ISSN: 2317-0840

Modelagem probabilística de dados de pagamentos de provedor de *internet* usando variável mista

Shirlley O. Silva[1], Divanilda Maia[2], Gustavo H. Esteves[2†]

[1] *Graduada em Estatística pela Universidade Estadual da Paraíba.*

[2] *Professor do Departamento de Estatística da Universidade Estadual da Paraíba.*

Resumo: *O objetivo deste trabalho é usar uma variável aleatória mista para modelar dados de pagamentos de provedor de internet feitos pelos clientes de uma empresa de uma cidade da Paraíba. Numa análise de dados, um dos primeiros passos é observar a natureza das variáveis envolvidas e fazer uma análise gráfica delas. Geralmente, essas variáveis podem ser classificadas como discretas ou contínuas. As discretas surgem preponderantemente de contagens, enquanto que as contínuas surgem de medidas. Mas existem ainda as variáveis mistas, que são obtidas fazendo-se uma soma ponderada de variáveis discretas e contínuas. No caso dos dados utilizados, a análise gráfica indicou um comportamento exponencial, que é um modelo contínuo para dados positivos. No entanto, havia uma grande quantidade de valores nulos, de modo que foi levantada a hipótese de se usar uma variável mista, sendo que a parte positiva foi modelada pela distribuição exponencial e os valores nulos por uma variável degenerada no ponto zero.*

Palavras-chave: Dados inflacionados de zeros; distribuição exponencial; variável degenerada em zero.

Abstract: *The objective of this work is to use a mixed random variable to model data from payments made by customers of a Paraíba city. In a data analysis, one of the first steps is to observe the nature of the variables involved and to make a graphical analysis of them. Generally, these variables can be classified as discrete or continuous. The discrete ones arise predominantly from countings, while the continuous ones arise from measures. But there are still the mixed variables, which are obtained by making a weighted sum of discrete and continuous variables. In the case of the data used, the graphical analysis indicated an exponential behavior, which is a continuous model for positive data. However, there was a large amount of null values, which gave rise to the hypothesis of using a mixed variable, where the positive part were modeled by the exponential distribution and the null ones by a degenerate variable at zero point.*

Keywords: Zero infflated data; exponential distribution; zero degenerated variable.

Introdução

Uma variável aleatória X é uma função que tem domínio em um espaço amostral Ω e contradomínio no conjunto dos números reais. De acordo com os valores que a variável aleatória pode assumir, pode-se classificá-la como discreta, contínua ou mista. Há ainda a variável singular, que é contínua em quase toda parte exceto em um conjunto de medida de Lebesgue nula (JAMES, 2002), não podendo ser classificada como discreta, contínua ou mista.

Uma parte importante das análises estatísticas é definir variáveis aleatórias que sejam capazes de mensurar as características de interesse. Depois de feita a avaliação exploratória dos dados, muitas vezes a análise é aprofundada através de inferências. Podem ser construídos intervalos de confiança, realizados testes de hipóteses ou são propostos outros modelos, como os de regressão por exemplo.

[†]Autor correspondente: gesteves@servidor.uepb.edu.br.

Uma etapa intermediária neste processo consiste em fazer suposições sobre um modelo probabilístico que seja adequado à(s) variável(is) em questão. Primeiramente, é preciso saber se será usado um modelo discreto ou contínuo. Boa parte dos dados que aparecem em situações práticas se enquadra nesta situação e então usa-se um modelo conhecido que seja adequado. Em outras situações, observa-se que a variável é formada por uma parte discreta e outra contínua, o que se chama variável aleatória mista (CHANDRA, 1977; WELD; LEEMIS, 2017).

Quando a variável aleatória é mista, ela não tem uma função de probabilidade associada (porque não é discreta), nem uma função densidade de probabilidade, pois é necessário que para cada ponto individual se tenha uma probabilidade zero para uma variável aleatória absolutamente contínua. No entanto, pode-se caracterizar a distribuição da variável aleatória mista a partir da função de distribuição acumulada, que neste caso é composta por partes distintas: uma discreta e outra contínua. Se X é uma variável aleatória mista, sua função de distribuição pode ser especificada através da média ponderada entre uma função de distribuição acumulada de duas variáveis, uma discreta e outra contínua, isto é

$$F(x) = \alpha F^d(x) + (1-\alpha)F^c(x),$$

onde $0 < \alpha < 1$ será o peso dado a cada componente da mistura e F^d e F^c são as distribuições do tipo discreta e contínua, respectivamente (CHANDRA, 1977).

