadas os *ranks* determinados no passo 3.

Os intervalos de confiança *bootstrap* descritos em Davison e Hinkley (1997) podem, então, ser utilizados para estimar a média populacional usando amostras por conjuntos ordenados com a reamostragem *bootstrap* descrita no Algoritmo 1.

Assim, o intervalo *bootstrap* normal pressupõe normalidade à distribuição de $(\bar{\bar{X}} - \mu_X)$ e é dado por

$$\left(\bar{\bar{X}} - z_{1-\alpha} \sqrt{Var_B(\bar{\bar{X}})}, \bar{\bar{X}} - z_{\alpha} \sqrt{Var_B(\bar{\bar{X}})} \right), \tag{3.2}$$

sendo

$$Var_B(\bar{\bar{X}}) = \frac{1}{B-1} \sum_{b=1}^{B} (\bar{\bar{X}}_b^* - \bar{\bar{X}}^*)^2 \tag{3.3}$$

e

$$\bar{\bar{X}}^* = \frac{1}{B} \sum_{k=1}^{B} \bar{\bar{X}}_k^* \tag{3.4}$$

e $\bar{\bar{X}}_k^*$ as estimativas de μ_X calculadas em cada reamostra *bootstrap*.

O intervalo *bootstrap* studentizado representa uma alternativa capaz de produzir estimativas mais acuradas, sendo obtido da seguinte forma

$$\left(\bar{\bar{X}} - z^*_{((B+1)(1-\alpha))} \sqrt{Var_B(\bar{\bar{X}})}, \bar{\bar{X}} - z^*_{((B+1)\alpha} \sqrt{Var_B(\bar{\bar{X}})} \right) \tag{3.5}$$

sendo $Var_B(\bar{\bar{X}})$ calculado como em (3.3) e $z^*_{((B+1)(1-\alpha)}$ e $z^*_{((B+1)\alpha}$ os quantis $(1-\alpha)$ e α da distribuição empírica obtida a partir de $z_1^*, z_2^*, ..., z_B^*$, calculadas em cada uma das reamostras da seguinte forma

$$Z_b^* = \frac{\overline{\overline{X}}_b^* - \overline{\overline{X}}}{\sqrt{Var(\overline{\overline{X}}_b^*)}}, b = 1,...,B \tag{3.6}$$

com $Var(\overline{\overline{X}}_b^*)$ estimada através de um duplo *bootstrap* por

$$Var_B(\overline{\overline{X}}_b^*) = \frac{1}{R-1}\sum_{k=1}^{R}(\overline{\overline{X}}_k^{**} - \overline{\overline{X}}^{**})^2 \tag{3.7}$$

e $\overline{\overline{X}}^{**} = \frac{1}{R}\sum_{k=1}^{R}\overline{\overline{X}}_k^{**}$, sendo R o número de reamostras geradas a partir de cada uma das B reamostras *bootstrap* originais.

Alguns intervalos baseiam-se na distribuição empírica das estimativas calculadas em cada reamostra *bootstrap*. O intervalo de confiança percentil, por exemplo, tem os seguintes delimitantes

$$\left(\overline{\overline{X}}_{(B+1(\alpha))}^*, \overline{\overline{X}}_{((B+1)(1-\alpha))}^*\right) \tag{3.8}$$

onde $\overline{\overline{X}}_{((B+1)(1-\alpha)}^*$ e $\overline{\overline{X}}_{((B+1)(\alpha)}^*$ são os quantis $(1-\alpha)$ e α da distribuição empírica de $\overline{\overline{X}}^*$, calculada por meio das B estimativas obtidas.

Já os limites do intervalo de confiança *bootstrap* básico levam em consideração também o valor da estimativa calculado na amostra original

$$\left(2\overline{\overline{X}} - \left(\overline{\overline{X}}_{((B+1)(1-\alpha)}^*\right), 2\overline{\overline{X}} - \left(\overline{\overline{X}}_{((B+1)(\alpha)}^*\right)\right). \tag{3.9}$$

Finalmente, o método BCa (forma abreviada *para "bias corrected and accelerated"*) constitui uma versão aperfeiçoada do método percentil, produzindo intervalos de confiança com os seguintes limites

$$\left(\overline{\overline{X}}^{*(\alpha_1)}, \overline{\overline{X}}^{*(\alpha_2)}\right), \tag{3.10}$$

onde

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{a}(\hat{z}_0 + z^{(\alpha)})}\right), \qquad (3.11)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right)$$

sendo

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#(\overline{\overline{X}}_b^* < \overline{\overline{X}})}{B}\right) \qquad (3.12)$$

e $\hat{a}$ obtido via *jackknife* da seguinte forma

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\overline{\overline{X}}_{(\bullet)} - \overline{\overline{X}}_{(i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\overline{\overline{X}}_{(\bullet)} - \overline{\overline{X}}_{(i)}\right)^2\right]^{3/2}}. \qquad (3.13)$$

onde $\overline{\overline{X}}_{(i)}$ é a média dos elementos da amostra original excluindo o i-ésimo e $\overline{\overline{X}}_{(\bullet)}$ é a média dos $\overline{\overline{X}}_{(i)}$.

A implementação dos intervalos de confiança *bootstrap* para amostras por conjuntos ordenados imperfeitamente em *R* (Ihaka e Gentleman, 1996) encontram-se no Apêndice.

Sua aplicação na amostra por conjuntos ordenados selecionada na seção 2 (Tabela 1) produziu, respectivamente com 90%, 955 e 99% de confiança:

- *intervalo normal:* (72,68;80,07), (71,97;80,78) e (70,59;82,16), sendo o número de reamostras (B) igual a 1000;

- *intervalo percentil:* (72,62;80,12), (72,00,80,75) e (70,87;82,12), sendo B=1000;
- intervalo básico: (72,62;80,12), (72,00;80,75) e (70,62,81,87), sendo B=1000;
- intervalo BCa: (72,38;80,00), (71,87;80,56) e (70,72;82,00), sendo B=1000; e
- intervalo studentizado: (72,59;79,88), (71,49;80,99) e (69,50;82,27), sendo B=400, e R=200 no duplo *bootstrap*.

Considerações a respeito do desempenho desses intervalos de confiança *bootstrap* para conjuntos ordenados imperfeitamente encontram-se a seguir.

4. Avaliação do procedimento de reamostragem *bootstrap* em conjuntos ordenados imperfeitamente na estimação intervalar da média populacional

O procedimento de reamostragem *bootstrap*, proposto neste trabalho para intervalos de confiança para a média populacional quando do uso de amostras por conjuntos ordenados imperfeitamente, foi avaliado através de estudos por simulação. Neles foram consideradas cinco distribuições teóricas (normal, exponencial, uniforme, lognormal e Gumbell) e uma distribuição de referência, com dados fornecidos pela Companhia de Tecnologia de Saneamento Ambiental - CETESB, referentes às medições médias diárias de MP_{10} (material particulado maior que 10 micra) realizadas na estação Ibirapuera, na cidade de São Paulo, durante o biênio de 2000-2001.

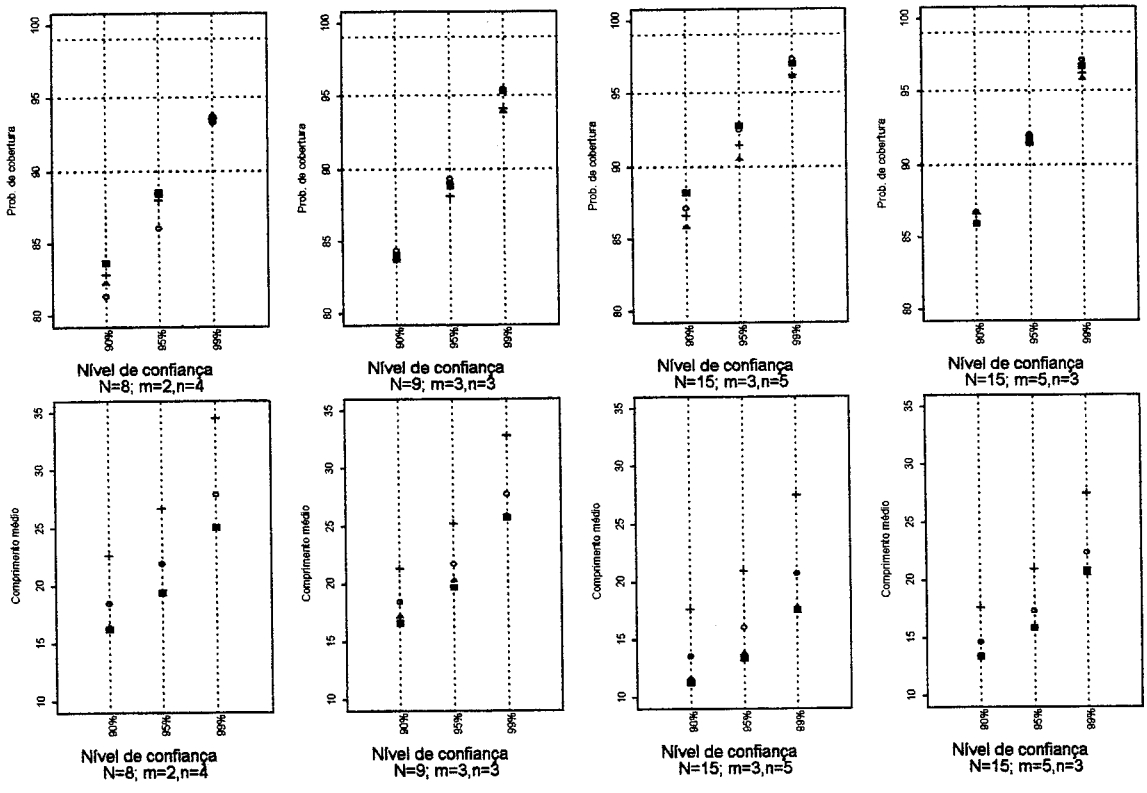
Também foram consideradas quatro combinações de tamanhos de amostra e números de replicações ($m=2,n=4$; $m=3,n=3$; $m=3,n=5$; $m=5,n=3$), além de dois valores para σ_ε^2, a variância dos erros de ordenação: $\sigma_\varepsilon^2=0{,}05$ (pequena incorreção) e $\sigma_\varepsilon^2=0{,}5$ (grau de imperfeição mais elevado), valores análogos aos utilizados em Dell e Cuttler(1972), no estudo da precisão do estimador $\overline{\overline{X}}$. A escolha dos tamanhos amostrais teve como critério a seleção de pequenas amostras (inferiores a 15), tomando valores moderados de m e n.

Os graus de imperfeição utilizados para a distribuição de referência são diferentes, devido à discrepância entre as grandezas destas observações em relação aos valores gerados a partir das distribuições teóricas. Buscou-se, entretanto, manter uma correspondência com a influência dos valores de σ_ε^2 adotados para as demais distribuições. O valor de σ_ε^2 responsável por uma maior imperfeição, neste caso, foi considerado $\sigma_\varepsilon^2 = \sigma_X^2 / 2 = 238{,}59$, enquanto $\sigma_\varepsilon^2 = \sigma_X^2 / 20 = 23{,}85$ foi escolhido para a situação em que os erros de ordenação têm menor interferência.

Para cada possível combinação dos tamanhos de amostras, graus de imperfeição e distribuição da variável de interesse foram extraídas K amostras por conjuntos imperfeitamente ordenados, responsáveis pela determinação de R intervalos de confiança *bootstrap* para a média populacional, com 90, 95 e 99% de confiança. Os intervalos foram avaliados de acordo com sua acurácia e precisão, baseado, respectivamente, na probabilidade de cobertura e no comprimento médio das estimativas geradas. Os valores atribuídos a K foram escolhidos através de um estudo relativo à convergência da probabilidade de cobertura enquanto a escolha de B e R se baseou na análise de convergência de $Var_B(\overline{\overline{X}})$, no caso dos intervalos *bootstrap* normal e studentizado, e dos quantis de interesse, para os demais intervalos de confiança. Os valores mínimos para os quais se verificou convergência foram $K=400$, $B=200$ e $R=200$. Foram adotados valores iguais ou superiores a estes em todas as simulações realizadas neste trabalho.

Os resultados alcançados para algumas das distribuições teóricas avaliadas são omitidos neste artigo, podendo, no entanto, serem verificados na íntegra em Taconeli (2005).

