DOI: 10.5533/JBG-0368-1416-2017110127102

Análise multivariada aplicada aos estudos clínicos: notas práticas para profissionais de saúde

Multivariate analysis applied to clinical studies: practice notes for health professionals

Saint Clair Gomes Junior[1]*, Rosimary Terezinha de Almeida[2]

RESUMO

Métodos estatísticos multivariados possibilitam o tratamento simultâneo de um conjunto de variáveis permitindo, dessa forma, uma visão mais abrangente e realista de um problema estudado. Este artigo teve por objetivo apresentar alguns dos métodos multivariados mais comumente utilizados em estudos clínicos que podem ser, de um modo geral, agrupados em métodos de dependência (regressão linear multivariada, regressão logística e o modelo proporcional de Cox) e de interdependência (análise de componentes principais, análise de agrupamentos e análise de correspondência). Para todos os métodos descritos serão apresentados seus fundamentos de utilização, os métodos de estimação dos parâmetros mais utilizados, a forma de avaliação da qualidade do ajuste desses parâmetros e um exemplo prático. Por fim, serão discutidas questões que devem ser consideradas quando se realiza uma análise multivariada, tais como: o tamanho da amostra, as falhas nos dados em decorrência tanto de dados faltantes como, também, discrepantes, a escolha de pacotes estatísticos e o estabelecimento de um modelo teórico para apoio das análises.

Palavras-chave: análise multivariada; interpretação estatística de dados; modelos estatísticos.

ABSTRACT

Multivariate statistical methods allow the simultaneous processing of a set of variables, thereby allowing a more comprehensive and realistic view of a problem studied. This article aimed to present some of the multivariate methods most commonly used in clinical studies. These can be roughly grouped into: dependency methods (multivariate linear regression, logistic regression and the Cox proportional model); and interdependence methods (principal component analysis, cluster analysis and correspondence analysis). For each method, the background, the parameter estimation and the quality of adjustment, and an example of application were presented. Finally, some aspects that should be considered when performing a multivariate analysis, such as the sample size, the missing data, the outliers, the choice of a statistical package and the establishment of a theoretical model to support the analysis.

Keywords: multivariate analysis; statistical data interpretation; statistical models.

INTRODUÇÃO

Métodos estatísticos multivariados são recursos poderosos de análise de dados, uma vez que possibilitam o tratamento simultâneo de um conjunto de variáveis, dando assim uma visão mais abrangente e realista de um problema estudado[1-5].

No entanto, a utilização desses métodos requer conhecimentos prévios relacionados aos pressupostos de utilização, ao tamanho da amostra, aos tipos de variáveis, aos métodos de estimação dos parâmetros, entre outros aspectos importantes de serem considerados em função do impacto que causam na precisão e na validade dos resultados[1,6].

Assim, este artigo teve por objetivo apresentar alguns dos fundamentos da análise multivariada como, também, as etapas que devem ser consideradas para a sua realização e interpretação[2,5].

[1]Instituto Nacional de Saúde da Mulher, da Criança e do Adolescente Fernandes Figueira, Fundação Oswaldo Cruz – Rio de Janeiro (RJ), Brasil.
[2]Programa de Engenharia Biomédica, Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa em Engenharia, Universidade Federal do Rio de Janeiro – Rio de Janeiro (RJ), Brasil.
*Autor correspondente: scgomes@iff.fiocruz.br
Fonte de financiamento: nenhuma.
Conflito de interesses: nada a declarar.
Recebido em: 01/12/2016. Aprovado em: 29/02/2017.

O que é análise multivariada?

A análise multivariada pode ser entendida, de um modo geral, como um conjunto de métodos estatísticos que possibilitam analisar múltiplas variáveis de um determinado fenômeno[1,2,6]. Esse fenômeno pode ser, por exemplo, a eficácia de um medicamento e sua relação com fatores tais como sexo, idade, peso corporal, raça, tabagismo, etilismo, presença de comorbidades etc. O fenômeno multivariado também pode ser entendido como a identificação de um conjunto de características comportamentais para o diagnóstico diferencial de crianças com suspeita de transtorno de déficit de atenção e hiperatividade (TDAH) que, por sua natureza, envolvem um grande conjunto de variáveis que se inter-relacionam para definirem o diagnóstico.

Quais são os métodos de análise multivariada?

De modo geral, os métodos de análise multivariada podem ser agrupados de acordo com sua relação de dependência (regressão linear múltipla, regressão logística, regressão de riscos proporcionais de Cox, dentre outras) ou de interdependência (análise de componentes principais, análise de agrupamentos, análise de correspondência, dentre outras). Os métodos de dependência têm como principal objetivo a análise da contribuição individual de um conjunto de variáveis para um determinado desfecho de interesse, enquanto que os de interdependência visam, entre outros objetivos, a redução da dimensionalidade e o agrupamento de indivíduos ou de variáveis com características semelhantes[1-3, 6-9].

Muitos desses métodos são extensões de análises uni ou bivariadas, como é o caso da regressão linear múltipla, que é uma extensão da regressão linear simples, enquanto que outros foram projetados para tratarem de questões de natureza exclusivamente multivariada, como a análise fatorial e a de agrupamentos, cujos pressupostos demandam um conjunto mínimo de variáveis para fornecerem resultados confiáveis[1].