Em muitas situações práticas muitos zeros aparecem para variáveis que tecnicamente deveriam assumir valores no intervalo positivo da reta real, ou seja, estritamente positivos e contínuos, o que pode trazer problemas na modelagem de tais variáveis, e o uso de variáveis mistas pode ajudar em situações assim.

Deste modo, este artigo tem o objetivo principal de apresentar o uso de uma variável mista para modelar um conjunto de dados referente a pagamentos realizados a um provedor de *internet*, bem como apresentar alguns resultados descritivos desta modelagem.

Material e métodos

O conjunto de dados usado neste trabalho é referente a pagamentos feitos por clientes de um provedor de *internet* de uma cidade da microrregião do Brejo Paraibano (Figura 1), que fica inserida na mesorregião do Agreste do estado, englobando ao todo oito municípios.

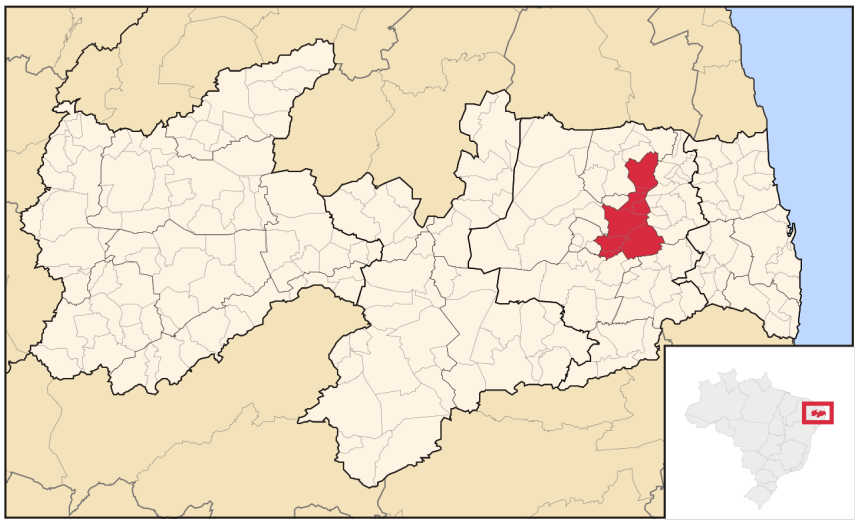


Figura 1: Região do Brejo no estado da Paraíba destacada em vermelho no mapa.
Fonte: Wikipedia (2020).

No total foram obtidos os valores de 302 pagamentos, em reais, feitos a uma empresa que fornece serviços de acesso à *internet* banda larga através de cabos coaxiais ou metálicos, e fibras óticas, sendo o preço estabelecido conforme a quantidade de *megabytes* contratada pelo cliente. Todas as informações foram obtidas durante o mês de março de 2017.

Foi usada uma variável aleatória mista para modelar a variável de interesse, referente aos pagamentos realizados, com a parte contínua sendo representada por uma distribuição exponencial e a parte discreta representada por uma variável degenerada no ponto zero.

Neste contexto, em princípio pode-se pensar em uma variável aleatória com distribuição de Bernoulli onde a probabilidade de sucesso p está associada à adimplência dos clientes, ou seja, $Y \sim \text{Bernoulli}(p)$ (ROSS, 2010) tal que

$$Y = \begin{cases} 0, & \text{se o cliente era inadimplente,} \\ 1, & \text{se o cliente era adimplente;} \end{cases}$$

e como a probabilidade de sucesso é p, a distribuição da variável aleatória Y fica dada da seguinte maneira

$$P(Y = y) = p^y(1-p)^{1-y},$$

com $y \in \{0, 1\}$ e $p \in [0, 1]$.