Figura 1- Probabilidades de cobertura e comprimentos médios de intervalos de confiança *bootstrap* percentil para a média das verificações médias diárias de MP10 na estação Ibirapuera (2000/2001), via SRS e RSS com ordenação perfeita e imperfeita

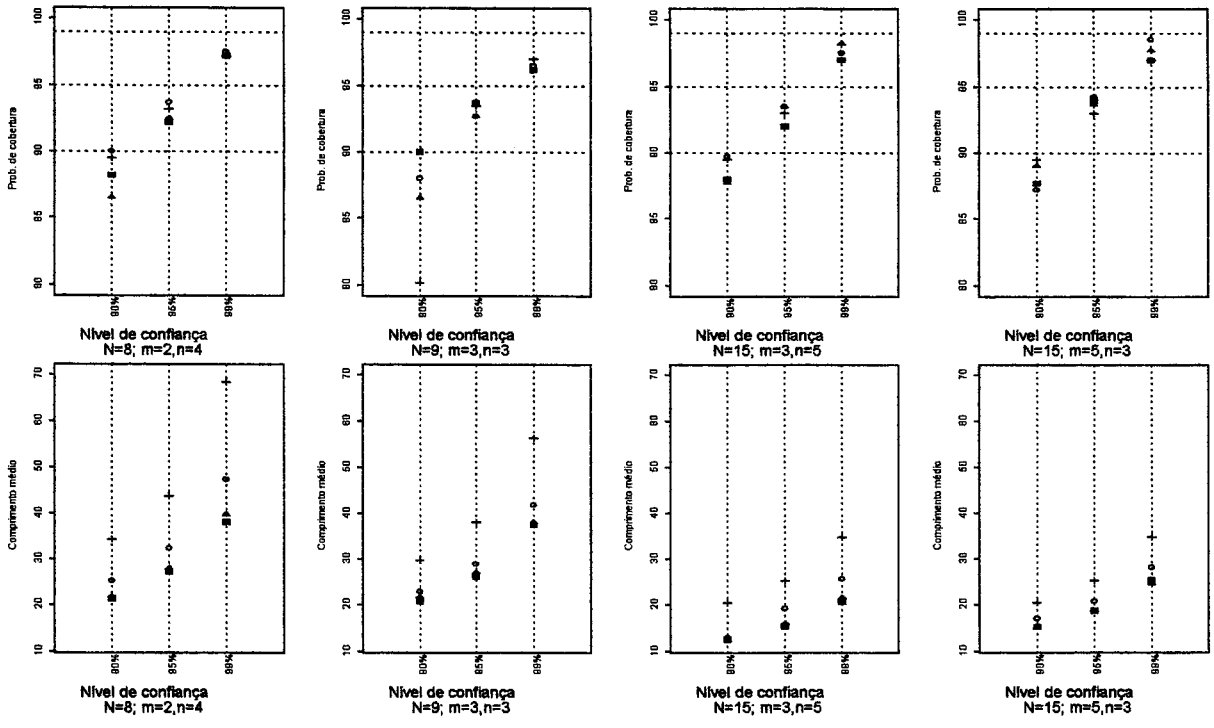


(+) *SRS* (■) *RSS* (Ordenação perfeita) (□) *RSS* (Ordenação imperfeita, $\sigma_\varepsilon^2 = 0{,}05$)

(○) *RSS* (Ordenação imperfeita, $\sigma_\varepsilon^2 = 0{,}50$)

As Figuras 1 e 2 permitem avaliar o desempenho do procedimento de reamostragem proposto na estimação intervalar da média de MP_{10}, através, respectivamente, de intervalos de confiança dos tipos percentil e studentizados, frente aos diferentes graus de imperfeição na ordenação das amostras. Verifica-se uma forte similaridade em relação à acurácia das estimativas concebidas via SRS e RSS com ordenação perfeita e imperfeita.

Figura 2- Probabilidades de cobertura e comprimentos médios de intervalos de confiança *bootstrap* studentizado para a média das verificações médias diárias de MP10 na estação Ibirapuera (2000/2001), via SRS e RSS com ordenação perfeita e imperfeita



Quanto à precisão dos intervalos de confiança, ambas as figuras apontam uma substancial redução dos comprimentos médios dos intervalos de confiança gerados via RSS, sobretudo ao se considerar ausência ou pequena influência dos erros de ordenação. Este ganho em precisão decresce à medida que aumenta a variância dos erros de ordenação, embora haja um significante ganho ainda na situação mais crítica, com o maior valor de σ_ε^2 considerado.

Tabela 2 - Probabilidade de cobertura de intervalos de confiança *bootstrap* para a média das medições médias diárias de MP10, obtida através de simulação, via amostragem aleatória simples (SRS) e amostragem por conjuntos ordenados (RSS), considerando ordenação perfeita (O.P.) e imperfeita (O.I.)

		N = 8; m = 2, n = 4				N = 9; m = 3, n = 3			
	Confiança	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS
Normal	90%	82,3	82,6	81,3	80,7	84,2	84,0	84,0	80,8
	95%	87,6	88,3	86,2	87,0	89,2	88,7	89,3	86,9
	99%	93,8	94,4	93,4	93,7	94,8	94,2	95,4	94,2
Básico	90%	83,0	82,8	81,7	81,4	83,7	83,9	84,0	83,4
	95%	88,0	87,9	86,1	86,1	89,1	88,2	88,9	88,1
	99%	94,7	93,3	92,1	92,7	95,0	93,5	94,3	93,7
Percentil	90%	83,6	82,2	81,3	82,8	84,0	83,6	84,3	83,7
	95%	88,5	88,3	86,0	88,0	88,9	88,7	89,3	88,1
	99%	93,5	93,8	93,3	93,6	95,3	93,9	95,4	94,1
BCa	90%	83,5	83,1	83,0	82,2	84,7	84,8	81,8	81,9
	95%	88,8	88,5	88,2	87,9	89,9	89,8	88,8	87,2
	99%	94,3	94,7	93,7	93,2	93,8	95,4	94,7	92,3
Studentizado	90%	88,2*	86,5	90,0*	89,5*	90,0*	86,5	88,0*	90,2*
	95%	92,2	92,5	93,7*	93,2*	93,7*	92,7	92,7*	93,5*
	99%	97,2	97,2	97,5	97,2	96,2	96,2	96,5	97,0

Tabela 2 - Probabilidade de cobertura de intervalos de confiança *bootstrap* para a média das medições médias diárias de MP10, obtida através de simulação, via amostragem aleatória simples (SRS) e amostragem por conjuntos ordenados (RSS), ordenação perfeita (O.P.) e imperfeita (O.I.) (*cont.*)

		N = 15; m = 3,n = 5				N = 15; m = 5,n = 3			
	Confiança	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS
Normal	90%	85,6	85,3	86,7	84,6	85,8	85,4	86,5	84,6
	95%	90,3	90,9	92,2	89,5	91,5	91,3	91,9	89,5
	99%	96,5	96,5	97,6	95,3	96,2	95,6	97,2	95,3
Básico	90%	85,4	84,9	86,4	85,1	87,2	85,6	86,6	85,1
	95%	91,9	90,3	91,3	89,4	91,4	90,7	91,4	89,4
	99%	96,1	95,5	96,9	94,3	96,0	95,1	96,6	94,3
Percentil	90%	88,2	85,8	87,1	86,6	85,9	86,0	86,7	86,6
	95%	92,8	90,5	92,5	91,5	91,8	91,4	92,0	91,5
	99%	97,0	96,1	97,3	96,2	96,7	95,8	97,1	96,2
BCa	90%	86,3	88,1	87,1	86,0	85,0	85,3	85,7	86,0
	95%	91,3	92,9	91,3	91,3	89,7	91,1	91,4	91,3
	99%	96,8	97,0	97,3	96,8	94,4	96,6	96,7	96,8
Studentizado	90%	88,0*	87,7*	89,7*	89,5*	87,7*	89,0*	87,2*	89,5*
	95%	92,0	93,5*	93,5*	93,0*	94,0*	93,7*	94,2*	93,0*
	99%	97,0	98,2*	97,5	97,0	97,0	97,7	98,5*	97,0

As Tabelas 2 e 3 apresentam os resultados de forma mais detalhada, permitindo verificar que o intervalo de confiança *bootstrap* do tipo studentizado é responsável pela geração de estimativas mais acuradas que os demais intervalos, que apresentam desempenhos bastante semelhantes em relação à acurácia e à precisão dos intervalos simulados. Na Tabela 2, valores acompanhados de um asterisco indicam que, para essa configuração, não se pode rejeitar a hipótese da probabilidade de cobertura obtida via simulação ser a inicialmente desejada (90, 95 ou 99%) a um nível de 95% de confiança. Tal hipótese é verificada através de um teste para proporções, baseado na distribuição binomial com parâmetros n e p, sendo n o número de intervalos de confiança simulados e p o nível de confiança considerado. Valores em negrito indicam que os intervalos gerados via RSS apresentam maior acurácia em relação aos intervalos obtidos via SRS, para aquela configuração de tamanho de amostra, número de réplicas e grau de imperfeição da ordenação.

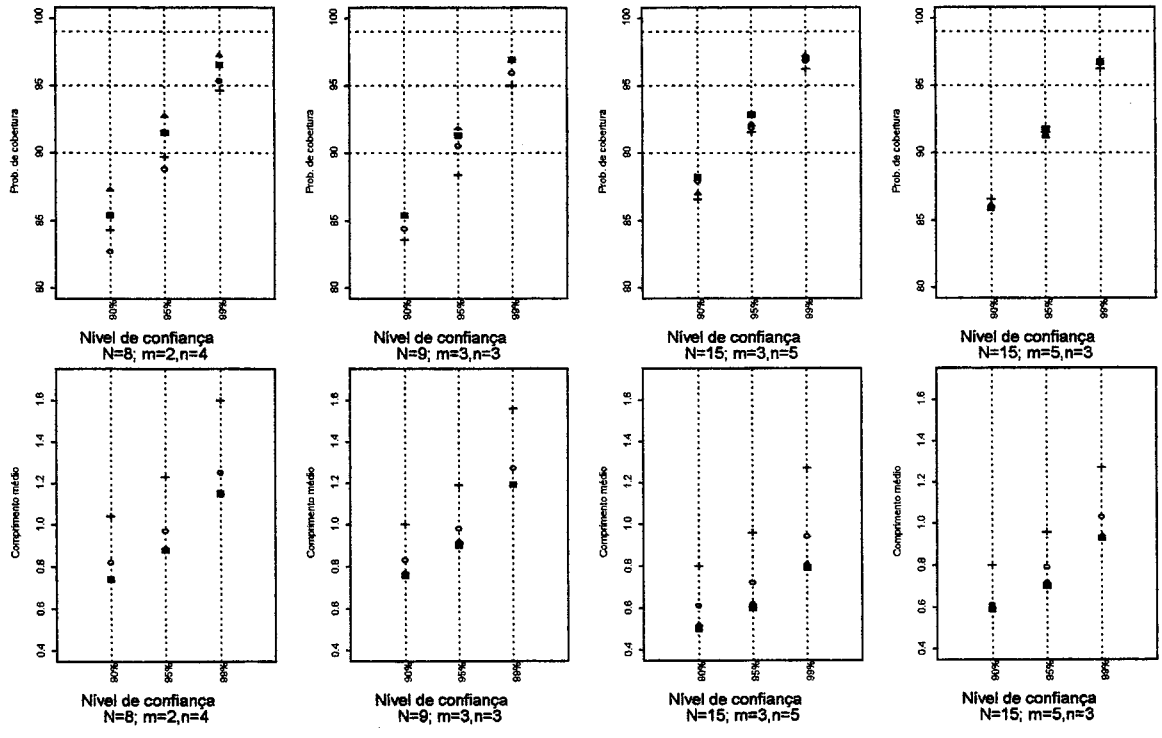
Tabela 3 - Comprimentos médios e desvio padrão dos comprimentos (entre parêntesis) de intervalos de confiança *bootstrap* para a média das medições médias diárias de MP10, obtidos através de simulação, via amostragem aleatória simples (SRS) e amostragem por conjuntos ordenados (RSS), ordenação perfeita (O.P.) e imperfeita (O.I.).