Regressão linear multivariada

A regressão linear multivariada permite analisar a relação de uma variável numérica com um conjunto de outras variáveis (numéricas ou categóricas), por meio de uma relação matemática (Equação 1).

$$Y = b_o + b_1X_1 + \ldots + b_kX_k + e \quad (1)$$

Em que:

Y é a variável dependente;

X_i são as variáveis independentes (ou preditoras);

b_i são os coeficientes de regressão; e

e é o erro de estimativa.

Os pressupostos básicos de um modelo de regressão linear multivariada são de que a variável dependente (Y) seja numérica, preferencialmente contínua, o seu valor médio esteja linearmente relacionado com as variáveis independentes e o erro de estimativa (a diferença entre os valores observados e estimados pelo modelo de regressão) tenha distribuição normal com média zero e variância constante[1,6,10].

Os coeficientes do modelo de regressão linear múltipla (b_i) fornecem a contribuição individual das variáveis independentes (X_i) para a alteração dos valores médios da variável dependente (Y). Esses coeficientes podem ser obtidos por meio de métodos de estimação como o dos mínimos quadrados ou da máxima verossimilhança, que se utilizam dos dados observados para a obtenção do melhor ajuste do modelo ao conjunto de dados[1,11]. A significância estatística dos coeficientes estimados pode ser avaliada por meio do teste F de Snedecor, que permite testar a significância estatística da contribuição do conjunto de variáveis independentes para a alteração do valor médio da variável dependente, e do teste *t* de Student, que permite testar a significância estatística da contribuição individual de cada uma das variáveis independentes para a alteração da variável dependente[1,11].

O poder de explicação das variáveis independentes de um modelo de regressão linear pode ser avaliado por meio do seu coeficiente de determinação (R^2), que é calculado dividindo-se a variância do erro de estimativa pela variância da variável dependente[1,11]. Essa estatística varia de 0 a 1 e, quanto maior o seu valor, melhor o ajuste do modelo aos dados observados[1,11].

O modelo de regressão linear multivariado foi utilizado por Costa et al.[12] para determinar o valor preditivo do apoio social na qualidade de vida relacionada com a saúde dos doentes com esclerose múltipla. Os coeficientes das 12 variáveis consideradas no modelo foram estimados pelo método dos mínimos quadrados utilizando dados de 150 pacientes.

Os autores observaram que a idade produzia uma alteração significativa na redução da vitalidade (b=-0,323, p<0,000), na funcionalidade social (b=-0,192; p=0,013) e no desempenho emocional (b =-0,271; p=0,001). Também verificaram que cuidados de saúde estavam relacionados com a vitalidade (b=-0,290; p<0,000), funcionalidade social (b=0,336; p<0,000), desempenho emocional (b=0,214; p=0,003) e saúde mental (b=0,447; p<0,000).

Regressão logística

A regressão logística é adequada para analisar fenômenos descritos por variáveis categóricas, mais frequentemente dicotômicas, e que podem ser descritas por uma combinação de um conjunto de variáveis independentes. Os pressupostos da regressão logística são de que a variável dependente seja descrita por categorias mutuamente

exclusivas e a chance de ocorrência de um evento esteja linearmente relacionada com as variáveis independentes[1,2,10,11].

Tal como ocorre na regressão linear múltipla, a regressão logística também é descrita a partir de uma equação matemática (Equação 2).

$$p = \frac{e^{b_0 + b_1 X_1 + \ldots + b_k X_k}}{1 + e^{b_0 + b_1 X_1 + \ldots + b_k X_k}} \quad (2)$$

Em que:
p representa a probabilidade de ocorrência do evento analisado;
b_i são os coeficientes de regressão; e
X_i são as variáveis independentes.

A Equação 2 pode ser reescrita como a Equação 3:

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X_1 + \ldots + b_i X_i \quad (3)$$

Em que:
$ln\left(\frac{p}{1-p}\right)$ representa o *logit* e fornece o logaritmo da chance de ocorrência do evento analisado.

A Equação 3 apresenta características desejáveis do ponto de vista computacional e permite que se obtenha estimativas dos coeficientes b_i por meio do método da máxima verossimilhança[1,2]. Esses coeficientes, quando reescritos para a Equação 2, fornecem a razão de chance (ou *odds ratio* – OR) ajustada para o conjunto de variáveis independentes incluídas no modelo[1,2].

A qualidade do ajuste do modelo de regressão logística pode ser avaliada a partir do teste da razão de verossimilhança, que consiste na comparação das verossimilhanças de diferentes modelos ajustados com diferentes números de variáveis. Assim, pode-se verificar o impacto que a inclusão ou exclusão de determinada variável causa no ajuste do modelo[1,2,6,10,11]. A significância estatística dos coeficientes de um modelo de regressão logística pode ser avaliada pelos testes de Wald ou de Score[1,2].

