Imputação de dados ausentes e análise não paramétrica de perfis bioquímicos e hematológicos em serpentes da espécie *Crotalus Durissus Terrificus*

Selene Maria Coelho Loibel [1], **Sabrina Gonçalves Cardoso** [2], **Rafael Amorim de Castro** [3], **Kathleen Fernandes Grego** [4]

Introdução

Segundo relatório SINITOX de 2018, em 2017 cerca de 1600 acidentes ofídicos foram registrados no Brasil O gênero *Crotalus* é responsável por $7,7\%$ dos acidentes, *Lachesis* por $1,4\%$ e *Bothrops* por $90,5\%$ sendo que o envenenamento crotálico possui a maior taxa de letalidade: $1,87\%$ (MS/FIOCRUZ/SINITOX, 2018). A manutenção de serpentes em cativeiro tornou-se uma alternativa para a obtenção do veneno para pesquisas imunológicas e produção de soro antiofídico. O Instituto Butantan, conta com o Laboratório de Herpetologia que desde a década de 1960 mantém o programa de criação de serpentes em sistema intensivo (GREGO e CASTRO, 2015).

Existem inúmeras dificuldades em relação ao diagnóstico precoce, prevenção e controle de doenças infecciosas em serpentes. Estes problemas estão diretamente relacionados à dificuldade de adaptação ao cativeiro. Serpentes provenientes da natureza estão amplamente sujeitas a diversos agentes estressantes relacionados ao cativeiro, que enfraquecem o sistema imunológico dos animais, levando-os a desenvolver doenças. As infecções parasitárias são mais comumente observadas em animais de natureza, embora serpentes mantidas por longos períodos em cativeiro possam apresentar alguma forma de doença parasitária (RAMEH-DE-ALBUQUERQUE, 2007 e GREGO e CASTRO,2015).

Para contribuir com estudos de pesquisadores do Laboratório de Herpetologia do Instituto Butantan, nesse trabalho foram utilizados dados coletados neste laboratório. Foi feita a comparação do perfil hematológico e bioquímico de serpentes, tratadas e não tratadas com vermífugo, da espécie *Crotalus durissus terrificus* utilizando métodos não paramétricos. Antes da análise dos perfis foi necessária a imputação de dados ausentes, ocorrência comum nesse tipo de conjunto de dados. Para isso foi utilizado um método de imputação múltipla.

Material

Os dados consistem de medidas de parâmetros bioquímicos e hematológicos tomadas em 5 coletas de sangue ao longo do tempo com o objetivo de avaliar o estado de saúde das serpentes em cativeiro no laboratório. Foram utilizadas cascavéis (*Crotalus durissus terrificus*) de ambos os sexos e naturalmente parasitadas para a realização do experimento. Como atualmente existem valores de referência para esses parâmetros somente para serpentes saudáveis (sem parasitas), os animais foram divididos em dois grupos. Grupo G1,

[1]DEMAC, UNESP, Rio Claro. e-mail:*smc.loibel@unesp.br*
[2]Depto de Matemática, UNESP, Rio Claro.
[3]Instituto Butantan, São Paulo.
[4]Instituto Butantan, São Paulo.

composto por 6 individuos que foram vermifugados e Grupo G2, composto por 6 individuos que não receberam vermífugo, de forma que a comparação dos perfis serve também para estabelecer outras referências para os parâmetros.

Os animais passaram por coletas periódicas de sangue e a cada coleta de amostra sanguínea foram avaliados os parâmetros bioquímicos e hematológicos, dos quais, neste trabalho citamos seis:

Hemácias: Células sanguíneas, responsáveis pela condução de oxigênio para os tecidos.

Leucócitos: Conjunto de células sanguíneas, denominadas células brancas. Atuam nos processos inflamatórios e imunológicos.

Trombócitos: Células sanguíneas, análogas as "plaquetas" dos mamíferos. Participam dos processos hemostáticos (mecanismo de coagulação) e auxiliam as células brancas em processos inflamatórios e/ou infecciosos.

Proteínas totais (PT): Associações das duas principais proteínas plasmáticas do organismo: Albumina e globulina.

