SMNE UFPR
Simpósio de Métodos Numéricos em Engenharia
25 a 27 de outubro, 2017

Dados faltantes em análises: uma revisão sobre métodos estatísticos flexíveis a incompletude

Melissa Mello de Carvalho
Universidade Federal do Paraná - Programa de pós-graduação em Bioinformática, Curitiba, Brasil.
email: melissa.mdecarvalho@gmail.com

Resumo— **Dados faltantes são um problema para a análise de informações, reconhecimento de padrões e tomada de decisão em todas as áreas do conhecimento. A complexidade do problema não afeta somente a extração de informações, como a aplicação de vários métodos estatísticos que necessitam de completude. Este trabalho tem o objetivo de apresentar de uma breve revisão a respeito da complexidade não trivial de causas, características e consequências da incompletude bem como apresentar métodos estatísticos flexíveis e robustos à análise de dados com lacunas.**

Palavras-chave— incompletude; inferência estatística; imputação; testes não paramétricos;

I. INTRODUÇÃO

Dados faltantes são um problema à análise de informações em todas as áreas do conhecimento. Incompletude é o termo que designa a falta de dados em uma amostra, conjuntos de dados incompletos, ou lacunas de informação. A incompletude interfere na análise estocástica, e pode ser devida a vários mecanismos de perda de informações e de erros nas seleções de informações por equipamentos ou por erros humanos. As lacunas na matriz de dados podem apresentar consequências às respostas obtidas. A análise incompleta promove impactos e podem causar dependências que afetam, consequentemente, a observação de características da amostra e a tomada de decisão, gerando incertezas quanto aos resultados obtidos [1]e[2]. Na estatística inferencial a incompletude produz consequências que se revelam pelo distanciamento de resultados na incompletude em relação a completude: a incompletude produz resultados incompletos ou "parciais" em relação a completude; a seleção não intencional de dados provocada pela incompletude pode gerar respostas com desvios, tendências e erros interpretativos em relação aos resultados com completude. A incompletude também reduz a precisão de resultados [1]. Esses fatores podem contribuir para má interpretação de estatísticas resultantes [3] e consequentes decisões errôneas, além representar uma barreira a aplicação de estatísticas descritivas. Dados com disposição longitudinal são geralmente mais robustos à incompletude em comparação com dados onde há somente uma observação no tempo, de disposição transversal. O efeito danoso da incompletude em estudos longitudinais se deve a "quebra" do vínculo causal entre intervenção e resposta, a não medição de uma série de respostas pode levar a conclusões enganosas [4].

No que diz respeito à taxonomia, a incompletude é dividida em duas formas: apresentação e quantificação da incompletude em variáveis, e na classificação de Rubin (1976) referente ao tipo randômico da incompletude e seus critérios de independência e dependência de dados observados e não observados. As formas taxonômicas da descrição de incompletude são listadas abaixo:

Quanto à apresentação da incompletude em variáveis:

a) *Falta não monótona: a incompletude ocorre em algumas variáveis e para alguns casos.*

b) Falta monótona: incompletude ocorre em uma variável em todos os casos, pode se tratar também de uma variável latente não observada.
c) Falta por unidade sem resposta: incompletude em todas as variáveis para alguns casos [5].

Quanto à mecanismos geradores da incompletude:

a) Incompletude completamente randômica (MCAR): onde a probabilidade de um valor estar faltando é independente de dados observados e não observados;
b) Incompletude randômica (MAR): gera um condicionamento nos dados observados, a probabilidade e os valores em perda são dependentes dos dados não observados;
c) Incompletude não randômica (MNAR): em que a probabilidade do valor estar ausente depende de dados observados e não observados [6].

II. MÉTODOS APLICÁVEIS À ANÀLISE COM INCOMPLETUDE

A maioria dos métodos estatísticos, computacionais e softwares estatísticos necessitam de completude para realizar medições [2]e[5]. Há diversos métodos que visam possibilitar análises estatísticas na presença de incompletude criando completude artificialmente. São eles: imputação simples, imputação múltipla, imputação por último caso observado [2]e[4], e exclusão de lacunas com consequente exclusão de informações associadas ou dependentes, são eles: *listwise* (exclusão de linhas na forma tabular) *pairwise* (exclusão de informações com dependência em relação às informações faltantes) [2] e *casewise* (exclusão de casos) [3]. Dependendo do tamanho da incompletude, métodos de exclusão podem reduzir drasticamente o tamanho da amostra por excluírem lacunas e suas dependências, o que pode resultar em desvios e perda de padrões [4]. A imputação é realizada pela aplicação de métodos estatísticos e não estatísticos diversos: imputação por média, imputação por correlação, imputação baseada em regressão linear, entre outras. De acordo com Burke (2001), pacotes de softwares estatísticos geralmente trabalham com as lacunas pelos seguintes métodos: *casewise, listwise, pairwise*, e *mean substitution* (imputação pela média) [3]. Dentre os métodos citados os mais populares entre pesquisadores são os de imputação e recuperação de informações, pois descartar dados implica no desperdício de recursos [5].