O modelo de regressão logística foi utilizado por Silva et al.[13] para identificar os principais determinantes da detecção de atipias celulares no programa de rastreamento do câncer do colo de útero no estado do Rio de Janeiro. Os autores utilizaram dados de 65.535 resultados de exames citopatológicos registrados no Sistema de Informação do Câncer do Colo do Útero (SISCOLO) e estimaram os coeficientes para 20 variáveis utilizando o método de máxima verossimilhança.

Foi possível observar, por meio da análise dos coeficientes, um aumento na chance de atipias quando os exames foram realizados em laboratórios de referência (OR=2,827 e intervalo de confiança de 95% — IC95% 2,256–3,082), na presença de elementos celulares (OR=3,897; IC95% 3,489–4,364), na ocorrência de metaplasia escamosa imatura (OR=2,196; IC95% 1,191– 2,518) e quando havia ausência de microrganismos da microbiota vaginal (OR=2,165; IC95% 1,964–2,386).

Modelos de riscos proporcionais de Cox

Os modelos de riscos proporcionais de Cox são utilizados quando o interesse é verificar o efeito de fatores de risco ou de prognósticos no tempo ou na velocidade de ocorrência de um evento de um indivíduo ou de um grupo de indivíduos[3,11,14-16]. Esses modelos fornecem as estimativas das razões de risco dos fatores estudados para diferentes momentos do tempo, permitindo avaliar o impacto desses fatores na taxa de ocorrência do evento de interesse ao longo do tempo[14,15].

Uma característica importante desse modelo é a presença de censuras, que ocorrem quando os participantes foram expostos a uma intervenção ou a fatores de risco, porém a ocorrência do evento não teve como ser verificada[15].

As principais suposições dos modelos de riscos proporcionais de Cox são de que os riscos são constantes e proporcionais entre os indivíduos[14,15]. Para avaliação desses pressupostos, normalmente utilizam-se as curvas de Kaplan-Meier, que estimam a curva de sobrevida considerando o número de ocorrências de eventos observados na amostra, e o teste do log-rank, que testa a hipótese de que a distribuição desses eventos não difere significativamente entre os grupos analisados[15].

Os coeficientes do modelo de riscos proporcionais de Cox consideram a ocorrência das censuras observadas e podem ser estimados pelo método da máxima verossimilhança[15]. Os testes de Wald e a razão de verossimilhança são utilizados para avaliar a significância dos coeficientes e a qualidade do ajuste do modelo selecionado[15].

O modelo de riscos proporcionais de Cox foi utilizado por Ayala[17] para analisar a sobrevida de 655 mulheres com câncer de mama atendidas no Sistema Único de Saúde (SUS) de Joinville, Santa Catarina, Brasil. O tempo de sobrevida foi definido como o período entre o diagnóstico de câncer de mama e a ocorrência do óbito, segundo a declaração de óbito. As mulheres sobreviventes até o momento do término do levantamento de dados foram incluídas no grupo censuras, uma vez que essas participantes foram expostas, porém, a ocorrência do evento não teve como ser verificada. O método de Kaplan-Meier e o teste de log-rank foram utilizados para avaliar o pressuposto de proporcionalidade nos grupos formados.

Os coeficientes do modelo foram estimados considerando as variáveis referentes ao nível de estadiamento e às faixas etárias, em que se verificou que a taxa de óbito apresenta um risco relativo com o estadiamento I de 3,26 (IC95% 1,29–8,22), com o estadiamento II de 15,40 (IC95% 6,22–38,12) e com o estadiamento III de 25,51 (IC95% 9,64–67,51). O modelo utilizado não identificou diferenças significativas no risco de óbito entre as faixas etárias consideradas.

Análise de componentes principais

A análise de componentes principais consiste em transformar um conjunto de variáveis (X1, X2,... ,Xp) em um novo conjunto de variáveis não correlacionadas — Y1(CP1), Y2(CP2),... ,Yp (CPp) —, independentes entre si e ordenadas a partir de suas variâncias[1,5,6].

A ideia principal desse procedimento é de que poucas componentes principais são capazes de incorporar a maior parte da variabilidade dos dados originais permitindo, dessa forma, descartar as demais componentes e reduzir o número de variáveis. Esse método vem sendo utilizado para construção de indicadores, eliminação de variáveis sobrepostas, reconhecimento de padrões, entre outras finalidades[6,18].

O principal pressuposto da análise de componentes principais é a possibilidade de expressar as características comuns das variáveis originais a partir de um conjunto menor de variáveis formadas pela combinação linear das variáveis originais[18]. Esse pressuposto pode ser verificado por meio dos índices de Kaiser-Meyer-Olkin (KMO), que avaliam a adequabilidade da amostra a uma análise de componentes principais, o qual deve estar entre 0,5 e 1,0 para ser aceito como adequado; do coeficiente de correlação de Pearson, que avalia a correlação bivariada; e o teste de esfericidade de Bartlett, que testa a hipótese de correlação multivariada[1,6].

As componentes principais são ordenadas de tal modo que a primeira concentra o maior percentual da variância total existente na amostra; a segunda, o segundo maior percentual da variância total; e assim sucessivamente[1,6,19].