Alanina amonitransferase (ALT): Enzima associada à lesão hepática (valor aumentado indica rompimento de hepatócitos). Associada a AST, define se a lesão é em fígado ou em músculo.

Fosfatase alcalina (FA): Presente em células ósseas, mucosa intestinal, rim e fígado. Eleva-se quando há lesão óssea ou intensa inflamação/infecção intestinal.

Métodos

Imputação de dados ausentes - algoritmo MICE

No conjunto de dados descrito na seção anterior há ausência de dados em todas as variáveis que representam os parâmetros (Hemácias, Leucócitos, Trombócitos, Proteínas totais, Alanina Amonitransferase e Fosfatase alcalina). O modelo de imputação deve levar em conta o processo que produziu a falta de dados, deve preservar as relações entre as variaveis e a incerteza sobre essas relações. Aseguir é apresentada a formulação da imputação múltipla utilizando o algoritmo MICE (VAN BUURREN, 2012)

Denotar por $Y = (Y_1, ..., Y_p)$ o conjunto de variáveis a ser estudado e $Y_j, j = 1, ..., p$; uma das variáveis incompletas. As partes observadas e não observadas de Y_j são Y_j^O e Y_j^A, respectivamente. Então $Y^O = (Y_1^O, ..., Y_p^O)$ é o conjunto com as partes completas e $Y^A = (Y_1^A, ..., Y_p^A)$ é o conjunto com as partes incompletas. O $h-$ésimo conjunto de dados imputados é denotado por Y^h,onde $h = 1, ..., m$. Seja $Y_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_p)$ e seja Q a quantidade de interesse (por exemplo o coeficiente em um modelo de regressão).

Seja Y seguindo uma distribuicão $p-$variada $P(Y|\theta)$. Assumir que a distribuição multivariada de Y seja completamente especificada por θ, um vetor de parâmetros desconhecidos. O problema é como obter a distribuição de θ. O algoritmo MICE obtém a distribuição a posteriori de θ gerando iterativamente das distribuições condicionais na forma

$$P(Y_1|Y_{-1}, \theta_1), ..., P(Y_p|Y_{-p}, \theta_p)$$

Os parâmetros $\theta_1, ..., \theta_p$ são específicos de cada distribuição condicional que não são necessariamente o produto de uma fatorização da verdadeira distribuição conjunta $P(Y|\theta)$. A $t-$ésima iteração das equações em cadeia é um amostrador de Gibbs que gera sucessivamente

$$\theta_1^{*(t)} \sim P(\theta_1 | Y_1^O, Y_2^{(t-1)}, ..., Y_p^{(t-1)})$$

$$Y_1^{*(t)} \sim P(Y_1 | Y_1^O, Y_2^{(t-1)}, ..., Y_p^{(t-1)}, \theta_1^{*(t)})$$

.....

$$\theta_p^{*(t)} \sim P(\theta_p | Y_p^O, Y_1^{(t)}, ..., Y_{p-1}^{(t)})$$

$$Y_p^{*(t)} \sim P(Y_p | Y_p^O, Y_1^{(t)}, ..., Y_p^{(t)}, \theta_p^{*(t)})$$

sendo que $Y_j^{(t)} = (Y_j^O, Y_j^{*(t)})$ é a $j-$ésima variável imputada na iteração t. Observa-se que as imputações anteriores $Y_j^{*(t-1)}$ somente entram em $Y_j^{*(t)}$ por meio das suas relações com as outras variáveis e não diretamente. A convergência portanto é bem rápida, com número de iterações entre 10 e 20, diferentemente de em outros métodos de MCMC. O nome "imputação múltipla por equações em cadeia" se refere ao fato de que o algoritmo MICE pode facilmente ser implementado como uma concatenação de procedimentos univariados para preencher conjuntos de dados incompletos.