Dessa forma, esta variável de Bernoulli foi adaptada para a construção de uma variável mista, em que a parte discreta, X_d, é dada por uma variável degenerada em zero (equivalente a $Y = 0$, para os clientes inadimplentes) e a parte contínua é dada por uma variável com distribuição exponencial. Assim, considera-se uma variável aleatória X_c exponencialmente distribuída, ou seja, $X_c \sim \text{Exp}(\lambda)$ (ROSS, 2010), tal que sua função de distribuição acumulada é dada por

$$F^c(x) = \begin{cases} 0, & \text{se} \quad x \leq 0, \\ 1 - e^{-\lambda x}, & \text{se} \quad x > 0. \end{cases}$$

Tal ideia está representada graficamente pela ilustração da Figura 2. Se a variável Y assume o valor 0, tem-se a situação dos clientes inadimplentes, cujo valor pago é igual a 0. Quando $Y = 1$ estão sendo considerados os clientes que realizaram efetivamente algum pagamento, sendo que o valor pago será modelado por uma variável aleatória $X_c \sim \text{Exp}(\lambda)$.

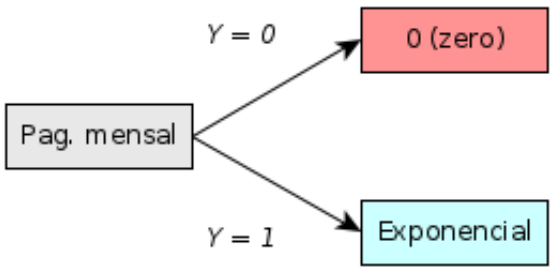


Figura 2: Esquema gráfico ilustrando a construção da variável mista a partir de uma variável com distribuição de Bernoulli. Fonte: os autores.

Os valores referentes aos pagamentos dos clientes foram inseridos no *software* R, versão 4.0.3 (R CORE TEAM, 2020). Todos os resultados apresentados na próxima seção foram obtidos diretamente neste programa, sem o uso de nenhum pacote adicional.

Resultados e discussão

A amostra foi composta de 302 observações referentes aos pagamentos de serviços de *internet*, sendo que 220 desses pagamentos foram efetivados, enquanto 82 deles não o foram, implicando em uma quantidade considerável de clientes inadimplentes e, consequentemente, na presença de vários zeros nos dados.

O problema da inadimplência tem sido objeto de estudos no Brasil. Daros e Pinto (2017), por exemplo, fazem uma boa revisão de trabalhos que tratam sobre o problema em diversas localidades do país. Porém, é incomum encontrar trabalhos que tratem desta abordagem de variáveis mistas nestes tipos de trabalhos.

Na Figura 3, pode-se observar dois histogramas dos dados, um com a inclusão dos valores iguais a zero (a) e outro sem estas observações (b). No primeiro histograma (a) observa-se a frequência relativa maior para a primeira categoria da variável em função da presença dos valores nulos, já no segundo (b) a forma do histograma indica que os dados seguem uma distribuição exponencial.