A estimação das componentes principais pode ser realizada considerando os valores originais ou seus respectivos valores padronizados[6,19]. Geralmente, a estimação das componentes principais a partir de valores padronizados tende a fornecer combinações mais equilibradas e com resultados de mais fácil interpretação, isso porque os métodos de padronização tendem a equilibrar a variabilidade e a homogeneizar os dados com relação à escala de medida[6,19].

O número de componentes principais pode ser definido considerando o percentual de variância total explicada por um determinado número de componentes ou pela conveniência do pesquisador[6]. Não existe uma regra para a determinação do número de componentes principais. A situação ideal é aquela em que sejam consideradas poucas componentes e que essas concentrem o maior percentual da variância da amostra[6].

A análise de componentes principais foi utilizada por Oliveira et al.[20] para analisar 650 respostas a 14 perguntas relativas às barreiras enfrentadas pelos médicos do Distrito Federal para promover a alimentação saudável entre seus pacientes. Os pressupostos da análise de componentes principais foram avaliados pelo índice de KMO e pelo teste de esfericidade de Barlett.

O índice de KMO forneceu um valor de 0,79, indicando adequabilidade da amostra a uma análise de componentes principais. O teste de esfericidade de Barlett forneceu um valor de 859,95 ($p<0,000$), logo, havendo evidência de correlação multivariada nos dados analisados. Os autores decidiram por uma solução com 4 componentes principais, o que resumiu 59% da variabilidade da amostra. A primeira componente principal (CP1) apresentou uma variabilidade de 19% e agrupou variáveis como hábitos culturais, resistência à mudança, falta de interesse, baixa instrução e condições precárias de moradia, e foi nomeada pelo autor como sendo a de barreiras socioculturais dos pacientes. A segunda componente principal (CP2) apresentou uma variabilidade de 14% e agrupou as variáveis falta de interesse dos profissionais e desorganização do serviço, e foi nomeada como sendo barreiras relacionadas ao processo gerencial. A terceira componente principal (CP3) apresentou uma variabilidade de 13% e agrupou variáveis relacionadas à quantidade de pacientes, à falta de espaço físico nos serviços, à falta de recursos humanos e à falta de integração interprofissional, sendo nomeada pelos autores como barreiras do serviço de saúde. Por fim, a quarta componente principal (CP4) apresentou uma variabilidade de 13% e agrupou as variáveis ausência de treinamento e reciclagem, falta de conhecimentos e falta de material didático, e foi nomeada como barreiras educacionais e de comunicação.

Análise de agrupamentos

A análise de agrupamentos é um conjunto de métodos cuja finalidade é a formação de grupos de acordo com suas similaridades ou dissimilaridades[1,6,8]. As decisões que envolvem uma análise de agrupamentos estão relacionadas ao cálculo da distância existente entre as observações, ao algoritmo de identificação desses grupos e à quantidade de grupos que serão identificados[6,8,9,19].

As medidas de distância permitem decidir o quão próximo (similar) ou distante (dissimilar) uma observação encontra-se da outra. O Quadro 1 faz uma breve apresentação das medidas de distância mais comumente utilizadas na análise de agrupamento e presentes nos pacotes estatísticos[1,6,19], não havendo consenso de qual dessas medidas fornece o resultado mais adequado (grupos com o máximo de homogeneidade entre si). Assim, recomenda-se a utilização de diferentes abordagens a fim de analisar qual das medidas atende melhor os pressupostos de uma análise de agrupamento para o conjunto de dados analisado[3,6].

Os agrupamentos são formados a partir de algoritmos que podem ser do tipo hierárquico ou não hierárquico (também conhecidos como particionais)[1,6]. Os métodos hierárquicos de agrupamento têm uma utilização maior nas análises exploratórias e auxiliam na identificação do número de grupos existentes no conjunto de dados

analisados[1,3,6,19]. Esses métodos são classificados em aglomerativos ou divisivos. Os métodos aglomerativos consideram a existência de "n" aglomerados formados por "n" elementos ou indivíduos que estão sendo agrupados de acordo com suas similaridades até a formação de um único grupo. Os métodos divisivos consideram a existência de um único aglomerado constituído de "n" elementos ou indivíduos e que vai sendo particionado em "n" aglomerados de acordo com as dissimilaridades[4,6].

Os métodos aglomerativos são mais utilizados e apresentam maior suporte computacional. Tal como as medidas de distância, diferentes propostas de métodos de agrupamentos hierárquicos vêm sendo propostas e a decisão está relacionada à medida de distância adotada, ao tipo de dado analisado, à presença de valores discrepantes, ao grau de correlação entre as variáveis, entre outros fatores[6]. O Quadro 2 apresenta os principais algoritmos hierárquicos presentes nos pacotes estatísticos com algumas observações relativas a sua utilização[1,6].

O número de grupos em uma análise de agrupamentos hierárquica pode ser definido a partir da análise do dendograma, que é um gráfico que ilustra a formação dos agrupamentos de acordo com o nível de similaridade previamente definido. A principal dificuldade para interpretar os resultados da análise de agrupamentos por meio da inspeção de dendrogramas se deve ao fato de não haver um critério objetivo, devendo os grupos definidos serem analisados a fim de verificar se atendem os pressupostos de uma análise de agrupamento[6].