		N = 8; m = 2, n = 4				N = 9; m = 3, n = 3			
	Confiança	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS
Normal	90%	15,73 (5,19)	16,44 (5,14)	18,49 (5,92)	22,42 (7,22)	16,58 (5,00)	17,13 (5,39)	18,40 (5,22)	21,67 (6,76)
	95%	18,80 (6,20)	19,65 (6,14)	22,09 (7,07)	26,80 (8,63)	19,81 (5,98)	20,48 (6,45)	21,99 (6,25)	25,42 (8,08)
	99%	24,70 (8,15)	25,82 (8,07)	29,03 (9,29)	35,20 (11,35)	26,03 (7,85)	26,91 (8,47)	28,89 (8,21)	33,40 (10,62)
Básico	90%	16,21 (5,30)	16,40 (5,08)	18,44 (5,85)	22,49 (7,71)	16,70 (5,05)	17,15 (5,40)	18,41 (5,23)	21,54 (6,70)
	95%	19,31 (6,39)	19,45 (6,08)	21,87 (7,00)	26,53 (9,03)	19,77 (5,96)	20,25 (6,32)	21,72 (6,16)	25,45 (7,85)
	99%	24,81 (7,90)	24,90 (7,57)	27,85 (8,71)	33,96 (11,49)	25,50 (7,57)	25,88 (7,90)	27,71 (7,71)	32,71 (10,00)
Percentil	90%	16,19 (5,13)	16,40 (5,08)	18,44 (5,85)	22,60 (7,50)	16,59 (5,16)	17,15 (5,40)	18,41 (5,23)	21,32 (6,60)
	95%	19,37 (6,29)	19,45 (6,08)	21,87 (7,00)	26,67 (8,81)	19,68 (6,09)	20,25 (6,32)	21,72 (6,16)	25,23 (7,74)
	99%	25,08 (7,78)	24,90 (7,57)	27,85 (8,71)	34,54 (11,33)	25,71 (7,95)	25,88 (7,90)	27,71 (7,71)	32,78 (9,85)
BCa	90%	16,21 (5,68)	16,87 (5,54)	18,68 (5,75)	22,84 (7,96)	16,88 (5,43)	16,91 (5,30)	18,44 (5,86)	22,10 (7,44)
	95%	19,28 (6,70)	20,06 (6,50)	22,20 (6,75)	27,19 (9,58)	20,10 (6,45)	20,15 (6,34)	21,95 (6,54)	26,37 (8,93)
	99%	24,79 (8,47)	25,69 (8,18)	28,39 (8,73)	34,51 (11,87)	25,87 (8,26)	25,86 (7,98)	28,06 (8,18)	33,71 (11,44)
Studenti-zado	90%	21,30 (12,99)	21,75 (12,94)	25,23 (13,47)	34,30 (17,88)	20,76 (11,20)	21,74 (10,86)	22,90 (9,15)	29,71 (13,78)
	95%	27,20 (17,62)	27,98 (17,33)	32,28 (18,52)	43,77 (23,41)	26,04 (13,87)	27,07 (13,75)	28,83 (12,28)	38,04 (19,05)
	99%	38,08 (29,46)	39,77 (29,01)	47,35 (28,73)	68,38 (48,82)	37,46 (23,93)	38,06 (21,76)	41,81 (22,52)	56,33 (33,57)

Tabela 3 - Comprimentos médios e desvio padrão dos comprimentos (entre parêntesis) de intervalos de confiança *bootstrap* para a média das medições médias diárias de MP10, obtidos através de simulação, via amostragem aleatória simples (SRS) e amostragem por conjuntos ordenados (RSS), ordenação perfeita (O.P.) e imperfeita (O.I.) (*cont.*)

		N = 15; m = 3,n = 5				N = 15; m = 5,n = 3			
	Confiança	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS	RSS O.P.	RSS- O.I. ($\sigma_\varepsilon^2 = 23{,}85$)	RSS- O.I. ($\sigma_\varepsilon^2 = 238{,}59$)	SRS
Normal	90%	11,33 (2,82)	11,62 (2,63)	13,51 (2,67)	17,34 (4,24)	13,28 (3,21)	13,39 (3,22)	14,60 (3,20)	17,34 (4,24)
	95%	13,54 (3,37)	13,88 (3,14)	16,15 (3,19)	20,73 (5,07)	15,64 (3,83)	16,00 (3,85)	17,45 (3,82)	20,73 (5,07)
	99%	17,79 (4,43)	18,24 (4,13)	21,21 (4,19)	27,24 (6,67)	20,54 (5,04)	21,03 (5,07)	22,93 (5,02)	27,24 (6,67)
Básico	90%	11,45 (2,90)	11,62 (2,63)	13,51 (2,67)	17,59 (4,47)	13,27 (3,20)	13,39 (3,22)	14,60 (3,20)	17,59 (4,47)
	95%	13,61 (3,44)	13,76 (3,10)	16,03 (3,18)	20,87 (5,31)	15,75 (3,79)	15,88 (3,81)	17,29 (3,79)	20,87 (5,31)
	99%	17,72 (4,43)	17,75 (3,96)	20,69 (4,08)	27,10 (6,82)	20,47 (4,83)	20,50 (4,90)	22,35 (4,86)	27,10 (6,82)
Percentil	90%	11,23 (2,66)	11,62 (2,63)	13,51 (2,67)	17,60 (4,17)	13,34 (3,23)	13,39 (3,22)	14,60 (3,20)	17,60 (4,17)
	95%	13,37 (3,16)	13,76 (3,10)	16,03 (3,18)	20,93 (4,93)	15,84 (3,84)	15,88 (3,81)	17,29 (3,79)	20,93 (4,93)
	99%	17,56 (4,11)	17,75 (3,96)	20,69 (4,08)	27,45 (6,49)	20,82 (5,04)	20,50 (4,90)	22,35 (4,86)	27,45 (6,49)
BCa	90%	11,37 (2,85)	11,79 (2,75)	13,68 (2,91)	17,80 (4,57)	13,36 (3,68)	13,60 (3,40)	14,89 (3,47)	17,80 (4,57)
	95%	13,56 (3,40)	14,06 (3,28)	16,31 (3,50)	21,28 (4,59)	15,97 (4,42)	16,20 (4,07)	17,76 (4,16)	21,28 (4,59)
	99%	17,54 (4,37)	18,19 (4,21)	21,10 (4,59)	27,45 (7,04)	20,61 (5,70)	20,99 (5,26)	22,96 (5,40)	27,45 (7,04)
Studenti-zado	90%	12,55 (4,31)	13,17 (4,50)	15,63 (4,04)	20,42 (6,23)	15,31 (5,15)	14,98 (4,36)	17,01 (5,22)	20,42 (6,23)
	95%	15,39 (5,65)	16,00 (5,62)	19,16 (5,18)	25,18 (8,01)	18,69 (6,38)	18,32 (5,65)	20,72 (6,62)	25,18 (8,01)
	99%	20,65 (7,84)	21,55 (8,89)	25,56 (7,33)	34,72 (12,74)	25,28 (9,91)	24,52 (8,13)	28,09 (10,33)	34,72 (12,74)

Figura 3 - Probabilidades de cobertura e comprimentos médios de intervalos de confiança *bootstrap* percentil para a média de uma variável com distribuição normal, via SRS e RSS com ordenação perfeita e imperfeita

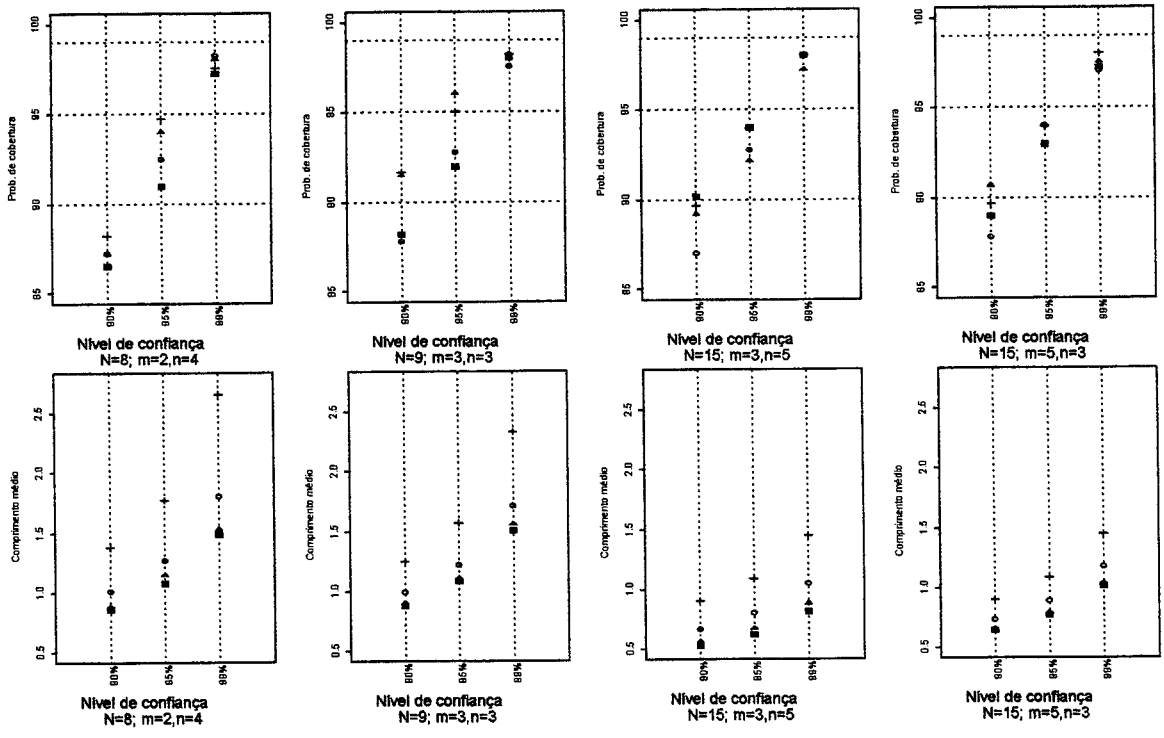


(+) *SRS* (■) *RSS* (Ordenação perfeita) (□) *RSS* (Ordenação imperfeita, $\sigma_\varepsilon^2 = 0{,}05$)

(○) *RSS* (Ordenação imperfeita, $\sigma_\varepsilon^2 = 0{,}50$)

Os resultados referentes à estimação da média de uma população com distribuição normal são apresentados nas Figuras 3 e 4. Mais uma vez verifica-se que o *bootstrap* baseado em conjuntos ordenados imperfeitamente possibilita a construção de intervalos de confiança mais precisos e tão acurados quanto àqueles originados via SRS, mesmo considerando erros de ordenação de maior magnitude. Novamente o intervalo de confiança *bootstrap* studentizado forneceu estimativas diferenciadas, mais acuradas que os demais.

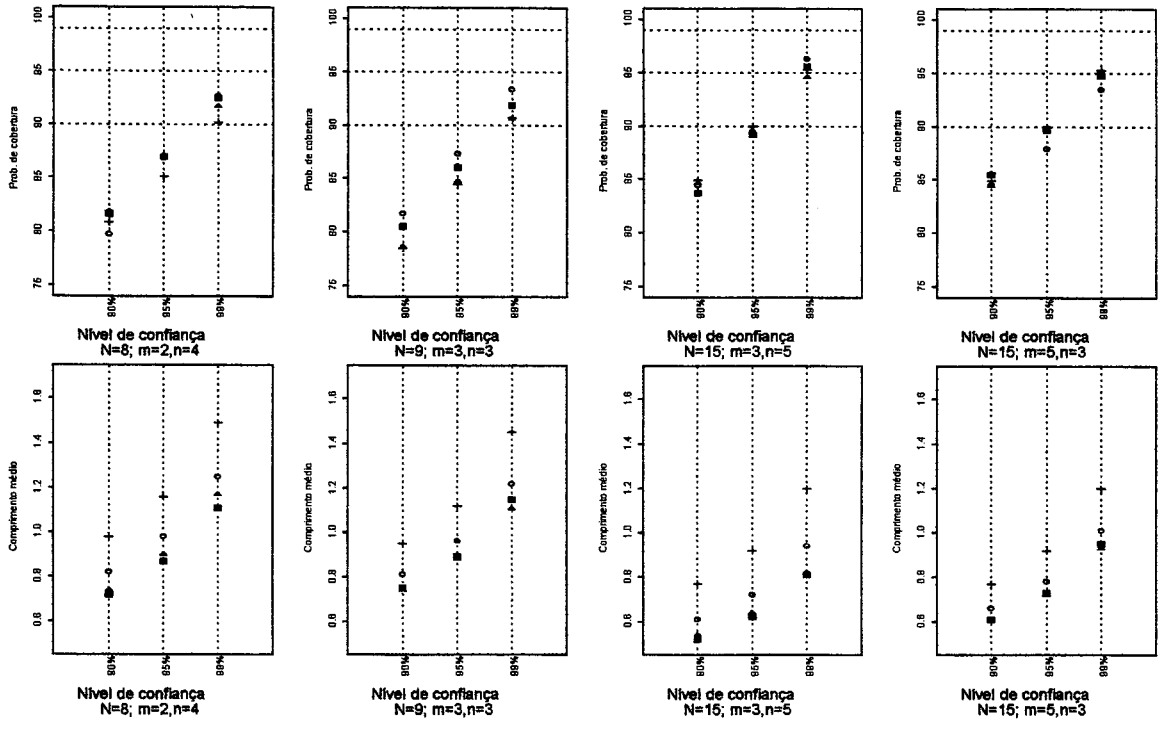
Figura 4 - Probabilidades de cobertura e comprimentos médios de intervalos de confiança *bootstrap* studentizado para a média de uma variável com distribuição normal, via SRS e RSS com ordenação perfeita e imperfeita



(+) ***SRS*** (■) ***RSS*** **(Ordenação perfeita)** (□) ***RSS*** **(Ordenação imperfeita,** $\sigma_\varepsilon^2 = 0{,}05$**)**

(○) ***RSS*** **(Ordenação imperfeita,** $\sigma_\varepsilon^2 = 0{,}50$**)**

Figura 5 - Probabilidades de cobertura e comprimentos médios de intervalos de confiança *bootstrap* percentil para a média de uma variável com distribuição exponencial, via SRS e RSS com ordenação perfeita e imperfeita