Análise não paramétrica dos perfis

O experimento apresentado pode ser classificado como do tipo F1-LD-F1, composto por 2 grupos diferentes de indivíduos, cada um recebendo tratamento diferente e observados repetidamente ao longo do tempo (NOGUCHI *et al*, 2012). A análise estatística dos perfis desses grupos e sua comparação é feita com 3 testes de hipóteses não paramétricos nos quais temos como hipóteses nulas:

Teste 1: H0.: Não há diferença entre os perfis dos grupos;

Teste 2: H0.: Não há diferença entre as medidas nas coletas dentro do grupo (não há efeito do tempo);

Teste 3: H0.: Não há efeito de interação entre grupo e coleta (tempo).

Supor que grupos diferentes de indivíduos homogêneos são observados repetidamente em diferentes pontos t de tempo e cada grupo recebe um tratamento distribuído aleatoriamente (tratamento 1, o tratamento 2, ..., tratamento a). O modelo estatístico pode ser descrito por vetores aleatórios independentes $X_{ik} = (X_{ik1}, ..., X_{ikt})^T, K = 1, ..., n_i$, com distribuições marginais $X_{iks} \sim F_{is}, i = 1, ..., a; s = 1, ..., t$. O número total de observações é dado por $N = n \cdot t$, sendo $n = \sum_{i=1}^{a} n_i$.

As hipóteses de nenhum efeito principal A, nenhum efeito do tempo principal T, e nenhuma interação (AT) entre A e T, são expressas em termos das funções de distribuição marginais:

$H_0^F(A) : \overline{F}_{1.} = ... = \overline{F}_{a.}$

$H_0^F(T) : \overline{F}_{.1} = ... = \overline{F}_{.t}$

$H_0^F(AT) : F_{is} = \overline{F}_{i.} - \overline{F}_{.s} + \overline{F}_{..}, i = 1, ..., a; s = 1, ..., t$

sendo que $\overline{F}_{i.} = \frac{1}{t}\sum_{s=1}^{t} F_{is}$ denota a distribuição média ao longo do tempo para o grupo de tratamento $i, i = 1, ..., a$, $\overline{F}_{.s} = \frac{1}{a}\sum_{i=1}^{a} F_{is}$ denota a distribuição média ao longo dos grupos de tratamento para o ponto de tempo $s, s = 1, ..., T$, e $\overline{F}_{..} = \frac{1}{at}\sum_{i=1}^{a}\sum_{s=1}^{t} F_{is}$ denota a distribuição média geral.

As hipóteses para os modelos longitudinais lineares clássicos (paramétricos) são expressas da mesma forma com μ_{is}. Para uma discussão de formulação de hipóteses por funções de distribuição, consultar o artigo de Akritas e Arnold (AKRITAS e ARNOLD, 1994)

Resultados e Discussão

Primeiramente, foi verificada a não normalidade dos dados, em seguida foi feita a imputação de dados ausentes e então foram feitos análises dos perfis dos dois grupos para comparação. Os cálculos foram feitos com o software livre R (R CORE TEAM, 2012), sendo utilizados os pacotes MICE para imputação e NparLD para a análise não paramétrica dos perfis.

As Figuras 1 e 2 apresentam os resultados da imputação múltipla (5 imputações) utilizando o algoritomo MICE com a opção do método "pmm: predictive mean matching". Os pontos em azul são os dados originais e os pontos em vermelho representam os valores imputados no lugar dos dados ausentes. Variáveis: V1: Coleta e V2: Identificação do indivíduo não tem imputação, já estavam completas; V3: Hemácias, V4: Leucócitos, V5: Trombócitos, V6: Proteínas totais (PT), V7: Alanina amonitransferase (ALT) e V8: Fosfatase alcalina (FA).