Quadro 1. Medidas de distância e de similaridade de acordo com o tipo de dado que se aplica.

Distância	Tipo de dado	Observações
Euclidiana	Numérico	Aplica-se a dados não padronizados e que estão mensurados na mesma escala
Euclidiana quadrada	Numérico	Aplica-se a dados mensurados em uma mesma escala, porém, existe o interesse em enfatizar as diferenças entre os objetos pela presença de *outliers*.
Euclidiana ponderada	Numérico	Aplica-se quando o interesse é dar maior peso para variáveis que se julga serem mais importantes para a classificação do que outras.
Minkowsky	Numérico	Aplica-se quando as características dos objetivos são independentes e de igual importância.
Mahalanobis	Numérico	Aplica-se quando as características dos objetivos são independentes e apresentam pesos diferentes.
Coeficiente de concordância	Categórico	Aplica-se quando os dados estão representados em uma escala binária (0 e 1) e o interesse é identificar a proporção de variáveis em que há concordância nos valores presentes e ausentes.
Coeficiente de Jaccard	Categórico	Aplica-se quando os dados estão representados em uma escala binária (0 e 1) e o interesse é identificar a proporção de variáveis em que há concordância apenas nos valores presentes.
Coeficiente de Gower e Legendre	Categórico	Aplica-se quando os dados estão representados em uma escala binária (-1 e 1) e o interesse é identificar a proporção de variáveis em que há concordância e discordância nos valores presentes e ausentes.

Quadro 2. Algoritmos de agrupamento hierárquico mais comumente utilizados de acordo com o tipo de dado que se aplica.

Algoritmo de agrupamento	Observações
Single Linkage	Inicia o procedimento pela procura dos dois objetos mais similares na matriz de similaridade. Em geral, grupos muito próximos podem não ser identificados. Muito sensível a valores discrepantes e dados faltantes.
Complete Linkage	Agrupa os elementos mais semelhantes e verifica a distância máxima do grupo para os objetos restantes. Tendência a formar grupos compactos e a isolar os valores discrepantes
Average Linkage	Utiliza a média aritmética das distâncias dos objetos de cada grupo para calcular a matriz de distâncias. Pouco sensível a valores discrepantes e dados faltantes. Tendência a formar grupos muito homogêneos.
Método de Ward	Utiliza o critério de aumento mínimo na variância do grupo para inclusão ou não de um elemento. Sensível à presença de valores discrepantes. Tendência a combinar grupos com o mesmo número de elementos. Alta homogeneidade interna. Exclusivo para variáveis numéricas.
Método do centroide	Utiliza vetores de médias dos grupos (centroides) que estão sendo analisados. Método robusto e que utiliza toda a informação do grupo para a comparação. Tende a suavizar diferenças entre os grupos. Exclusivo para variáveis numéricas

Os métodos não hierárquicos de agrupamentos formam grupos de "n" elementos, tendo como requisitos básicos a coesão interna e o isolamento dos grupos formados[1,21]. Esses métodos dividem um conjunto de dados otimizando alguma medida de qualidade previamente definida[6]. Os métodos não hierárquicos normalmente assumem que o número de agrupamentos finais seja conhecido, embora alguns algoritmos permitam que esse número possa variar durante a análise[6].

O método *k-means* é um dos métodos não hierárquicos mais utilizados e divide os "n" elementos nos "k" grupos previamente definidos de modo que a heterogeneidade interna dos agrupamentos seja minimizada[1,6,21]. Esse método tem um algoritmo que calcula a distância entre cada um dos valores existentes no banco de dados, geralmente a partir da distância euclidiana, porém, outras medidas encontram-se implementadas nos pacotes estatísticos[21]. Após o cálculo das distâncias, o algoritmo calcula os centroides para cada um dos agrupamentos previamente informados e esse valor é corrigido a cada interação do algoritmo, até que não se observem mais alterações significativas nos valores dos centroides[6,21].

A análise de agrupamentos foi utilizada por Feitosa e Almeida[22] para classificar os 850 municípios de Minas Gerais, Brasil, com relação à produção de 981.316 exames citopatológicos (*Papanicolaou*) registrados no Sistema de Informação do Câncer da Mulher do Sistema de Informação Ambulatorial do SUS (SISCAM-SAI/SUS) no ano de 2002. Foram consideradas as seguintes variáveis:

1. razão de exames realizados na população-alvo;
2. percentual de exames apresentando efeito citopático compatível com o vírus do papiloma humano (HPV);
3. percentual de exames apresentando neoplasia intraepitelial cervical (NIC I – displasia leve);
4. percentual de exames apresentando neoplasia intraepitelial cervical (NIC II – displasia moderada);
5. percentual de exames apresentando neoplasia intraepitelial cervical (NIC III – displasia acentuada);
6. percentual de exames apresentando carcinoma escamoso invasivo;
7. percentual de lâminas consideradas com adequabilidade "satisfatória, mas limitada por";
8. percentual de lâminas consideradas com adequabilidade insatisfatória.

Os autores utilizaram a abordagem hierárquica, com método de Ward, para identificação do número de agrupamentos, e a não hierárquica, pelo método de *k-means*, para confirmação do número de agrupamentos identificados.