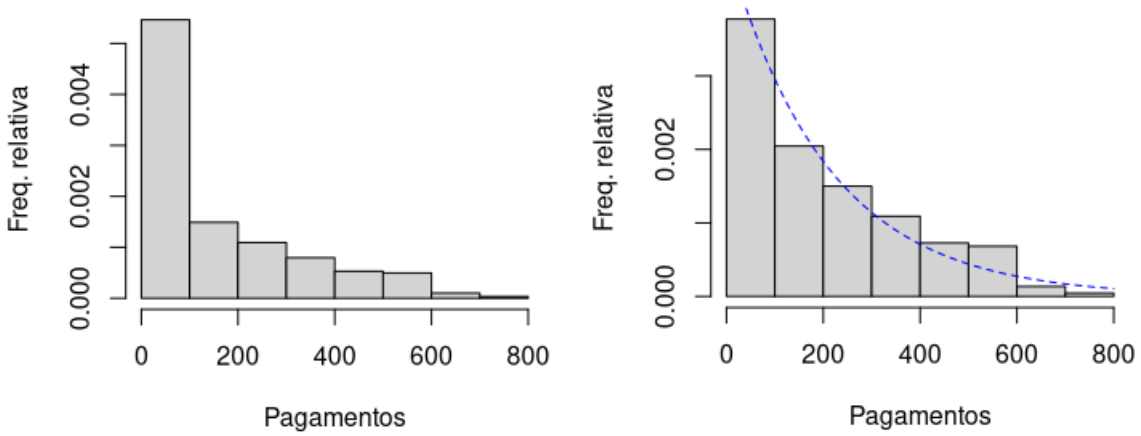


Figura 3: Histogramas dos dados referentes a pagamentos dos serviços de *internet* com a inclusão dos zeros (a) e sem a inclusão dos zeros (b).

Conforme mencionado anteriormente, ao longo do período observado 82 clientes não pagaram, o que representa $27,15\%$ de inadimplência. Porém, quando se observa o histograma sem a inclusão destes valores nulos verifica-se o indício de um comportamento exponencial conforme linha tracejada em azul na Figura 3 (b). Mas quando se considera o conjunto de dados completo, há uma quantidade considerável de zeros, o que atrapalha a modelagem probabilística por meio da distribuição exponencial, dado que ela é definida apenas para valores positivos. Quando se compara a primeira barra (valores menores que 100) entre os gráficos (a) e (b) da Figura 3 vê-se que a frequência no primeiro gráfico é praticamente o dobro da frequência para o mesmo intervalo de variação no segundo histograma, o que mostra o impacto dos valores iguais a zero no banco de dados.

Deste modo, percebe-se que há duas variáveis distintas: uma discreta, degenerada no ponto zero associada aos clientes inadimplentes, e outra contínua e estritamente positiva, referente aos clientes que cumpriram com seus compromissos financeiros junto à empresa. Em outras palavras, a partir da quantidade de dinheiro paga pelo cliente à empresa, é possível se definir uma variável aleatória mista X tal que sua função de distribuição acumulada seja dada por

$$F(x) = \alpha F^d(x) + (1-\alpha)F^c(x), \tag{1}$$

sendo $\alpha \in [0,1]$, F^d a função de distribuição acumulada de uma variável discreta degenerada em

zero, $X_d \equiv 0$, e F^c a função de distribuição acumulada de uma variável aleatória contínua com distribuição exponencial, $X_c \sim \text{Exp}(\lambda)$.

Uma vez proposto o modelo, o passo seguinte foi estimar os parâmetros envolvidos, α e λ. O parâmetro α é interpretado diretamente com uma taxa de inadimplência para os clientes da empresa, cuja estimativa se dá pela frequência de zeros, o que neste caso implica que

$$\widehat{\alpha} = \frac{82}{302} \approx 0,2715.$$

O parâmetro da distribuição exponencial (λ) pode ser interpretado como o inverso do valor médio de pagamentos realizados pelos clientes adimplentes e foi calculado pelo estimador de máxima verossimilhança da exponencial, ou seja, $\widehat{\lambda} = 1/\overline{X}_c$, sendo $\overline{X}_c \approx 210,28$ a média amostral dos dados estritamente positivos. Daí,

$$\widehat{\lambda} = \frac{1}{210,28} \approx 0,0048.$$

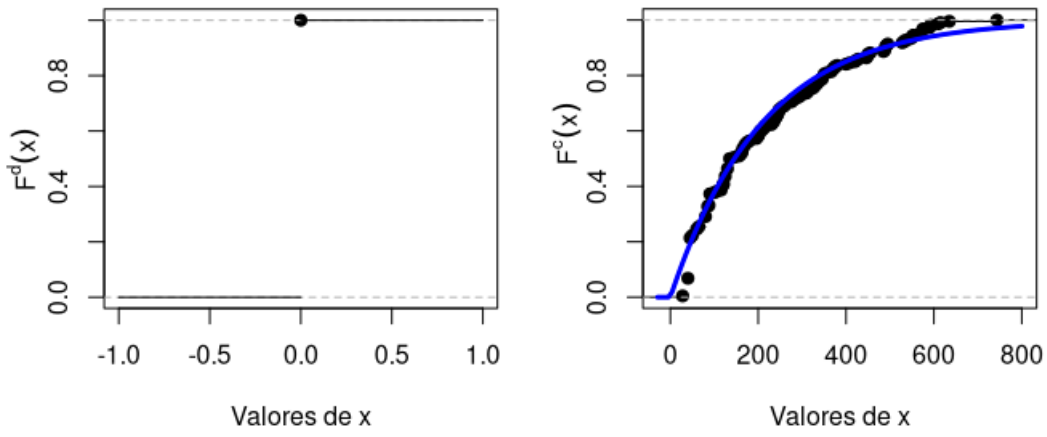
O componente discreto da variável mista tem apenas um valor, que é o 0, e assim a função de distribuição acumulada para o componente discreto fica,