São considerados robustos a incompletude os métodos que não necessitam de completude para o cálculo, são eles: estimativa maximizada [5], análise de máxima verossimilhança [7]; análise probabilística parcial [4]e[5]; métodos que ignoram lacunas para fins de cálculo, alguns métodos de correlação não paramétrica [4], [5]e[7]. Os métodos não paramétricos que ignoram lacunas e são flexíveis à incompletude são exemplificados a seguir.

A. Métodos paramétricos

Métodos paramétricos não são aptos à falta de informações, pois necessitam de completude para a realização de cálculos. Algoritmos baseados em métodos paramétricos necessitam obrigatoriamente da aplicação de métodos de imputação ou de exclusão de lacunas para a obtenção de resposta [3]. Assim, a escolha de métodos de imputação tem importância na qualidade das respostas de estatísticas [1]. A imputação pela média, por exemplo, pode alterar resultados obtidos por regressão linear produzindo correlações mais fortes [3]. Alguns softwares substituem lacunas por zeros para a aplicação de métodos paramétricos. A substituição por zeros pode gerar tendência do resultado por zero [8].

B. Metodos não paramétricos

A maioria dos métodos não paramétricos são considerados uma alternativa robusta aos testes paramétricos [3], por não necessitarem de completude. A exceção são métodos baseados em "posto ordem" ou *ranking* [7] como: Mann-Whitney, Kruskal-Wallis, Kendall e Spearman posto ordem, entre outros. A aplicação de métodos posto, tal como no caso dos paramétricos, requer imputação ou exclusão de lacunas, o que pode gerar problemas relacionados com a escolha do método de imputação ou de exclusão de lacunas. São flexíveis à incompletude métodos não paramétricos de correlação: correlação tau de Kendall [3] por exemplo, onde a ordem com que os dados se apresentam não é empecilho à resposta. A matriz de correlação requer a disposição de todos os dados para a obtenção dos valores de correlação, em métodos como correlação tau de Kendall que afere força de correlação, as lacunas são ignoradas e a força de correlação é proporcional aos dados presentes, mas para isso é necessário que a variável que classifica a população de pertencimento de dados esteja em completude. Testes de sensibilidade são considerados critério de confiabilidade das respostas para métodos não paramétricos robustos à incompletude [4].

III. DISCUSSÕES

A dependência entre dados observáveis e não observáveis e a quantidade de lacunas são fatores de grande importância para a confiabilidade de análises com dados incompletos. Métodos de exclusão e imputação de dados aplicados sem o devido cuidado podem ocasionar desvios e vieses interpretativos mais danosos aos resultados do que teria a incompletude não tratada analisada por métodos flexíveis. Métodos não paramétricos não posto ordem possuem flexibilidade à incompletude e podem ser uma alternativa a exploração do conhecimento incerto sem a necessidade da criação de informações artificialmente. De acordo com Karpenter e Kenward (2007), não há método universalmente aplicável para tratamento da incompletude. Recomenda-se estudar a forma da incompletude e aplicar métodos com cautela se possível testando vários métodos. O estímulo a discussão de efeitos da incompletude tem importância informativa à pesquisadores, e a declaração da presença de incompletude e formas de tratamento das lacunas em pesquisas científicas deve ser incentivada como forma de estabelecer a confiabilidade dos resultados encontrados.

REFERÊNCIAS

[1] A. WOOD, I. WHITE, S. THOMPSON. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. SAGE Clinical Trials, n.1, p.368-76, ago. 2004.

[2] G. MOLENBERGHS et Al. Analyzing Incomplete Longitudinal Clinical Trial Data. Biostatistics. Biometrika Trust, v.5, n.3, p.445-64, jul, 2004.

[3] S. BURKE. Missing values. outliers. robust statistics & non-parametric methods. LC-GC Europe Online Supplement, Statistics & Data Analysis. Buckinghamshire, UK. v.2, n.0, p.19-24, 2001.

[4] J. KARPENTER, M. KENWARD. Missing data in randomized controlled trials — a practical guide, London School of Hygiene & Tropical Medicine, London, UK, Spring, 2007.

[5] P. ALLISON. Missing data. Design and Inference. Sage University Papers Series on quantitative applications in the social sciences. Thousand Oaks (CA). Cap. 4, p.72-89, 2001.

[6] D. RUBIN. Inference and missing data. Biometrika, n.63, p.581-592, dec. 1976.

[7] S. SIEGEL, J. CASTELLAN. Estatística Não-Paramétrica para Ciências do Comportamento, Métodos de Pesquisa, Artmed, Bookman, 2º ed, p. 287. 294, p. 318, 325-326, 2006.

[8] P. BRACARENSE COSTA. Um enfoque segundo a teoria de conjuntos difusos para a meta-análise. f.155. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina (UFSC), Florianópolis, 1999.