As duas abordagens utilizadas permitiram identificar cinco agrupamentos de municípios com perfis semelhantes entre si. Pelo método *k-means* foi possível verificar que a variável percentual de lâminas consideradas com adequabilidade "satisfatória, mas limitada por" foi a que melhor discriminou os grupos. A análise dos resultados revelou que o grupo 1 apresentava 70% dos exames realizados concentrados nos municípios do centro-sul do estado; o grupo 2 concentrava 26% dos exames realizados nos municípios das regiões sul/sudeste do estado; o grupo 3 concentrava 20% dos exames realizados nos municípios das regiões sul/sudoeste do estado; o grupo 4 concentrava 23% dos exames realizados nos municípios da região norte; e o grupo 5 concentrava 23% dos exames realizados nos municípios da região do Jequitinhonha e leste do estado.

Análise de correspondência

A análise de correspondência tem por objetivo analisar um conjunto de dados categóricos a partir das similaridades ou dimissilaridades dos elementos que podem ser visualizadas em um gráfico. Esse método não requer suposições a respeito das distribuições dos dados e possibilita identificar relações que não seriam facilmente percebidas em uma análise bivariada com variáveis categóricas[1,6,23,24].

As etapas analíticas da análise de correspondência envolvem a padronização do conjunto de dados categóricos, o cálculo da matriz de distância, a definição do número de dimensões a serem analisadas e a análise dessas dimensões[23,24]. A padronização do conjunto de dados categóricos consiste em transformar os dados de contagem das variáveis categóricas em frequências relativas, que irão fornecer os perfis de linhas e colunas. Esses perfis representam as respostas dos indivíduos a cada uma das variáveis consideradas na análise e são denominados valores de frequência relativas das linhas e colunas[1,23,24].

Um conceito importante em análise de correspondência é o de inércia, que envolve tanto a representatividade dos objetos analisados com relação aos seus respectivos perfis de linhas ou colunas, como também a distância que esses apresentam a um centroide (que é o perfil médio das linhas ou das colunas). O cálculo da inércia é realizado pela decomposição da estatística χ^2 e possibilita a formação de uma nuvem de pontos que poderão ser representados em um gráfico[23,24].

A análise gráfica da nuvem de pontos permite a identificação de padrões e de relações entre as variáveis consideradas[6,23,24]. Os pontos localizados próximos à origem do gráfico indicam que as variáveis ou categorias apresentam baixas associações entre si, enquanto que os pontos mais afastados, como também os próximos um dos outros, indicam que as variáveis ou categorias apresentam maior associação entre si[23,24]. A aglomeração de pontos dentro de um mesmo quadrante no gráfico indica a presença de categorias ou variáveis com características comuns[23,24].

A análise de correspondência foi utilizada por Aranha[24] para caracterizar o perfil de mulheres na pós-menopausa e o uso da terapia de

reposição hormonal. Foram analisadas respostas de 195 mulheres com relação aos seus aspectos socioeconômicos, demográficos, saúde geral e saúde reprodutiva.

A pesquisadora verificou que as duas primeiras dimensões da análise de correspondência concentraram 37% da variabilidade, sendo que a primeira dimensão sofreu maior influência das variáveis relativas ao nível socioeconômico, enquanto a segunda dimensão recebeu influência das variáveis relativas ao planejamento familiar e cuidado reprodutivo. A análise do gráfico revelou que o primeiro quadrante correspondia às mulheres não usuárias de terapia de reposição hormonal e que essas apresentam índice de massa corporal (IMC) superior a 25 km/m^2, ganho ponderal superior a 20 kg, idade entre 52 e 67 anos, menopausa acima de 48 anos, renda e escolaridade baixa. O segundo quadrante agrupou as mulheres usuárias de terapia de reposição hormonal, com idade entre 38 e 52 anos, cor branca, IMC entre 21 e 22,9 kg/m^2 e ganho ponderal inferior a 10 kg.

O que deve ser considerado antes da realização de uma análise multivariada?

Independente do método multivariado selecionado, alguns aspectos importantes devem ser considerados sob risco de prejudicar a confiabilidade e a validade dos resultados. O tamanho da amostra, por exemplo, sempre é uma questão complexa e difícil de ser solucionada em uma análise multivariada[1]. Um tamanho de amostra muito pequeno pode comprometer o poder de estimação dos parâmetros e resultar em ajustes pouco confiáveis e não generalizáveis. Por outro lado, amostras muito grandes podem resultar em modelos muito sensíveis e poucos práticos[2]. Uma estratégia é calcular a amostra para cada uma das variáveis que serão consideradas na análise e, a partir de algum critério de conveniência do pesquisador, definir o tamanho final da amostra entre o menor e o maior valor calculado[2]. Outra estratégia seria considerar dez observações para cada uma das variáveis incluídas na análise. No entanto, essa regra pode ser insuficiente para tratar problemas com muitas variáveis categóricas[25].