$$F^d(x) = \begin{cases} 0, & \text{se} \quad x < 0, \\ 1, & \text{se} \quad x \geq 0. \end{cases}$$

Já para a componente contínua da variável mista, sua função de distribuição acumulada é

$$F^c(x) = \begin{cases} 0, & \text{se} \quad x < 0, \\ 1 - e^{-\frac{x}{210,28}}, & \text{se} \quad x \geq 0. \end{cases}$$

Na Figura 4, são apresentados os gráficos das funções de distribuição acumulada para as componentes discreta (a) e contínua (b). Na Figura 4 (b), a curva da função de distribuição acumulada da componente exponencial foi sobreposta aos dados estritamente positivos, onde pode-se perceber que a curva se ajusta razoavelmente bem aos dados, o que corrobora com o que já havia sido observado na Figura 3 (b).



(a) (b)

Figura 4: Gráficos das funções de distribuição acumulada para os componentes discreto (a) e contínuo (b).

Assim, voltando para a construção da variável aleatória mista dada pela expressão (1), a função de distribuição acumulada de X, com a combinação das componentes discreta e contínua dadas anteriormente, tem a forma

$$\begin{aligned} F(x) &= \alpha\, F^d(x) + (1-\alpha)\, F^c(x) \\ &= 0,2715 + 0,7285\left(1 - e^{-\frac{x}{210,28}}\right), \end{aligned}$$

ou seja,

$$F(x) = \begin{cases} 0, & \text{se } x < 0, \\ 0,2715 + 0,7285\left(1 - e^{-\frac{x}{210,28}}\right), & \text{se } x \geq 0. \end{cases} \tag{2}$$

A Figura 5, apresenta os valores de probabilidade acumulados a partir dos dados completos (ou seja, com a inclusão dos zeros), juntamente com a sobreposição em vermelho do gráfico da função de distribuição acumulada obtido na Equação (2). Neste caso, destaca-se o salto no ponto zero, que corresponde a parte discreta da variável mista e a curva exponencial para o componente contínuo, onde nota-se um bom ajuste da curva aos valores observados. É importante notar que o tamanho do salto no ponto $x = 0$ corresponde diretamente à taxa de inadimplência dos clientes, ou seja, quanto maior o tamanho deste salto, maior é o número de inadimplentes nos dados, o que certamente é uma informação de grande interesse para a empresa.

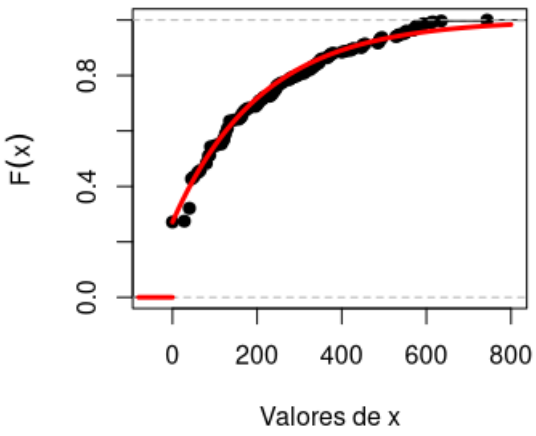


Figura 5: Gráfico da função de distribuição acumulada da variável aleatória mista.