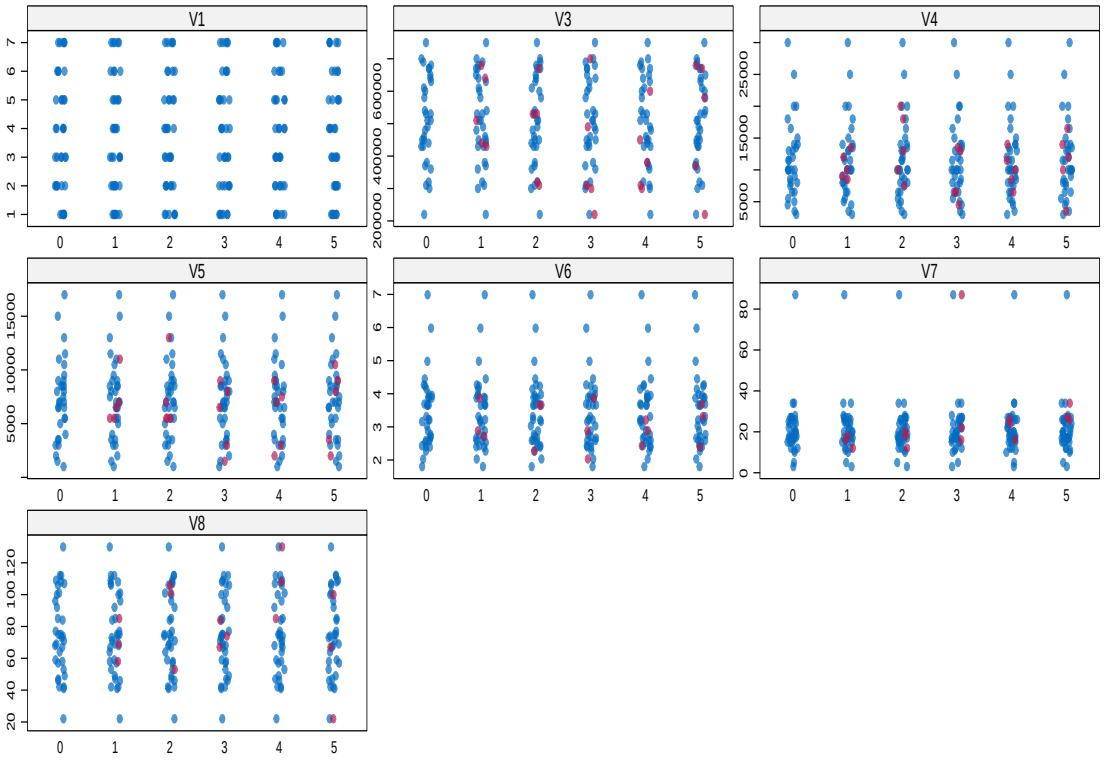

Figura 1: Imputações Grupo 1

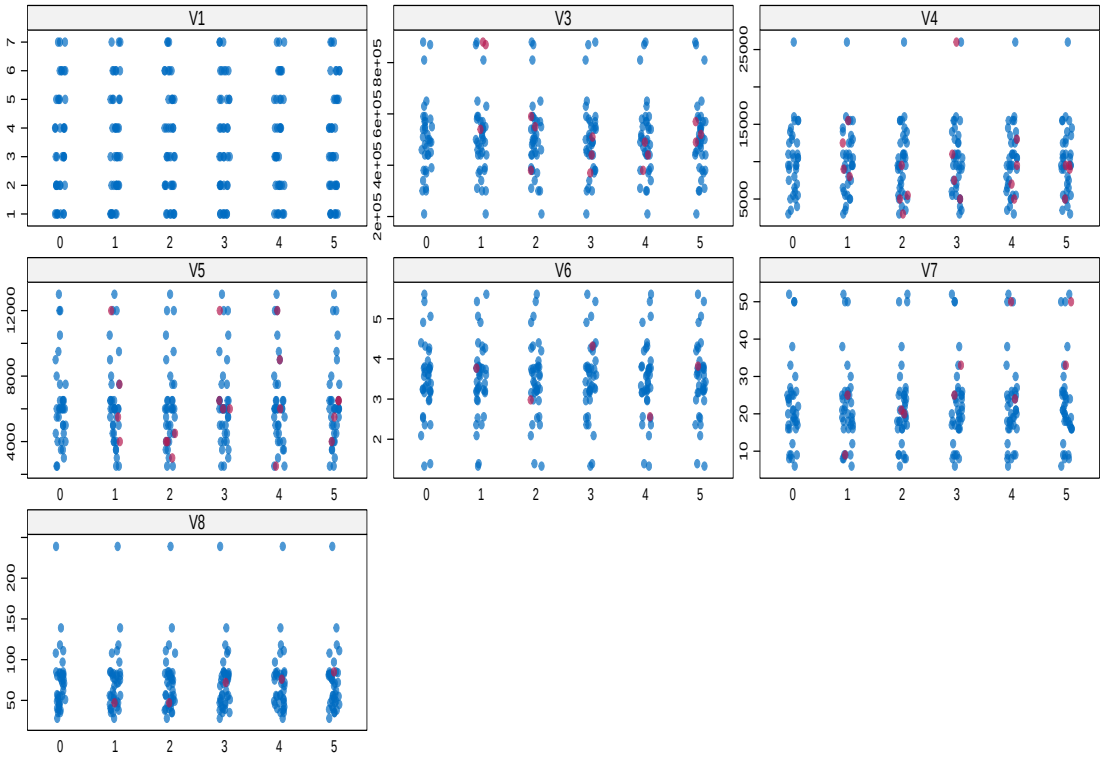

Figura 2: Imputações Grupo 2

A metodologia de imputação múltipla aplicada nesse trabalho pode ser aprimorada utilizando outra abordagem para dados longitudinais e deverá ser o próximo trabalho deste grupo. Como não são muitos os valores ausentes nesse conjunto de dados, o resultado apresentado é satisfatório.

Os resultados dos 3 testes para as variáveis hemácias, leucócitos e trombócitose fosfa-

tase alcalina mostraram que não há diferença significativa entre grupos (p-$value = 0,7674$ para as 3 primeiras e p-$value = 0,3311$ para última); não há diferença significativa entre coletas, dentro dos grupos (p-$value = 0,4892$ para as 3 primeiras e p-$value = 0,1428$ para última).

Na Tabela 1 são apresentados os resultados dos testes de comparação de perfis que mostraram diferença significativa entre as coletas dentro dos grupos para as variáveis Proteínas Totais (PT) e Alanina amonitransferase (ALT), sendo G-C o teste que verifica se há interação nos efeitos de grupos e coletas.

Tabela 1: Resultados dos testes de comparação de perfis das variáveis PT e ALT

Teste	Grupos PT	Coletas PT	G-C PT	Grupos ALT	Coletas ALT	G-C ALT
Est.	$0,5589$	$9,4905$	$1,6176$	$0,7051$	$5,4348$	$2,1972$
p-value	$0,4547$	$4,5 \times 10^{-6}$	$0,1851$	$0,4011$	$0,0084$	$0,1234$