Outro problema bastante comum em uma análise multivariada, e que pode afetar seriamente a consistência dos parâmetros de boa parte dos métodos multivariados, é a ocorrência de dados faltantes (também conhecidos como dados ausentes, dados perdidos ou *missing values*)[1,5,19]. As soluções para lidar com esse tipo de problema são diversas, mas precisam ser cuidadosamente avaliadas antes de serem implementadas devido aos impactos que podem causar nos resultados[1,5,19].

A solução mais simples para lidar com dados faltantes é a exclusão dos registros ou das variáveis que apresentarem esse tipo de problema da análise[1]. No entanto, dependendo da frequência de ocorrência, o tamanho da amostra resultante pode comprometer a confiabilidade dos parâmetros do método utilizado (no caso da exclusão de registros) ou pode-se perder informação sobre alguma relação relevante (no caso de exclusão de variáveis)[1]. Outra solução pode ser a substituição dos valores faltantes pela média ou mediana das observações[1]. Esse método é restrito a variáveis numéricas (ou pelo menos que estejam em uma escala intervalar) e pode aumentar, artificialmente, a homogeneidade da variável, podendo gerar um viés importante na estimativa do parâmetro[26]. Por fim, pode-se utilizar métodos de imputação de dados, por meio de modelos de regressão ou de métodos de simulação, para completar os registros faltantes[1]. Os métodos de imputação costumam ser bastante eficientes no tratamento dos dados faltantes, principalmente quando esses apresentam um padrão de ocorrência aleatório[1].

Dados discrepantes, atípicos, ou *outliers* também são bastante comuns em uma análise multivariada e devem ser analisados com bastante cuidado antes de serem excluídos, pois muitas vezes o seu valor representa um padrão não usual ou uma tendência, não sendo, dessa forma, falha de registro[1].

O tamanho da amostra, os dados faltantes e os dados discrepantes têm um impacto direto nos métodos de estimação, sendo recomendável a realização de análises da robustez, de modo a avaliar o quanto esses parâmetros estão sendo afetados por esses problemas. Essa análise pode ser feita a partir de análises estratificadas, no entanto, o tamanho da amostra pode ser um complicador importante e métodos como *cross-validation* e reamostragem podem ser estratégias interessantes para a análise da robustez do método multivariado utilizado.

Os pacotes estatísticos como, por exemplo, SAS®, SPSS®, PSPP®, STATA®, S-PLUS®, R®, entre outros, apresentam rotinas que permitem avaliar a robustez dos parâmetros e são ferramentas importantes para a realização de análises multivariadas[1,2,5,6,19]. Muitas vezes, a escolha do pacote estatístico mais adequado passa por critérios subjetivos, mas sempre deve-se levar em consideração questões como:

1. custo: existem opções gratuitas e outras que podem custar alguns milhares de dólares, dependendo da configuração e do número de opções do pacote selecionado;
2. nível de conhecimento do usuário: determinados pacotes estatísticos são específicos para determinadas análises, enquanto outros são mais genéricos;
3. capacidade de processamento e de armazenamento: geralmente os pacotes estatísticos suportam grandes volumes de dados, porém, algumas outras soluções vão demandar uma capacidade de processamento e de armazenamento superior ao que normalmente se utiliza;
4. interatividade: geralmente essa questão influencia na facilidade de uso e no custo do *software*;

5. facilidade de programação: essa opção visa permitir ao usuário otimizar tarefas repetitivas de importação e tratamento dos dados.

Um ponto em comum nos pacotes estatísticos é a ênfase na significância estatística, no entanto, os resultados também devem ser avaliados sob o ponto de vista da significância prática, uma vez que resultados estatisticamente significativos podem não ter relevância prática ou clínica[1,27]. Isso porque a significância estatística pode ter sido observada devido a um tamanho de amostra muito grande ou a uma pequena variabilidade entre as observações[1,27].

Finalmente, independente do método multivariado selecionado, todas os métodos multivariados demandam uma etapa inicial que é o desenvolvimento do problema de pesquisa em termos conceituais, de forma a identificar os principais aspectos que descrevem um determinado evento estudado. Essa etapa visa a aprimorar a compreensão do problema estudado, identificar as relações entre exposição e desfecho, fatores de confundimento e variáveis que serão consideradas, visando a minimizar a probabilidade do que se entende por erro de especificação.

A elaboração do modelo conceitual também tem por finalidade reduzir a chance de inclusão indiscriminada de variáveis que pouco acrescentariam no entendimento do problema em questão. A inclusão dessas variáveis pode mascarar os verdadeiros efeitos devido a fenômenos como superajustamento e multicolinearidade que, de um modo geral, aumentam a probabilidade de que qualquer efeito de qualquer variável poderia ser previsto ou explicado pelo conjunto de variáveis do banco de dados.