Existem trabalhos que tratam deste tipo de abordagem envolvendo dados inflacionados de zeros (MIN; AGRESTI, 2002), porém, são muito mais frequentes os modelos para dados de contagem, tais como Poisson ou Binomial negativo (ZAMRI; ZAMZURI, 2017). Desta forma, não é muito comum se encontrar trabalhos na literatura que tratem de dados contínuos estritamente positivos inflacionados de zeros, como abordado aqui.

Ainda assim, em outro contexto Hazra et al. (2018) apresentam um trabalho que buscou modelar dados provenientes a índices pluviométricos na Índia, usando uma modelagem exponencial inflacionada de zeros, porém, seguindo uma abordagem bayesiana. Em outro trabalho, Huang et al. (2019) também usa uma modelagem exponencial inflacionada de zeros para modelar taxas de vítimas em acidentes de navio.

Entretanto, como pode ser observado, embora estes trabalhos citados adotem uma modelagem parecida ao que foi feito aqui, nenhum deles está diretamente relacionado com o tipo de problema tratado neste trabalho, que envolve dados financeiros e problemas relacionados à inadimplência dos clientes.

Conclusão

Diversos conjuntos de dados reais modelam diretamente alguma variável aleatória de interesse, que pode ser discreta ou contínua. Muitas vezes, com base apenas em uma análise visual rápida em alguns gráficos descritivos, é possível se adotar algum modelo probabilístico que pode esconder alguns detalhes importantes, e que acabam direcionando para algumas interpretações equivocadas.

Neste caso em particular, observou-se um conjunto de dados referente a pagamentos de serviços de *internet* que indica um modelo exponencial, mas que está inflacionado de zeros (referentes aos clientes inadimplentes). Como a taxa de inadimplência foi relativamente alta (pouco mais de 27%), o problema foi facilmente percebido. Porém, em outras situações reais em que a taxa de presença de zeros seja menor, tal situação pode não ser percebida.

Assim, neste trabalho foi apresentado um ajuste na definição da variável, com base no conceito de variável aleatória mista, que combina uma parte discreta no ponto zero com um componente contínuo exponencial para os valores que foram efetivamente pagos. Além disso, modela adequadamente a variável em questão facilitando inclusive a análise descritiva dos dados e possíveis direcionamentos para novas análises.

Referências

AITCHISON, J. On the distribution of a positive random variable having a discrete probability mass at the origim. *Journal of the American Statistical Association*, vol. 50 n. 271, p. 901-908. 1955.

CHANDRA, S. On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, vol. 4 n. 3, p. 105-112. 1977.

DAROS, M.; PINTO, N. G. M. Inadimplência no Brasil: uma análise das evidências empíricas. *Revista de Administração IMED*, vol. 7 n. 1, p. 208-229. 2017.

HAZRA, A.; BHATTACHARYA, S.; BANIK, P. A Bayesian zero-inflated exponential distribution model for the analysis of weekly rainfall of the eastern plateau region of India. *MAUSAM*, vol. 69 n. 1, p. 19-28. 2018.

HUANG, D.; HU, H.; LI, Y. Zero-Inflated Exponential Distribution of Casualty Rate in Ship Collision. *J. Shanghai Jiao Tong Univ.*, vol. 24 n. 6, p. 739-744. 2019.

JAMES, B. R. *Probabilidade: um curso em nível intermediário.* 2ª Edição. Rio de Janeiro: IMPA, 2002. (Projeto Euclides).

MIN, Y.; AGRESTI, A. Modeling nonnegative data with clumping at zero: a survey. *JIRSS.*, vol. 1 n. 1-2, p. 7-33. 2002.

R CORE TEAM. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2020. Disponível em: http://www.R-project.org/.

ROSS, S. *Probabilidade: um curso moderno com aplicações.* [S.l.]: Bookman, 2010.

WELD, C.; LEEMIS, L. Modeling mixed type random variables. *Proceedings of the 2017 Winter Simulation Conference*, p. 1595-1606. 2017.

WIKIPEDIA. *Brejo Paraibano*. Acesso em: 18 mar 2020. Disponível em: https://bit.ly/3uv1m8r.

ZAMRI, N.S.N.; ZAMZURI, Z.H. A review on models for count data with extra zeros. *AIP Conference Proceedings*, vol. 1830. 2017.