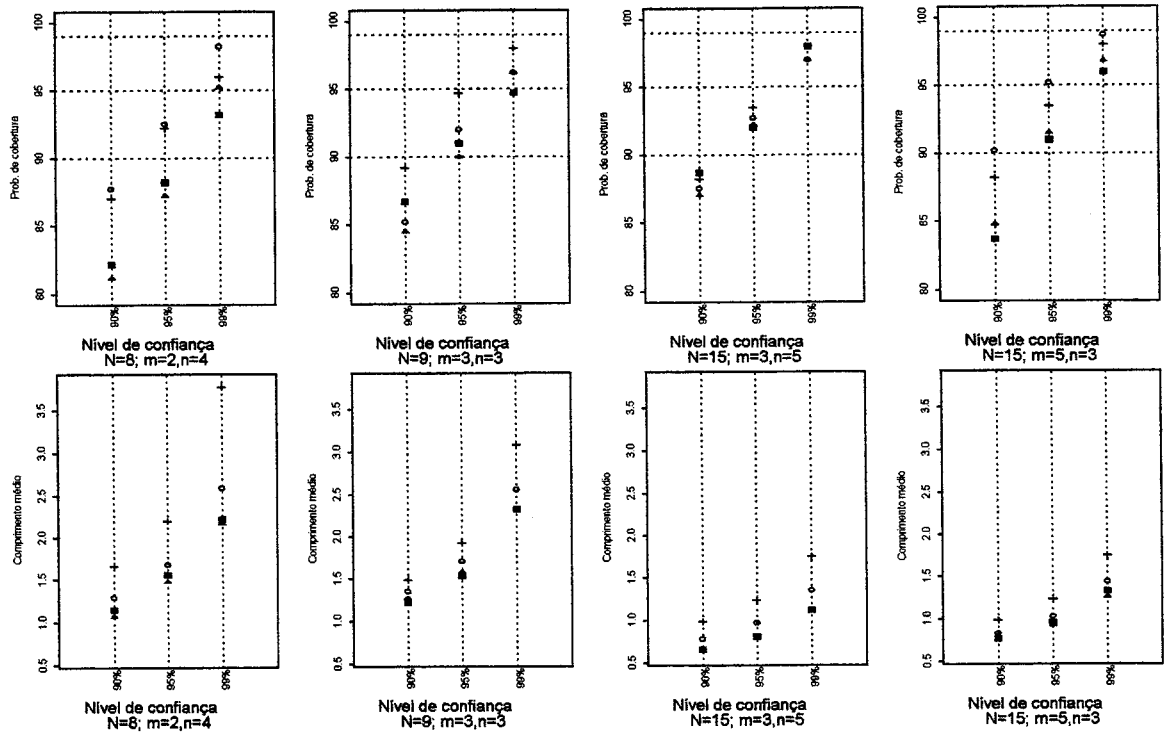


(+) *SRS* (■) *RSS* (Ordenação perfeita) (□) *RSS* (Ordenação imperfeita, $\sigma_\varepsilon^2 = 0{,}05$)

(○) RSS *(Ordenação imperfeita,* $\sigma_\varepsilon^2 = 0{,}50$)

As Figuras 5 e 6 ilustram os resultados alcançados na geração de intervalos de confiança para a média de uma distribuição exponencial. Tais resultados são semelhantes aos verificados para a distribuição de referencias de MP_{10}.

Figura 6 - Probabilidades de cobertura e comprimentos médios de intervalos de confiança *bootstrap* studentizado para a média de uma variável com distribuição exponencial, via SRS e RSS com ordenação perfeita e imperfeita



(+) ***SRS*** (■) ***RSS*** **(Ordenação perfeita)** (□) ***RSS*** **(Ordenação imperfeita,** $\sigma_\varepsilon^2 = 0{,}05$**)**

(○) ***RSS*** **(Ordenação imperfeita,** $\sigma_\varepsilon^2 = 0{,}50$**)**

Entre as distribuições estudadas, a distribuição lognormal ocasionou o maior distanciamento das probabilidades de cobertura simulada e teórica, tanto em relação aos intervalos propostos, como para aqueles baseados em amostragem aleatória simples.

5. Conclusão

Neste artigo apresentou-se o método de reamostragem *bootstrap* em conjuntos ordenados imperfeitamente proposto por Taconeli (2005) com o objetivo de se obter estimativas intervalares da média populacional. A partir de uma população de árvores de cerejas pretas, obteve-se uma amostra de conjuntos ordenados com a finalidade de se estimar a altura média, usando como critério de ordenação o diâmetro na altura do peito, onde foram calculas os intervalos de confiança percentil, normal, básico, BCa e

studentizado usando o método de reamostragem proposto. No estudo por simulação do desempenho desses intervalos, verificou-se que eles produzem estimação mais precisa, sem perda de acurácia, em relação à gerada pelo procedimento de reamostragem *bootstrap* tradicional, o que foi verificado para diferentes distribuições, tamanhos de amostras e graus de imperfeição no processo de ordenação. Tal resultado recomenda a utilização do método proposto em detrimento ao *bootstrap* tradicional, baseado em SRS, sobretudo quando as amostras a serem extraídas têm tamanho reduzido, ainda que o critério adotado para ordenação das amostras seja suscetível a falhas.

Referências bibliográficas

BRANDÃO, J.S.; Intervalos de confiança assintóticos para o parâmetro de locação em distribuições assimétricas na família locação-escala usando amostragem em conjuntos ordenados. ***Relatório de Iniciação Científica***. Des – Universidade Federal de São Carlos.

CANTY, A.;The bootstrap and confidence intervals. Disponível em ***www.mathstat.concordia.ca/canty/teaching/mast679t.html***. Acesso em 25 agosto 2004.

CESÁRIO, L.C.; BARRETO, M. C. M. (2003). Um estudo sobre o desempenho de intervalos de confiança bootstrap para a média de uma distribuição normal usando amostragem por conjuntos ordenados perfeitamente. ***Rev. Mat. Estat.***, 21, p. 7-20.

DAVISON, A.C.; HINKLEY, D.V.; ***Bootstrap methods and their application***. New York: Cambridge University Press, 1997. 582p.

DELL, T.R.; CUTTLER, J.L. (1972). Ranked set sampling theory with order statistics background. ***Biometrics***, 28, p.545-555.

IHAKA, R.;GENTLEMAN, R.; A language for data analysis and graphics. ***Journal of Computational and Graphics Statistics***, 5, p.299-314, 1996.

MCINTYRE, G.A.; A method for unbiased seletive sampling, using ranked sets. ***Australian Journal of Agricultural Research***, 3, p.385-390. 1952.

NAHHAS, R.W., WOLFE, D.A., CHEN, H.; Ranked set sampling: cost and optimal set size. ***Biometrics***, 58, 964-971, 2002.

RYAN, T. A., JOINER, B. L., RYAN, B. F. (1976). ***The Minitab Student Handbook***. Duxbury Press.

STOKES, S.L., SAGER, T.W. (1998). Characterization of a ranked-set sample with application to estimating distribution functions. ***J. Am. Stat. Assoc.***, 83, p. 374-381.

TACONELI, C.A. (2005). Reamostragem bootstrap em amostragem por conjuntos ordenados e intervalos de confiança não paramétricos para a média. 120f. ***Dissertação (Mestrado em Estatística)*** – Universidade Federal de São Carlos, São Carlos.

TACONELI, C.A., BARRETO, M.C.M. (2003). Intervalos de confiança para a média populacional usando amostragem em conjuntos ordenados. ***Rev. Mat. Estat.***, 21, p. 41-66.

TACONELI, C.A., BARRETO, M.C.M. (2005). Avaliação de uma proposta de intervalos de confiança bootstrap em Amostragem por Conjuntos Ordenados, . *Rev. Mat. Estat.*, 23, n.3, p. 33-53.

TAKAHASHI, K., WAKIMOTO, K. (1968). On biased estimates of population mean based on sample stratified by means of ordering. *Ann. Inst. Stat. Math.*, v.20, p.1-31.

Abstract

The design of ranked set samples is more efficient than single random samples in the estimation of several parameters, mainly for small samples. The aim of this work is to present the extension of the methodology of nonparametric bootstrap resampling introducing by Taconeli (2005) when the original sampling is drawn by imperfect ranked set design, with application to percentil, basic, normal, BCa and studentized confidence intervals for population mean. The simulation study has investigated the behavior of these bootstrap intervals under six distinct populations, four sample size and replications. The results point out the best precision of bootstrap confidence intervals in imperfect ranked set design over single random and between the different bootstrap confidence intervals. The studentized intervals are most accurate. The procedure of bootstrap resampling is applying in the estimation of the height of back cherry trees using the diameter of high of breast as rank variable.

Keywords: Ranked set sampling; imperfect ranked; confidence intervals for population mean; bootstrap resampling methods.

Agradecimentos

Agradecimentos – à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pela concessão de uma bolsa de mestrado.

Aos pareceristas da Revista Brasileira de Estatística, que avaliaram e contribuíram para o aperfeiçoamento do presente artigo.

APÊNDICE

```
# Função RSSBCa
# Este programa constrói intervalos de confiança BCa para a média em delineamentos por
# conjuntos imperfeitamente ordenados conforme proposto em Barreto e Taconeli (2006).
# Argumentos obrigatórios:
# Amostra: vetor contendo a amostra por conjuntos ordenados.
# m: é o número de réplicas do delineamento.

# n: é o tamanho das amostras do delineamento.
# Argumentos opcionais:
# Ranks: vetor contendo os ranks resultantes da ordenação da amostra por conjuntos
# ordenados através de julgamento visual ou contendo os valores associados da variável
```

```
# concomitante. Se não informados, considera-se ordenação perfeita.
# B: é o número de reamostras bootstrap (default: B=1000).
# alfa: é o nível de significância do intervalo (default: alfa=0.05).
#############################################################################
RSSBCa<-function(amostra,m,n,ranks,B,alfa){
if (missing(ranks))
ranks<-seq(1:m*n)
if (missing(B))
B<-1000
if (missing(alfa))
alfa<-0.05
vetmed<-numeric()
jack<-numeric()
amostra<-amostra[order(ranks)]
    for (p in 1:(m*n)){vetmed[p]<-mean(amostra[-p])}
a<-(1/6)*sum((mean(vetmed)-vetmed)^3)/((sum((mean(vetmed)-vetmed)^2))^1.5)
contador<-0
med<-numeric()
for(u in 1:B){reamostra<-numeric()
    for (i in 1:m){
    for (j in 1:n){  d<-sample(seq(1:(m*n)),n,replace=T)
                reamostra<-c(reamostra,amostra[(sort(d))[j]])}}
    med[u]<-mean(reamostra)
    if(med[u]<=mean(amostra))
    contador<-contador+1}
w<-qnorm(contador/B)
zalfa1<-qnorm(alfa/2)+w
zalfa2<-qnorm(1-alfa/2)+w
alfatio1<-pnorm(w+(zalfa1)/(1-a*(zalfa1)))
alfatio2<-pnorm(w+(zalfa2)/(1-a*(zalfa2)))
tetamin<-quantile(med,alfatio1)
tetamax<-quantile(med,alfatio2)
cat("Intervalo de confiança bootstrap BCa (", 100*(1-alfa),"%) para a média:
    (",tetamin,",",tetamax,")","\n")}
#############################################################################
# Para os dados das árvores
diâmetro<-c(8.3,12.9,17.9,18.0,8.8,13.7,13.3,20.6)
altura<-c(70,74,80,80,63,71,86,87)
exemplo1<-RSSBCa(amostra=altura,m=2,n=4,ranks=diâmetro)
#############################################################################
# Função ICBRSS
# Este programa constrói intervalos de confiança bootstrap normal, percentil e básico para a


# média em delineamentos por conjuntos imperfeitamente ordenados, conforme proposto
# em Barreto e Taconeli (2006).
# Argumentos obrigatórios:
# Amostra: vetor contendo a amostra por conjuntos ordenados.
# m: é o número de réplicas do delineamento.
# n: é o tamanho das amostras do delineamento.
# Argumentos opcionais:
# Ranks: vetor contendo os ranks resultantes da ordenação da amostra por conjuntos
# ordenados através de julgamento visual ou contendo os valores associados da variável
# concomitante. Se não informados, considera-se ordenação perfeita.
# B: é o número de reamostras bootstrap (default: B=1000).
# alfa: é o nível de significância do intervalo (default: alfa=0.05).
#############################################################################
```

```
ICBRSS<-function(amostra,m,n,ranks,B,alfa){
if (missing(ranks))
ranks<-seq(1:m*n)
if (missing(B))
B<-1000
if (missing(alfa))
alfa<-0.05
amostra<-amostra[order(ranks)]
teta.star<-rep(0,B)
for (i in 1:B){ reamostra<-numeric()
for (t in 1:m){ for (p in 1:n){      d<-sample(seq(1:(m*n)),n,replace=T)
                                reamostra<-c(reamostra,amostra[(sort(d))[p]])}}
teta.star[i]<-mean(reamostra)}
teta.star<-sort(teta.star)
var.boot<-sum((teta.star-mean(teta.star))^2)*1/(B-1)
lim.infbnormal<-mean(amostra)-qnorm(1-alfa/2)*sqrt(var.boot)
lim.supbnormal<-mean(amostra)+qnorm(1-alfa/2)*sqrt(var.boot)
lim.infpercent<-teta.star[(B+1)*alfa/2]
lim.suppercent<-teta.star[(B+1)*(1-alfa/2)]
lim.infbasic<-(2*mean(amostra)-(teta.star[(B+1)*(1-alfa/2)]))
lim.supbasic<-(2*mean(amostra)-(teta.star[(B+1)*alfa/2]))
cat("Intervalo de confiança bootstrap normal (", 100*(1-alfa),"%) para a média:
   (",lim.infbnormal,",",lim.supbnormal,")","\n")
cat("Intervalo de confiança bootstrap percentil (", 100*(1-alfa),"%) para a média:
   (",lim.infpercent,",",lim.suppercent,")","\n")
cat("Intervalo de confiança bootstrap básico (", 100*(1-alfa),"%) para a média:
   (",lim.infbasic,",",lim.supbasic,")","\n")}
exemplo2<-ICBRSS(amostra=altura,m=2,n=4,ranks=diâmetro)
##########################################################################
# Função RSSt
# Este programa constrói intervalos de confiança t-bootstrap para a média em
# delineamentos por conjuntos imperfeitamente ordenados, conforme proposto em Barreto #e Taconeli
   (2006).
# Argumentos obrigatórios:
# Amostra: vetor contendo a amostra por conjuntos ordenados.
# m: é o número de réplicas do delineamento.
# n: é o tamanho das amostras do delineamento.
# Argumentos opcionais:
# Ranks: vetor contendo os ranks resultantes da ordenação da amostra por conjuntos
# ordenados através de julgamento visual ou contendo os valores associados da variável
# concomitante. Se não informados, considera-se ordenação perfeita.
# B: é o número de reamostras bootstrap (default: B=500).
# R: é o número de reamostras para o duplo bootstrap (default: B=200).
# alfa: é o nível de significância do intervalo (default: alfa=0.05).
##########################################################################

RSSt<-function(amostra,m,n,ranks,B,R,alfa){
if (missing(ranks))
ranks<-seq(1:m*n)
if (missing(B))
B<-500
if (missing(R))
R<-200
if (missing(alfa))
alfa<-0.05
amostra<-amostra[order(ranks)]
teta.chap<-numeric()
```

```
z.star<-numeric()          #variância da amostra.
teta.star<-numeric()       #teta.star vai armazenar os estimadores prov. das amostras bootstrap.
for (p in 1:B){ reamostra<-numeric()
        for (i in 1:m){ for (j in 1:n){ d<-sample(seq(1:(m*n)),n,replace=T)
                reamostra<-c(reamostra,amostra[(sort(d))[j]]) }}
teta.chap<-c(teta.chap,mean(reamostra))
teta.chapest<-numeric()
for (k in 1:R){ reamostra2<-numeric()
        for (i in 1:m){ for (j in 1:n){ d<-sample(seq(1:(m*n)),n,replace=T)
                reamostra2<-c(reamostra2,reamostra[(sort(d))[j]]) }}
teta.chapest<-c(teta.chapest,mean(reamostra2))}
v.star<-(1/(R-1))*sum((teta.chapest-mean(teta.chapest))^2)
z.star[p]<-(teta.chap[p]-mean(amostra))/sqrt(v.star) }
z.star<-sort(z.star)
vezinho<-(1/(B-1))*sum((teta.chap-mean(teta.chap))^2)
lim.inft<-((mean(amostra)-quantile(z.star,1-alfa)*sqrt(vezinho)))
lim.supt<-((mean(amostra)-quantile(z.star,alfa)*sqrt(vezinho)))
cat("Intervalo de confiança t-bootstrap (", 100*(1-alfa),"%) para a média: (",lim.inft,",",lim.supt,")","\n")
exemplo3<-RSSt(amostra=altura,m=2,n=4,ranks=diâmetro)
```