Com esses resultados, sabe-se que há efeito do tempo nesses parâmetros, independentemente de se tratar de serpentes tratadas ou não com vermífugos. Conclui-se que a metodologia estatística utilizada é viável, destacando que a imputação múltipla pode ser utiizada com outros métodos mais aprimorados.

Referencias Bibliográficas

AKRITAS,M. G.; ARNOLD, S. F. Fully Nonparametric Hypothesis for Factorial Designs I: Multivariate Repeated Measures Designs *Journal of the American Statistical Association*, v.89, 336-343,1994.

GREGO, K. F. ; CASTRO, R.A. Efeitos do endoparasotismo em Crotalus durissus terrificus mantido em cativeiro. *Projeto de pesquisa*, Butantan 2015.

MS/FIOCRUZ/SINITOX *Relatório Anual 2018*.
https://sinitox.icict.fiocruz.br/sites/sinitox.icict.fiocruz.br/files//Brasil8.pdf

NOGUCHI, K. ;.GEL,Y.R.; BRUNNER,E.; KONIETSCHKE,F. nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data *Journal of Statistical Software* v. 50, Issue 12, 2012.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RAMEH-DE-ALBUQUERQUE, L.C. Aspectos hematológicos, bioquímicos, morfológicos e citoquímicos de células sanguíneas em Viperídeos neotropicais dos gêneros Bothrops e Crotalus mantidos em cativeiro. *Tese de* Prog. pós-grad. em patologia experimental e comparada da Fac. de Med. Vet. e Zootec-USP, 2007.

VAN BUUREN, S. Flexible Imputation of Missing Data, Chapman & Hall, 2012.