REFERÊNCIAS

1. Hair JF, Anderson RE, Tatham RL, Black WC. Análise multivariada de dados. Porto Alegre: Bookman; 2005. 600 p.
2. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. 3rd ed. Cambridge: Cambridge University Press; 2011. 233 p.
3. Reboldi G, Angeli F, Verdecchia P. Multivariable analysis in cerebrovascular research: practical notes for the clinician. Cerebrovascular Dis. 2013;35(2):187-93. DOI: 10.1159/000345491
4. Jackson J. Multivariate techniques: advantages and disadvantages. [Internet]. 2015. [cited 2015 Out 9]. Available from: http://www.ehow.com/info_8247893_multivariate-techniques-advantages-disadvantages.html
5. Hunt L. Multivariable analysis. Lancet. 1996;348(9033):1017-8. DOI: 10.1016/S0140-6736(05)64926-4
6. Mingoti SA. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora UFMG; 2005.
7. Possoli S. Técnicas de análise multivariada para avaliação das condições de saúde dos municípios do Rio Grande do Sul, Brasil. Rev Saúde Pública. 1984;18(4):288-300.
8. Monteiro C, Ladeira R, Silva B. A utilização de técnicas multivariadas na análise do posicionamento e segmentação de empresas do sistema suplementar de saúde. In: Simpósio de Excelência e Gestão em Tecnologia. [Internet]. [citado 2017 Out 10]. Disponível em: http://www.aedb.br/seget/arquivos/artigos07/1324_servicos_fatorial%20e%20cluster.pdf
9. Alves LB, Belderrain MC, Scarpel RA. Tratamento multivariado de dados por análise de correspondência e análise de agrupamentos. In: Anais do 13º Encontro de Iniciação Científica e Pós-Graduação do ITA. José dos Campos, SP, Brasil. [Internet]. 2007. [citado 2017 Out 10]. Disponível em: http://www.bibl.ita.br/xiiiencita/MEC17.pdf
10. Hidalgo B, Goodman M. Multivariate or multivariable regression? Am J Public Health. 2013;103(1):39-40. DOI: 10.2105/AJPH.2012.300897
11. Núñez E, Steyerberg EW, Núñez J. Regression modeling strategies. Rev Esp Cardiol. 2011;64(6):501-7. DOI: 10.1016/j.rec.2011.01.017
12. Costa DC, Sá MJ, Calheiros JM. The effect of social support on the quality of life of patients with multiple sclerosis. Arq Neuropsiquiatr. 2012;70(2):108-13.
13. Silva MG, Almeida RT, Bastos EA, Nobre FF. Determinants of cellular atypia detection in the cervical screening program in Rio de Janeiro, Brazil. Rev Panam Salud Publica. 2013;34(2):107-13.
14. Bustamante-Teixeira MT, Faerstein E, Latorre MR. Técnicas de análise de sobrevida. Cad Saúde Pública. 2002;18(3):579-94. DOI: 10.1590/S0102-311X2002000300003
15. Carvalho M de S. Análise de sobrevida: teoria e aplicações em saúde. Rio de Janeiro: Editora Fiocruz; 2005.
16. Botelho F, Silva C, Cruz F. Epidemiologia explicada – análise de sobrevivência. Acta Urológica. 2009;26(4):33-8.
17. Ayala AL. Sobrevida de mulheres com câncer de mama, de uma cidade no sul do Brasil. Rev Bras Enferm. 2012;65(4):566-70. DOI: 10.1590/S0034-71672012000400003
18. Moreira AC. Comparação da análise de componentes principais e da CATPCA na avaliação da satisfação do passageiro de uma transportadora aérea. Inv Op. 2007;27(2):165-78.
19. Manly BF. Multivariate statistical methods: a primer. 3rd ed. Florida: Chapman & Hall/CRC; 2005. 214 p.
20. Oliveira KS, Silva DO, Souza WV. Barreiras percebidas por médicos do Distrito Federal para a promoção da alimentação saudável. Cad Saúde Colet. 2014;22(3):260-5. DOI: 10.1590/1414-462X201400030007
21. Pichiliani M. Data mining na prática: algoritmo K-Means. 2006. [Internet]. [citado 2017 Out 10]. Disponível em: http://imasters.com.br/artigo/4709/sql-server/data-mining-na-pratica-algoritmo-k-means
22. Feitosa TM, Almeida RT. Perfil de produção do exame citopatológico para controle do câncer do colo do útero em Minas Gerais, Brasil, em 2002. Cad Saúde Pública. 2007;23(4):907-17. DOI: 10.1590/S0102-311X2007000400018
23. Infantosi AF, Costa JC, Almeida RM. Análise de correspondência: bases teóricas na interpretação de dados categóricos em ciências da saúde. Cad Saúde Pública. 2014;30(3):473-86. DOI: 10.1590/0102-311X00128513

24. Aranha RN, Faerstein E, Azevedo GM, Werneck G, Lopes CS. Análise de correspondência para avaliação do perfil de mulheres na pós-menopausa e o uso da terapia de reposição hormonal. Cad Saúde Pública. 2004;20(1):100-8.
DOI: 10.1590/S0102-311X2004000100024

25. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.

26. Melo WA, Carvalho MD. Análise multivariada dos fatores de riscos para prematuridade no sul do Brasil. Rev Eletronica Gestão Saúde. [Internet]. 2014 [citado 2017 Out 10]; 5: 398-9. Disponível em: http://www.convibra.com.br/upload/paper/2013/79/2013_79_7817.pdf

27. Loureiro L, Gameiro M. Interpretação crítica dos resultados estatísticos: para lá da significância estatística. Rev Enf Ref. 2011; serIII(3):151-62.
DOI: 10.12707/RIII1009