Análise de correspondência múltipla e análise de agrupamento na redução de dimensionalidade de indicadores de eventos de vida*

Raquel de Vasconcellos Carvalhaes de Oliveira[*]
Anderson Martins Silva[**]
Simone Gonçalves de Assis[***]
Nilton César dos Santos[****]

* A pesquisa recebeu apoio financeiro do UNICEF, CNPq e FAPERJ.

* Endereços para correspondências: Estatístico.Pesquisadora colaboradora do Centro Latino Americano de Estudos de Violência e Saúde Jorge Careli – CLAVES. Escola Nacional de Saúde Pública. Fundação Oswaldo Cruz. Mestre em Estudos Populacionais e Pesquisas Sociais ENCE – IBGE, raquel.carvalhaes@br.inter.net.

** Graduação em Estatística. Bolsista de Iniciação Científica CNPq no CLAVES/ENSP/FIOCRUZ.

*** Pesquisadora Titular do Departamento de Epidemiologia e Métodos Quantitativos em Saúde e do CLAVES/ENSP/FIOCRUZ.

**** Estatístico. Eletrobrás. Mestre em Estudos Populacionais e Pesquisas Sociais ENCE – IBGE.

Resumo

A resiliência tem se revelado como um conceito importante na área de conhecimento da saúde, especialmente no campo infanto-juvenil, sendo um conceito importante, consistindo principalmente na superação dos traumas vivenciados pelo indivíduo. Eventos de vida são situações traumáticas que podem acontecer na vida nos âmbitos: familiar, escolar, entre outros. O estudo contemplou 1 923 alunos de 11 a 19 anos da cidade de São Gonçalo (RJ), em 2003. Através da Análise de Correspondência Múltipla procura-se construir dimensões para os eventos de vida e ulteriormente relacioná-las à escala de Resiliência. Nesse estudo, pretende-se definir mais precisas dimensões de eventos de vida através da Análise de Correspondência nas 29 questões, utilizando-se também a Análise de Agrupamento para corroborar com a interpretação das plotagens da Análise de Correspondência bidimensionais e/ou tridimensionais. Nas dimensões construídas, constatou-se que a ausência de eventos relaciona-se ao comportamento resiliente, todavia a ocorrência de traumas não se relaciona àquele não-resiliente.

Palavras-chave: Análise de correspondência múltipla, análise de agrupamento, resiliência, evento de ida, psicometria.

1. Introdução

Esse artigo analisa dados inéditos, provenientes de uma pesquisa mais ampla sobre resiliência realizada com adolescentes (Assis *et al.*, 2006). O conceito da resiliência foi introduzido há poucas décadas na área da psicologia, sendo posteriormente inserido em outras áreas, como saúde pública e educação. Significa a capacidade de resistir às adversidades, a força necessária para uma pessoa possuir saúde mental durante sua vida, mesmo após a exposição a riscos. É fruto de um conjunto de processos sociais e intrapsíquicos que possibilitam o desenvolvimento de uma vida sadia, mesmo vivendo em um ambiente não sadio (Rutter, 1981; Tavares, 2001).

Viver eventos adversos ou estressantes é uma parte do processo de se tornar uma pessoa resiliente (Garmezy; Rutter, 1988). Outra parte do processo é a existência de fatores protetores que "tamponariam" os efeitos dos eventos de vida. Nesse dinâmico interagir, adversidade e proteção – atributos essencialmente proporcionados pelo meio

social - são administradas por cada indivíduo de forma a se tornar mais ou menos resiliente.

Este artigo tem como objetivo classificar, a partir da análise de correspondência múltipla e de agrupamento, a resiliência a partir de técnicas estatísticas que possibilitem explorar relações simultâneas com os eventos adversos ocorridos na dimensão dos pares, da escola e da comunidade. A principal justificativa da escolha dessas técnicas se dá pelo fato de que outras técnicas (associação e correlação) utilizadas preliminarmente não encontraram associação estatística significante entre resiliência e a ocorrência de todos esses eventos.

2. Métodos

2.1. População e amostra

A população foi constituída por 1 923 escolares de 11 a 19 anos de escolas pública e particular de 7ª e 8ª séries do Ensino Fundamental e 1º e 2º ano do Ensino Médio no Município de São Gonçalo-RJ, no ano de 2003. A escolha do município estudado deve-se ao fato deste apresentar altos índices de violência no ano de 2000, sendo o vigésimo terceiro colocado no *ranking* estadual, baseado em taxas de homicídios calculadas com os dados do DATASUS.

A população foi dividida em quatro estratos segundo o grau de escolaridade e natureza da instituição, tais como: Estrato 1- 7ª/8ª séries da Escola Pública; Estrato 2 - 7ª /8ª séries da Escola Particular; Estrato 3 - 1º /2º ano da Escola Pública; Estrato 4- 1º /2º ano da Escola Particular, segundo dados fornecidos pela Secretaria Municipal de Educação. Após a estratificação, o plano amostral utilizado foi o de amostragem conglomerada em dois estágios, onde a seleção de unidades primárias (escolas) baseou-se na Amostragem Seqüencial de Poisson (Ohlsson, 1998), com a finalidade de produzir uma maior sobreposição entre as escolas da amostra. Na etapa final de seleção, utilizou-se a amostra aleatória simples das turmas (unidade secundária de amostragem) dentro de cada unidade primária, exceto no caso da escola possuir uma única turma,

totalizando 71 turmas amostradas. Assim, o procedimento de seleção passou por duas etapas: a) seleção das escolas (representantes do conjunto de turmas e alunos dos estratos selecionados); b) seleção das turmas por amostragem aleatória simples dentro das escolas ou, em alguns casos, pesquisa da única turma da escola. O principal motivo da escolha do desenho de amostra é a constante dificuldade da obtenção de um cadastro de identificação de alunos, e também pela falta de viabilidade prática na execução de uma seleção aleatória de alunos. O efeito do plano amostral influencia a estimação das variâncias, e por conseqüência, algumas estatísticas de teste, como a estatística de *Wald* numa regressão. Entretanto, o efeito do plano amostral não foi levado em consideração nas duas técnicas multivariadas, como correspondência múltipla e agrupamento, empregadas no presente estudo.

O instrumento de avaliação compõe-se de diversas escalas e indicadores psicométricos. Um estudo piloto foi aplicado a 203 respondentes num intervalo de 7 a 10 dias entre as aplicações com a finalidade de verificar a confiabilidade e a validade.

2.2. Escala de resiliência

A escala de Resiliência desenvolvida por Wagnild e Young (1993) é uma escala constituída por 25 itens com sete opções de resposta: Discordo Totalmente, Discordo Muito, Discordo Pouco, Nem Concordo Nem Discordo, Concordo Pouco, Concordo Muito e Concordo Totalmente, onde todas as perguntas têm um sentido positivo na mensuração do conceito de resiliência. A análise desta escala é feita através da soma dos escores de resposta de um dado indivíduo, onde altos escores totais indicam um indivíduo resiliente (tipo Guttman).

A escala utilizada no estudo passou pela adaptação transcultural descrita por Herdman *et al.* (1998). Bons índices psicométricos foram obtidos no presente estudo: alpha de Cronbach - 0,838, Coeficiente de correlação intraclasse: ICC - 0,75 e kappa predominantemente moderado. Utilizou-se como ponto de corte para definir os menos resilientes um desvio padrão abaixo da média aritmética.

2.3. Indicador de eventos de vida ocorridos na escola, comunidade, amigos e namorados

Constitui-se a partir da concatenação de indicadores consolidados com a finalidade de estudar "eventos estressantes" ocorridos durante a vida da criança e adolescente nas esferas da escola, comunidade, amigos e namorados. As respostas são dicotômicas com o objetivo de estudar a presença e ausência de potenciais eventos de vida. Uma das escalas utilizadas para a formulação das questões foi a formulada por Pitzner e Drummond (1997) para Eventos Negativos de Vida da qual se retiraram as questões que "correspondem" às dimensões Amigos e Namorado e Comunidade e as demais questões foram aquelas propostas no trabalho de Trombeta *et al.* (2002).

As questões que fizeram parte do estudo são:

Subescala escola: mudar-se muitas vezes de escola; envolver-se em confusão na escola, indo para a direção ou sendo suspenso das aulas; ter conflitos sérios com professor;

Subescala comunidade: já ter visto alguém gravemente ferido; viver situação de perigo e insegurança na vizinhança; e

Subescala amigos/namorados: separar-se de amigo próximo por brigas, mudança de residência ou morte; sofrer por término de namoro.

2.4. Violência na escola e na localidade

As escalas de violência na escola e na localidade foram utilizadas originalmente pelo Instituto Latino-Americano das Nações Unidas para a Prevenção de Delito e Tratamento do Delinqüente (ILANUD/ONU) sendo aplicadas a escolares do ensino público da cidade de São Paulo (ILANUD, 2000). O questionário procura investigar a vitimização sofrida pelo adolescente no âmbito escolar e da comunidade ocorrida no último ano através de oito perguntas dicotômicas para cada uma das áreas. As perguntas incluem ter sido: humilhado, ameaçado, agredido fortemente (a ponto de necessitar cuidados médicos); ter tido seus objetos danificados propositalmente, conviver com portadores de arma branca e de fogo; ter sido vítima de furto e roubo. Os pontos de corte utilizados para definir os graus de vitimização na escola e na

comunidade foram os *tercis* que definiram os pontos abaixo de 33,33% como baixa violência, entre 33,33% e 66,67% como média ocorrência de violência e acima de 66,67% como vítima de alta violência.

2.5. Processamento e análise de dados

A máscara de entrada de dados foi criada no Epi-info 6.0, com a posterior construção do banco de dados no SPSS 8.0, atendendo quatro rigorosas etapas durante seu processamento: codificação, digitação, correção e análise. Com a finalidade de reduzir a escala de evento de vida em dimensões interpretáveis, utilizou-se a análise de correspondência e a análise de agrupamento através do *software* XLSTAT 4.0.

Primeiramente, testou-se isoladamente a análise do indicador de eventos de vida. Cada dimensão (comunidade, amigos e namorados, escola) foi analisada através de uma Análise de Correspondência Múltipla independente a fim de verificar o padrão de constituição da dimensão do Indicador de Eventos de Vida. Após isso, foram retiradas variáveis que não apresentavam um padrão de agregação com as demais, configurando uma observação discrepante (*outlier),* pelo baixo padrão de respostas ou por não abarcar o conceito de eventos estressantes naquele contexto de ocorrência (gostar de alguém sem ser correspondido; assalto/roubo da residência; sentir-se mal pelo mau desempenho escolar; já ter fracassado em competição). Posteriormente, incluiu-se a variável resiliência na análise, visando reduzir a dimensionalidade do indicador de eventos de vida.

A **análise de correspondência múltipla** é uma técnica multivariada, descritiva ou exploratória que examina relações geométricas do cruzamento das variáveis categóricas que são utilizadas na associação de grupos, entre categorias das linhas com as categorias das colunas, através de dimensões ou eixos ortogonais. A técnica se baseia na observação e critério de interpretação do pesquisador em relação à plotagem destas distâncias geométricas, onde cada ponto representa uma das categorias das variáveis analisadas. A principal vantagem da técnica é a isenção de conhecimento prévio da distribuição dos dados, dispensando qualquer ajustamento *à priori* ou tratamento estatístico. Outra de suas vantagens é a possibilidade de trabalhar com grande número

de dados facilitando e resumindo sua interpretação em um único gráfico (ou mapa perceptual), entretanto ao se utilizar poucos dados deve-se preferir os modelos log-lineares (Clausen, 1988). Uma de suas desvantagens é a dificuldade da realização do teste de hipóteses pelo caráter descritivo da técnica, além de geralmente seus achados não permitirem generalização. Logo, as conclusões são válidas somente para aquele grupo estudado tal como qualquer outro estudo de natureza qualitativo, advindo o crescente interesse desses pesquisadores por essa técnica. (Aranha *et al.*, 2004; Bouvier *et al.*, 1999, Tenenhaus *et al.*, 1985, Young, 1984).

Alguns conceitos de análise de correspondência são essenciais para o entendimento dos resultados e para a correta interpretação. Todos os termos constantemente utilizados ao se dissertar sobre a técnica se baseiam numa tabela de *crosstabs*, onde os perfis das linhas e colunas são somente a representação absoluta das freqüências relativas com o total da base linha ou coluna, e as respectivas freqüências marginais obtidas nessa tabela são as massas das linhas e colunas. As massas são as marginais das variáveis em estudo, ao se tomar a linha como base e as massas seriam os perfis das linhas.

O conceito mais citado na bibliografia é o de inércia que nada mais é do que a proporção explicada pelo autovalor em relação ao traço total, ou seja, é a proporção do total de inércia explicada naquela correspondente dimensão (Bendixen, 1996). A análise de correspondência se baseia nas distâncias qui-quadrado obtidas através das distâncias euclidianas, portanto esta funciona como uma decomposição do qui-quadrado através das categorias e dimensões analisadas (Clausen, 1988; Salvador Figueras, 2003).

Um dos conceitos importantes para a interpretação da análise de correspondência é o de contribuições absoluta e relativa. A contribuição absoluta mede a construção de cada eixo através das variáveis categóricas, ou seja, verifica o quanto cada categoria da variável contribuiu para a inércia daquela dimensão, com isso há uma melhor interpretação dos eixos (Greenacre, 1981). Na contribuição relativa há uma melhor explicação da variabilidade categórica que está sendo compreendida através do eixo, portanto, a contribuição relativa explica a quantidade da inércia do ponto que é descrita por uma dita dimensão (Salvador Figueras, 2003). Entretanto, existe uma implicação

direta entre as contribuições absolutas com as contribuições relativas, conquanto o contrário nem sempre ocorre (Greenacre, 1981).

Uma variante da análise de correspondência múltipla, chamada de análise de correspondência múltipla ajustada ou conjunta, pode ser utilizada ignorando as submatrizes da diagonal da matriz de Burt e, por conseguinte, não inflacionando a distância qui-quadrado entre os perfis e a inércia total. Neste método é sugerida a aplicação de um algoritmo de mínimos quadrados, onde as matrizes diagonais de Burt são tratadas como dados faltantes (Greenacre, 2005). Entretanto, a diferença entre os dois tipos de análise de correspondência múltipla (não-ajustada e ajustada) é justificada apenas pela mudança em escala, sugerindo que um reescalonamento dos resultados provenientes da análise de correspondência múltipla solucionaria o problema da variabilidade das inércias, aliando a definição de inércia média fora das diagonais obtida na análise de correspondência múltipla ajustada em substituição à inércia total obtida na análise de correspondência múltipla tradicional. Portanto, Greenacre (2005) recomenda um simples ajuste das inércias, levando em consideração o número de variáveis utilizadas (Q) e o número total de categorias (J) de todas as variáveis. Na sua abordagem, o ajuste deve ser aplicado para cada autovalor $\geq 1/Q \geq$. Para tal, calculou-se cada autovalor ajustado pelos autovalores ajustados e pelo número Q de variáveis, utilizando-se a soma quadrada dos autovalores brutos para a obtenção de um novo valor médio ajustado (inércia média fora das diagonais), possibilitando o cálculo das percentagens através da razão entre os autovalores ajustados e o valor médio ajustado. Assim, é possível ponderar a inércia pelo número de categorias e variáveis utilizadas, obtendo um resultado aprimorado e com melhor qualidade de ajuste, melhorando o poder de explicação da variabilidade das dimensões.

A **técnica de análise de agrupamento** foi empregada como um complemento à interpretação do gráfico obtido na análise de correspondência, auxiliando a identificação dos grupos. Seria como uma análise multivariada confirmatória dos resultados obtidos, dando uma maior segurança nas afirmações, tentando trazer uma objetividade maior na análise subjetiva da plotagem da correspondência. A modalidade de análise de agrupamento empregada foi a hierárquica, cuja distância utilizada consiste na menor

distância euclidiana entre dois objetos para a formação do agrupamento, sendo os elementos agrupados de acordo com suas proximidades geométricas num processo iterativo através da junção inicial de dois elementos que formem um nó de agregação (Mota, 2004).

A análise desses nós é resumida a um gráfico do tipo dendograma, no qual se observa todo esse processo de agregação e a conseguinte construção dos agrupamentos propriamente ditos, para tal a análise de agrupamento leva em conta os dados da variabilidade de d dimensões da análise de correspondência determinadas pelo inverso do número total de dimensões analisadas $(1/n)$ (Salvador Figueras, 2003).

3. Resultados

3.1. Resiliência e eventos adversos ocorridos na escola

Cinco variáveis foram analisadas e poderão ser observadas nas Tabelas 1 e 2 e no Gráfico 1. Três delas correspondem às variáveis que compõem a subescala escola (seis dimensões); resiliência (presença e não-presença de resiliência, totalizando duas dimensões); e violência escolar (três dimensões: ausência, moderada e alta). A distância entre os pontos foi representativa das associações entre as variáveis. O eixo 1 ao eixo 3 explica 58% da variabilidade das dimensões (Tabela 1) e ficam bem-definidos na plotagem dos dados (Gráfico 1). Na construção dos grupos, utilizou-se a plotagem tri-dimensional das coordenadas do eixo 1, 2 e 3, verificando-se a constituição de três grupos:

1. Ausência de todos os eventos de vida na escola, inclusive violência escolar, que se agregam no mesmo grupo dos resilientes;
2. Presença de eventos de vida na escola, em conjunto com violência escolar moderada; e
3. Não-resilientes formam um grupo separado dos demais juntamente com aqueles que sofrem elevada violência escolar.

Na análise das contribuições absolutas (Tabela 2), verifica-se que as categorias da questão *a* e as da resiliência têm uma maior contribuição para a explicação da inércia da dimensão 2 e as categorias das questões *c* e *d* têm maiores contribuições absolutas para a constituição da inércia da dimensão 1, assim como a ausência da violência escolar. Já a violência escolar alta e moderada apresentam maior peso na formação do eixo 3.

No ajuste das inércias, obtêm-se novos valores para as duas primeiras dimensões (autovalores≥1/5) e um novo valor médio ajustado de 0,02, o que conduz a um poder de explicação de 83,51% já na primeira dimensão, diminuindo para 0,63% ao se observar a segunda dimensão.

Para confirmar os achados, procedeu-se a análise de agrupamento com as coordenadas de três (d) dimensões, determinadas pelo inverso do número total de dimensões analisadas na correspondência ($1/n=1/6$), correspondendo a dimensões com valores de inércia não-ajustada parcial até 17%. Durante a análise do dendograma (Gráfico 4), observa-se que com índice de corte igual a 0,25, confirmam-se os três grupos obtidos pela análise da plotagem da análise de correspondência, com diferenças em três pontos: a alta violência escolar substituiria a violência escolar moderada no grupo de presença de eventos de vida na escola, enquanto a violência escolar moderada se uniria à violência escolar baixa no grupo composto por indivíduos resilientes e com ausência de eventos de vida, sendo assim os não-resilientes estariam isolados em um grupo à parte.

3.2. Resiliência e eventos adversos ocorridos com amigos e namorados

Esta subescala de evento de vida é composta por três eixos, analisados com duas variáveis (quatro dimensões), mais a resiliência (presença e não-presença). A distância entre os pontos é de fundamental importância para associação e interpretação dos grupos. Na forma tri-dimensional há explicação de 100% de variabilidade dos eixos, neste caso pode-se ter uma visão e informação completa da formação de todos os grupos. Entretanto, trabalhou-se com as duas primeiras dimensões que explicam 71% da inércia total (Tabela 3).

Na interpretação do gráfico gerado na análise de correspondência para as duas primeiras dimensões (Gráfico 2), observou-se que:

1. Ausência de eventos de vida na esfera de amigos e namorados se agrupa com o indivíduo resiliente;
2. Presença dos eventos de vida se agrupam separadamente, conformando um segundo grupo; e
3. Indivíduo não-resiliente forma um ponto em separado dos dois outros grupos.

Quanto à contribuição absoluta (Tabela 4), verifica-se que as categorias das questões *a* e *c* têm maior importância na explicação da variabilidade do eixo 1 enquanto as categorias da resiliência têm maiores contribuições na inércia do eixo 2. Ao se analisar as inércias ajustadas para as duas primeiras dimensões (auto-valores > = 1/3), obtém-se um poder de explicação de 71% já na primeira dimensão, enquanto a segunda explica somente 0,42% de todo o processo. A análise de agrupamento com as coordenadas de duas dimensões da análise de correspondência, determinadas pelas dimensões com valores de variabilidade não-ajustada de no máximo o inverso do número de dimensões utilizadas (1/3 = 33,33%), confirmou esses três grupos. No Gráfico 5, ao se utilizar o índice de 0,2, obtêm-se a formação dos três grupos de forma mais clara do que na análise dos gráficos de correspondência, visto que a localização da presença de resiliência encontrava-se ligeiramente distante da ausência dos eventos de vida.

3.3. Resiliência e eventos adversos ocorridos na localidade

Nesta subescala de evento de vida há uma composição tri-dimensional composta por duas questões de eventos ocorridos na localidade (quatro dimensões), a variável resiliência (duas dimensões) e violência na localidade (três dimensões). Optou-se por trabalhar com os três primeiros eixos que retêm 67% da variabilidade total (Tabela 5). Ao se analisar a dimensão da comunidade podem-se verificar três aglomerações:

1. Ausência de eventos de vida e de violência na localidade se associam ao indivíduo resiliente;
2. Presença de eventos de vida e elevada violência na comunidade, associadas ao indivíduo não-resiliente; e
3. Indivíduos vítimas de moderada violência na localidade.

As contribuições absolutas das categorias das questões a e b, além das categorias ausência/elevada violência na localidade são mais notáveis na dimensão 1. A variável resiliência se faz presente através de maiores contribuições absolutas na dimensão 2 e aqueles que sofreram moderada violência na localidade são os que mais contribuem para a dimensão 3 (Tabela 6). O ajuste das inércias faz-se necessário nas duas primeiras dimensões, correspondendo a 86,76% de poder de explicação pelas "novas" inércias, destacando-se a primeira dimensão com 86,25% de inércia explicada. A análise de agrupamento das coordenadas de três dimensões (variabilidades não-ajustadas até 20%) da análise de correspondência, confirma a formação dos três grupos observados com corte de 0,3, excetuando-se a ausência de resiliência que se isola num quarto grupo (Gráfico 6). Todavia, devido à limitação da apresentação gráfica preferiu-se não considerar essa quarta dimensão. A proximidade gráfica sugere que esta pode ser incluída na dimensão da presença de eventos de vida.

4. Discussão

Os dados da área comportamental, muitas vezes, são de mais difícil compreensão estatística pelo seu caráter de subjetividade, principalmente em um conceito tão recentemente estudado como a resiliência. Neste trabalho, conseguiu-se reduzir na análise os eventos estressantes e constatar-se que a ausência de eventos ocorridos durante a vida da criança/adolescente relaciona-se a ser um adolescente resiliente. Todavia, a ocorrência desses eventos de vida não implica necessariamente em ser um indivíduo não-resiliente.

Através dos resultados verifica-se que não se conseguiria realmente encontrar padrões de associação por intermédio de simples testes estatísticos. A relação eventos

de vida e resiliência é mais complexa, como já sinalizam vários autores (Rutter, 1981; Tavares, 2001; Garmezy; Rutter, 1988), misturando-se fatores de risco e proteção num arranjo único feito por cada indivíduo. O fato de que o adolescente não-resiliente não necessariamente é aquele que vivenciou várias situações traumáticas em qualquer das esferas estudadas, se deve aos mecanismos protetores oferecidos pelo meio e existentes no próprio indivíduo, que podem reduzir o efeito dos eventos traumáticos.

A dimensão da comunidade é a única que apresenta resultados diferenciados. Nela, a ocorrência de eventos de vida e de elevada violência é característica marcante dos não-resilientes. Esse dado talvez sugira que a comunidade possa influenciar mais negativamente o comportamento de superação dos traumas vivenciados, atingindo diretamente o oferecimento de suporte social. Os demais contextos de ocorrência não chegam a influenciar negativamente, apesar de sua ausência funcionar como um fator protetivo, a ocorrência de traumas não configura ausência de resiliência.

A análise de correspondência contribuiu especialmente no que se refere à formação das dimensões dos eventos de vida, observando-se que a inclusão da resiliência fez com que a distância relativa dos pontos ao centróide se reduzisse e, portanto, os dados estivessem melhores agrupados. Trouxe contribuições importantes para o entendimento do mecanismo trauma-resiliência e auxilia na restrição de foco nos futuros estudos. Indica também a necessidade de estudos futuros realizarem análises conjuntas de eventos de vida e fatores de proteção. Além disso, a utilização do ajuste das inércias indica como a qualidade do ajuste pode ser aprimorada pela aplicação de ponderações que levem em consideração o número de variáveis e categorias empregadas na análise, já que em muitos casos só a primeira dimensão explica de 71% a 86% da variabilidade total. Por exemplo, na análise da área escolar verifica-se o grande incremento no poder de explicação, todavia as inércias não-ajustadas de três dimensões explicavam apenas 58%, ao passo que a inércia ajustada da primeira dimensão alcançou um poder de 84% de explicação. Assim sendo, verifica-se também a grande importância na utilização de inércias ajustadas, a fim de trazer percentagens explicativas mais representativas e elevadas do complexo fenômeno resiliência e eventos de vida.

Finalizando, lembra-se que este trabalho tem caráter exploratório e que a tentativa de classificar os eventos de vida se restringe ao presente estudo, não podendo ser generalizado.

Referências bibliográficas

ARANHA, R.N. (2004). *et al.* Análise de correspondência para avaliação do perfil de mulheres na pós-menopausa e o uso da terapia de reposição hormonal. Cademos de Saúde Pública, v. 20, n.1, p.100-108.

ASSIS, S.G.; PESCE, R.P.; AVANCI, J.Q. (2006). *Resiliência*: enfatizando a proteção dos adolescentes. Porto Alegre/RS: Artmed.

BENDIXEN, M. (1996). A practical guide to the use of correspodence analysis in marketing research. *Marketing Research On-Line*, v.1, p.17-38.

BOUVIER, P. (1999). *et al.* Typology and correlates of sexual abuse in children and youth: multivariate analyses in a prevalence study in Geneva. *Child Abuse & Neglect*, v.23, n.8, p. 779-790.

CLAUSEN, S.E. *Applied correspondence analysis:* an introduction. Thousand Oaks, CA: Sage, 1988. (Sage University papers Series on Quantitative Applications in the Social Sciences).

GARMEZY, N.; RUTTER, M. (1998). *Stress, coping and development in children*. New York: McGraw-Hill.

GREENACRE, M.J. (2005). *From Correspondence Analysis to Multiple and Joint Correspondence Analysis*. Disponível em 1° de Setembro na Social Science Research Network - SSRN: http://ssrn.com/abstract=847664.

GREENACRE, M.J. *Practical correspondence analysis*. In: *Looking at Multivariate Data*. New York: J.Wiley & Sons, 1981. p.81-107.

HERDMAN, M; FOX-RUSHBY, J. & BADIA, X. (1998). *A model of equivalence in the cultural adaptation of HRQol instruments: the universalist approach*. Quality of Life Research, v.4, p.323-335.

ILANUD. Instituto Latino Americano das Nações Unidas para a Prevenção do Delito e Tratamento do Delinqüente. Violências nas escolas. *Revista do ILANUD*, n.16, 2000.

MOTA, J.C. (2004). *Violência contra a mulher praticada pelo parceiro íntimo:* estudo em um serviço de atenção especializado. Tese (Mestrado) – Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz.

OHLSSON, E. (1998). *Sequential Poisson sampling*. Journal of Official Statistics, v.14, p.149-162.

PITZNER, J. K.; DRUMMOND, P. D. (1997). *The Reliability and Validity of Empirically Scaled Measures of psychological/Verbal control and Physical/Sexual Abuse: Relationship between current negative mood and a history of abuse independent of other negative life events*. Journal of Psychosomatic Research, v.43, n.2, p.125-142

RUTTER, M. (1981). Stress, coping and development: some issues and some questions. *Journal of Child Psychology and Psychiatry*, v.22, n.4, p. 323-356.

SALVADOR FIGUERAS, M. Análisis de correspondências. Disponível em:
< http://www.5campus.com/leccion/correspondencias>. Acesso em: 2003.

TAVARES, J. (2001). (org.). A resiliência na sociedade emergente. In: *Resiliência e Educação*. São Paulo: Cortez.

TENENHAUS, M.; YOUNG, F.W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psycometrika*, v. 50, p.91-119.

TROMBETA, L.H.; GUZZO, R.S.L. (2002). *Enfrentando o cotidiano adverso:* estudo sobre resiliência em adolescentes. Campinas/SP: Alínea.

WAGNILD, G.M; YOUNG, H.M. (1993). Development and psychometric evaluation of the resilience sacle. *Journal of Nursing Measurement*, n.1, p.165-178.

YOUNG, F.W. (1984). Scaling. *Annual Review of Psychology*, v.35, p.55-81.

ANEXOS

Tabela 1 - Autovalores, Variabilidade por eixo e Inércia Total – Vida escolar

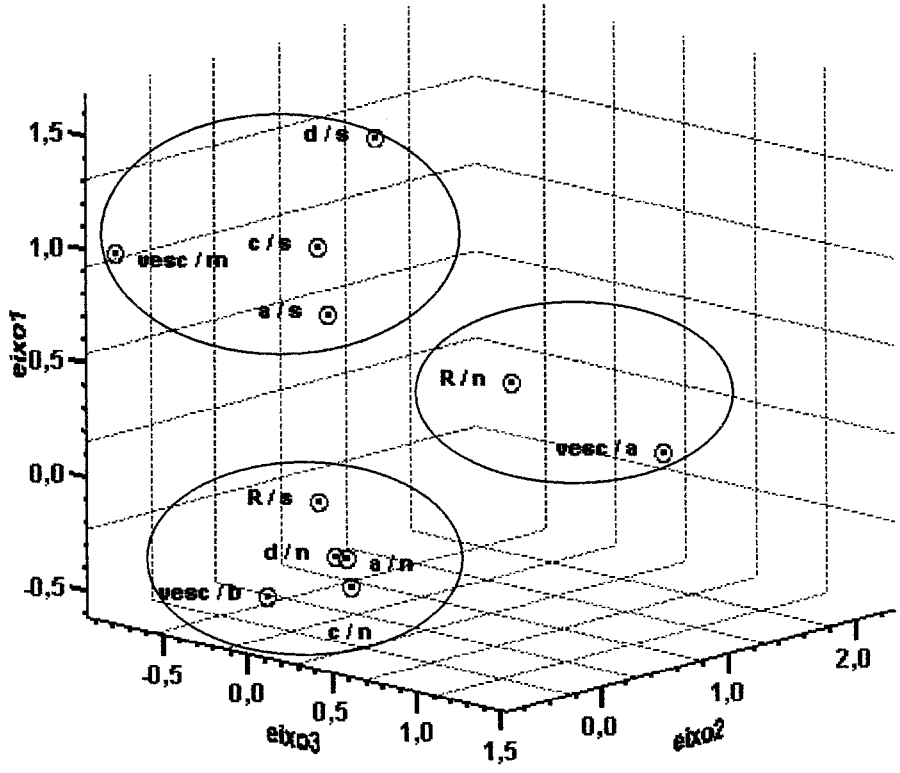
Autovalores	1	2	3	4	5	6
Valores	0,2945	0,2082	0,1992	0,1865	0,1682	0,1433
Inércia Parcial (%)	25%	17%	17%	16%	14%	12%
% Inércia Total	25%	42%	58%	74%	88%	100%

Tabela 2 - Contribuições Absolutas de cada ponto pelas dimensões de Vida Escolar

Variáveis	N	%	Absoluta					
			1	2	3	4	5	6
a / S	349	5,74	6,01	12,89	10,29	39,64	0,43	2,06
a / N	868	14,26	2,42	5,18	4,14	15,94	0,17	0,83
c / N	864	14,20	9,43	0,33	0,00	0,03	7,87	11,34
c / S	353	5,80	23,07	0,81	0,00	0,08	19,27	27,76
d / N	1020	16,76	5,46	0,02	0,24	1,27	0,89	8,30
d / S	197	3,24	28,27	0,13	1,26	6,58	4,62	42,95
Vesc / ausência	645	10,60	10,21	0,24	7,95	0,08	26,34	2,18
vesc / alta	249	4,09	13,44	7,81	15,30	4,66	36,60	1,73
Vesc / moderada	323	5,31	1,68	9,90	55,03	2,22	3,76	0,87
R / N	144	2,37	0,01	55,27	5,11	25,99	0,03	1,75
R / S	1073	17,63	0,00	7,42	0,69	3,49	0,00	0,24

Codificação: a - Mudanças de escola
c - Suspensão por confusões na escola
d - Conflitos sérios com professor
R - Resiliência
S - Sim (presença)
N - Não (ausência)
vesc –Violência escolar

Gráfico 1 - Plotagem das coordenadas pela Análise de Correspondência dos eixos 1 e 2 da dimensão Escola



Codificação: a - Mudanças de escola
c - Suspensão por confusões na escola
d - Conflitos sérios com professor
R - Resiliência
S - Sim (presença)
N - Não (ausência)
vesc /b–Violência Escolar ausência
vesc /m–Violência Escolar moderada
vesc/a –Violência Escolar elevada

Tabela 3 - Autovalores, Variabilidade por eixo e Inércia Total – Dimensão amigos e namorados

Autovalores	1	2	3
Valores	0,3693	0,3361	0,2945
Inércia Parcial (%)	37%	34%	29%
% Inércia Total	37%	71%	100%

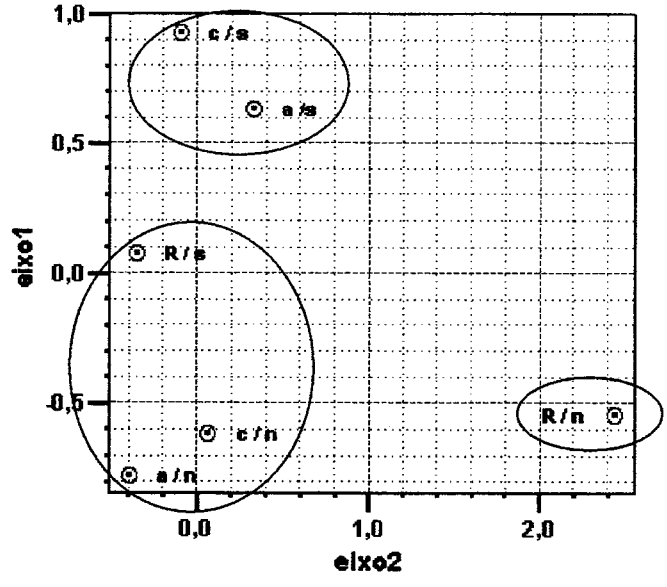
Tabela 4 - Contribuições Absolutas de cada ponto pelas dimensões de Amigos e Namorado

			Absoluta		
Variáveis	N	%	1	2	3
a / S	703	18,28	20,15	5,89	19,13
a / N	579	15,05	24,47	7,15	23,22
c / S	508	13,21	31,09	0,4	28,89
c / N	774	20,12	20,41	0,26	18,96
R / N	163	4,24	3,39	75,33	8,56
R / S	1119	29,1	0,49	10,97	1,25

Codificação: a - Separação de amigo por morte, mudança de residência ou briga
c - Sofre ou Sofreu com o término de namoro

S- Sim (presença)
N- Não (ausência)
R- Resiliência

Gráfico 2 - Plotagem das coordenadas pela análise de correspondência dos eixos 1 e 2 da dimensão amigos e namorado



Codificação: a - Separação de amigo por morte, mudança de residência ou briga
c - Sofre ou Sofreu com o término de namoro

S - Sim (presença)
N - Não (ausência)
R - Resiliência

Tabela 5 - Autovalores, variabilidade por eixo e inércia total – dimensão comunidade

Autovalores	1	2	3	4	5
Valores	0,3400	0,2569	0,2440	0,2147	0,1944
Inércia Parcial (%)	27%	21%	20%	17%	16%
% Inércia Total	27%	48%	67%	84%	100%

Tabela 6 - Contribuições absolutas de cada ponto pelas dimensões de comunidade

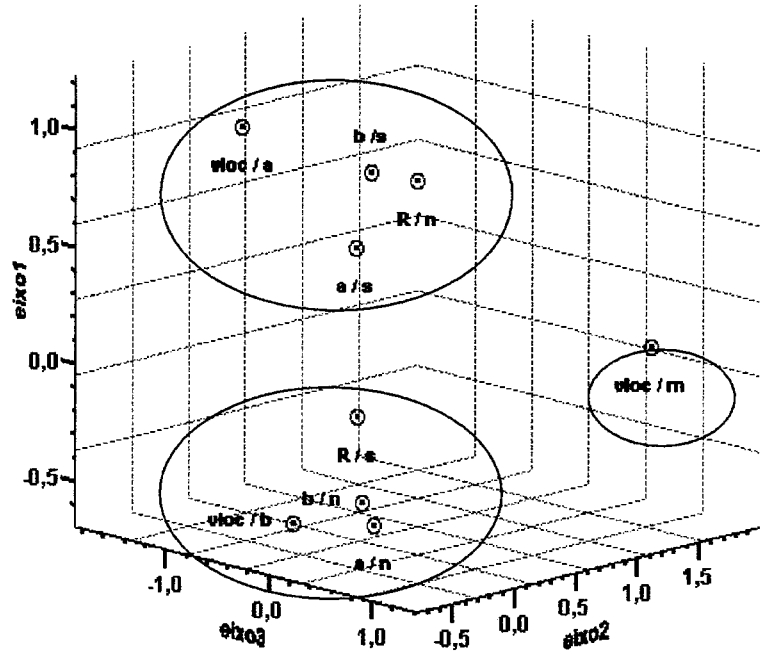
Variáveis	N	%	Absoluta				
			1	2	3	4	5
a / N	621	12,71	**14,65**	1,65	0,93	22,15	9,76
a / S	600	12,29	**15,16**	1,71	0,97	22,92	10,10
b / N	800	16,38	**10,60**	0,36	0,25	16,69	6,58
b / S	421	8,62	**20,14**	0,68	0,47	31,72	12,51
vloc / ausência	647	13,25	**13,53**	2,99	9,20	0,01	21,29
vloc / elevada	287	5,88	**23,46**	10,54	3,53	0,09	38,86
vloc / moderada	287	5,88	**0,46**	34,13	**41,39**	0,03	0,48
R / N	147	3,01	**1,75**	**42,16**	38,06	5,62	0,37
R / S	1074	21,99	**0,24**	**5,77**	5,21	0,77	0,05

Codificação: a - Já viu alguém ser gravemente ferido
b - Vive/viveu situação de perigo na vizinhança

S- Sim (presença)
N- Não (ausência)
R- Resiliência

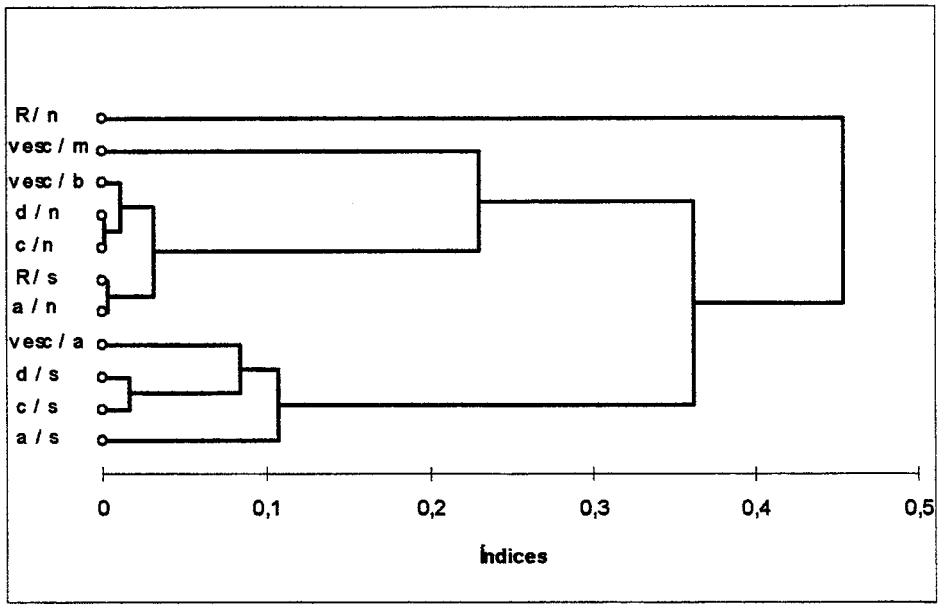
Vloc -Violência na comunidade

Gráfico 3 - Plotagem das coordenadas pela análise de correspondência dos eixos 1 e 2 da dimensão comunidade



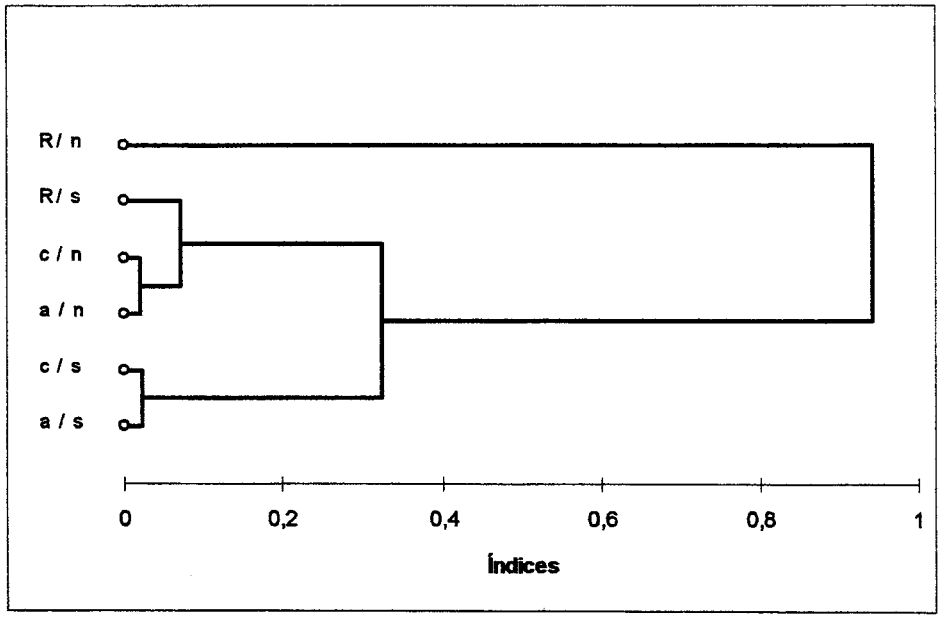
Codificação: a - Já viu alguém ser gravemente ferido
b - Vive/viveu situação de perigo na vizinhança
Vloc/b - Violência na comunidade ausência
N - Não(ausência) Vloc/a -Violência na comunidade elevada
S - Sim(presença) Vloc/m -Violência na comunidade moderada
R- Resiliência

Gráfico 4 - Dendograma das coordenadas da análise de correspondência da dimensão Escola



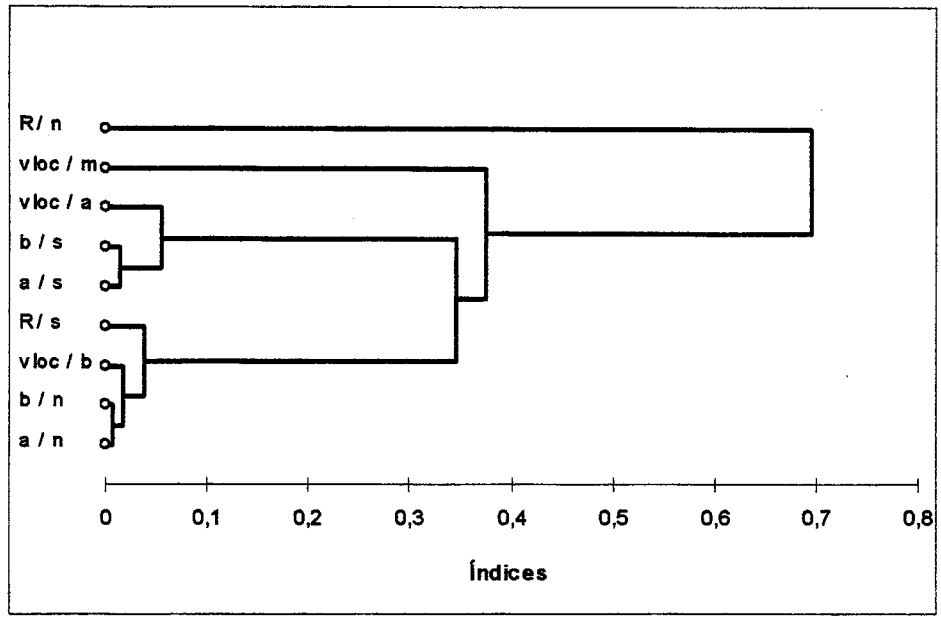
Codificação: a - Mudanças de escola; c - Suspensão por confusões na escola; d - Conflitos sérios com professor; R- Resiliência; S- Sim (presença); N- Não (ausência); vesc –Violência Escolar

Gráfico 5 - Dendograma das coordenadas da análise de correspondência da dimensão Amigos e Namorado



Codificação: a - Separação de amigo por morte, mudança de residência ou briga; c - Sofre ou Sofreu com o término de namoro; S - Sim (presença); N - Não (ausência); R - Resiliência

Gráfico 6 - Dendograma das coordenadas da análise de correspondência da dimensão comunidade



Codificação: a - Já viu alguém ser gravemente ferido
b - Vive/viveu situação de perigo na vizinhança
S - Sim (presença)
N - Não (ausência)
R - Resiliência
Vloc -Violência na comunidade

Abstract

The resilience has been revealed as an important concept of Health Area, especially in children and youth field. It is an important concept, consisting mainly in the ability to cope with stressful events. Life stress events are stressors that happens in families, schools, and so on. The study interviewed 1923 students from 11 to 19 years from São Gonçalo city (RJ) in 2003. Through *multiple correspondence analysis*, dimensions for life stress events was built and related with the resilience scale. In this study, we intend to define more necessary life stress events' dimensions through correspondence analysis from 29 questions. *Cluster analysis* was used to corroborate with the interpretation of the plots from two-way and/or three-way correspondence analysis. In built dimensions, the absent of events becomes related it the behavior resilience, however the presence of stressors does not become related to no resilience.

Key-words: multiple correspondence analysis, cluster analysis, resilience, life stress events, psychometric

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística Aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará, também, artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográficos e econométricos. Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores seja brasileiro e nem resida no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo 'duplo cego', isto é, os artigos são avaliados sem identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

O processo editorial da RBEs é eletrônico. Os artigos devem ser submetidos via e-mail para: hortega@ibge.gov.br

Após a submissão, o autor correspondente receberá um código para acompanhar o processo de avaliação do artigo. Caso não receba um aviso com este número no prazo de uma semana, fazer contato com a secretaria da revista no endereço:

Revista Brasileira de Estatística
IBGE – Diretoria de Pesquisas - Coordenação de Métodos e Qualidade
Av. República do Chile, nº 500, 10º andar
Centro, Rio de Janeiro – RJ
CEP: 20031-170
Tel.: 55 21 2142-0472
55 21 2142-4549
Fax: 55 21 2142-0039

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais entregues para publicação devem obedecer às normas seguintes.

1. Originais processados pelo editor de textos Word for Windows são preferidos. Entretanto, serão aceitos também, originais processados em LaTeX desde que sejam encaminhados acompanhados de versões em pdf, conforme descrito no item 3 a seguir;

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, também, devem figurar nesta página;

3. No caso da submissão não ser em Word for Windows, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, Latex). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página, como descrito no item 2, e outro contendo apenas o título, sem identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;

4. A segunda página do original deve conter resumos em português e inglês (*abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser digitado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com, no máximo, 150 palavras;

5. O artigo deve ser dividido em seções, numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado;

6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

7. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;

8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções de trabalho; e

9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos.