UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Shirlley Oliveira da Silva

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

Campina Grande - PB

Junho de 2018

Shirlley Oliveira da Silva

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

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Divanilda Maia Esteves

Coorientador: Gustavo Henrique Esteves

Campina Grande - PB

Junho de 2018

S586m Silva, Shirlley Oliveira da.
Modelagem probabilística de dados de pagamentos de provedor de internet usando variável mista [manuscrito] : / Shirlley Oliveira da Silva. - 2018.
31 p.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2018.
"Orientação : Profa. Dra. Divanilda Maia Esteves , Coordenação do Curso de Estatística - CCT."
"Coorientação: Prof. Dr. Gustavo Henrique Esteves , Coordenação do Curso de Estatística - CCT.""

1. Variáveis aleatórias. 2. Variável aleatória mista. 3. Distribuição exponencial. 4. Distribuição Bernoulli.

21. ed. CDD 519.5

Elaborada por Giulianne Monteiro Pereira - CRB - 15/714 BC/UEPB

Shirlley Oliveira da Silva

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

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 15 de junho de 2018.

BANCA EXAMINADORA

Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Juarez Fernandes de Oliveira
Universidade Estadual da Paraíba

Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Ao meu adorado DEUS, minha amada filha, Rayssa Brenda O. dos Santos, ao meu amado marido, Luciclaudio dos S. Lima, aos meus amados pais, Walderez O. da Silva e Severino Felinto da Silva e amada irmã, Sheylla O. da Silva.

Agradecimentos

A DEUS, o dono da minha vida, por me proporcionar todas as coisas que me emprestou até aqui, porque DELE, por ELE e para ELE são todas as coisas, toda Honra e toda Glória seja dada a ELE por ter feito com que eu terminasse esse curso!

Aos meus amados pais, Walderez Oliveira e Severino Felinto, e irmã Sheylla Oliveira, por seu amor, carinho ajuda e compreensão.

Ao meu amado marido Luciclaudio e minha amada filha Rayssa por sempre me ajudarem e me compreenderem, meu eterno amor a vocês.

Ao meu cunhado Jany de Oliveira, por me fornecer os dados de sua empresa.

A minha amada avó Irnailda Oliveira, por sua enorme fé em que conseguiria esse feito de terminar esse desafiador curso.

A minha amada tia Irta Maria, por seus conselhos, orações e por acreditar que eu terminaria o curso.

A toda minha amada família o meu muito obrigada.

Ao entrar nesse curso que para muitos é um grande desafio terminá-lo, logo no primeiro período pensei em desistir, pois o peso do desafio estava enorme em minhas costas, foi ai que DEUS colocou uma benção em forma de professora que me aconselhou como uma grande amiga e disse para que eu seguisse em frente e sempre que pensava em desistir me lembrava de seus conselhos e segui em frente. Jamais esquecerei da senhora professora Diana Maia Esteves e do seu marido o professor Gustavo Henrique Esteves, a vocês todo o meu carinho, respeito, amor e gratidão. Aprendi muito com vocês, não só como aluna, mas como pessoa também!

Não poderia jamais deixar de agradecer ao um grande professor e amigo Tiago Almeida de Oliveira, ao senhor todo o meu carinho, respeito, amor e gratidão por tudo jamais esquecerei o quanto o senhor me ajudou também, tanto na minha vida acadêmica quanto na amizade, perdoe-me as pertubações!

Ao professor Gil Luna, que faz falta na UEPB, nenhum aluno mais saberá o legado que ele deixou, a minha admiração, carinho, respeito e amor.

A professora Vitória sempre pronta a me ajudar, infelizmente ou felizmente conheci a pouco tempo, mas pelo pouco tempo é como se fosse há anos, meu respeito, carinho, admiração e amor.

A professora Pollyanna, por sua dedicação e esforço, sempre pronta a ajudar se possível indo de aluno a aluno verificando e se preocupando se realmente o aluno teria

absorvido o conteúdo ou não, o meu respeito, carinho e amor.

A todos os professores e professoras do departamento de Estatística, pois são excelentes profissionais e além de tudo humanos e humildes que trata muito bem os alunos.

Ao professor do Departamento de Matemática Onildo Freire, jamais esquecerei o humilde convite pra confraternização em sua residência, foi ai que descobri que aluno também poderia ser amigo de professor, o meu carinho, respeito, gratidão e amor.

E por fim e sem deixar jamais de citar em especial uma turma da pesada em que todos estudávamos juntos, tudo se combinava e que todos se preocupávamos uns com os outros, já mais a UEPB terá uma turma tão unida como a nossa, o meu sinceros agradecimentos à: Filipe, Fátima, Ângela, Carlos e Vitória, por tudo, aprendi muito com vocês em tudo, o meu carinho, respeito, gratidão e amor. Jamais esquecerei de vocês o quanto me ajudaram em minha vida acadêmica e pessoal, o meu muito obrigada, amo vocês!

Ao meu amado avó Manoel Felinto e a minha amada prima Beatriz in memória.

Por fim, agradeço à banca pelas valorosas sugestões.

JESUS deu indicações precisas da proximidade da SUA volta a este mundo. Disse ELE: "Aprendei, pois a parábola da figueira: Quando já os seus ramos se renovam, e as folhas brotam, sabeis que está próximo o verão. Assim também vós, quando virdes todas estas coisas, sabei que está próximo, às portas."

(Mateus 24:32,33)

Resumo

O objetivo deste trabalho é usar uma variável aleatória mista para modelar dados de pagamentos de provedor de *internet* feitos pelos clientes da empresa Brejo Net. 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 categorizações ou 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 aqui 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, donde surgiu a ideia de usar uma variável mista, sendo que a parte positiva será modelada pela distribuição exponencial e os zeros por uma Bernoulli com probabilidade de sucesso igual a 0. Assim, neste trabalho será feita uma revisão básica da teoria de variáveis aleatórias, dos modelos Bernoulli e exponencial e, por fim, a aplicação das variáveis mistas aos dados citados anteriormente.

Palavras-chaves: Variável aleatória mista, Distribuição exponencial, Pagamentos de provedor de Internet.

Abstract

The objective of this work is to use a mixed random variable to model data from payments made by customers of Brejo Net. 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 categorizations or counts, 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 here, 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 idea of using a mixed variable, and the positive part will be modeled by the exponential distribution and the zeros by a Bernoulli with probability of success equal to 0. Thus, in this work we performed a basic revision of the theory of random variables, Bernoulli and exponential models, and finally the application of the mixed variables to the data previously mentioned.

Key-words: Mixed random variable, exponential distribution, internet provider payments.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Uma parte importante das análises estatísticas é aquela que são de definir variáveis aleatórias que sejam capazes de mensurar as características de interesse. Isso é importante porque as análises estatísticas são feitas com o uso de ferramentas matemáticas. Depois de feita a análise exploratória dos dados, muitas vezes a análise é aprofundada através de inferências. Por exemplo, são construídos intervalos de confiança, realizados testes de hipótese, são propostos modelos de regressão. 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 melhor se enquadre na situação. Em outras situações, observa-se que a variável é formada por uma parte discreta e outra contínua, o que se chama variável aleatória mista.

O conjunto de dados analisados neste trabalho são referentes a pagamentos feitos por clientes da empresa de Brejo Net, situada em Alagoa Grande-PB. O que foi observado é que apesar da análise gráfica sugerir o uso de um modelo exponencial, havia uma quantidade considerável de clientes inadimplentes, o que implicou na presença de vários zeros nos dados. Daí, surgiu a ideia de usar uma variável aleatória mista, com a parte contínua sendo representada pela distribuição exponencial e a parte discreta representada por uma variável degenerada no ponto 0.

Assim, este trabalho está organizado da seguinte maneira. O Capítulo 2 visa fazer uma breve revisão de variáveis aleatórias, incluindo as variáveis mistas, as quais não são comumente contempladas nos cursos básicos de probabilidade, bem como relembrar as distribuições Bernoulli e exponencial. No Capítulo 3, serão apresentados os dados utilizados, bem como a distribuição mista com os respectivos parâmetros estimados. Toda a parte computacional foi feita usando o *software* R (R Core Team, 2018). Por fim, as conclusões do trabalho serão apresentadas.

2 Fundamentação Teórica

Um experimento aleatório é aquele que não se conhece o resultado exato antes de sua realização. Em contrapartida tem-se os experimentos determinísticos que por sua vez são aqueles que já se conhece o resultado antes mesmo de sua ocorrência. Os exemplos dos dois tipos de experimentos são inúmeros. São experimentos aleatórios lançar um dado e observar a face obtida, medir a duração (em minutos) de uma viagem de ônibus entre Campina Grande e João Pessoa, a porcentagem de votos com a qual será eleito o próximo presidente do Brasil, por exemplo. Pode-se citar entre os experimentos determinísticos, a temperatura sob a qual a água ferve em condições normais de pressão e a idade que teremos de hoje a exatamente um ano.

Os experimentos aleatórios acontecem a toda hora no cotidiano de todas as pessoas e busca-se modelá-los para melhor compreendê-los. Inicialmente, define-se como **espaço amostral** o conjunto de todos os resultados possíveis de um experimento e como **evento** um subconjunto qualquer do espaço amostral. O espaço amostral pode ser um conjunto enumerável ou não. Quando temos um espaço amostral enumerável, cada elemento do espaço amostral é chamado de **evento simples**.

Além de conhecer os resultados possíveis, é bastante útil associar valores a esses resultados que reflitam suas possibilidades de ocorrência. Tais valores são chamados **probabilidade**. A seguir serão apresentadas três definições de probabilidade. A primeira delas (Definição Clássica) é a mais simples e intuitiva e a terceira (Definição Axiomática) é a mais geral. As duas primeiras são apresentadas desde o ensino médio e a última, em geral, só é vista em cursos mais específicos de probabilidade. Elas podem ser encontradas em diversos livros e aqui são apresentadas baseando-se em Meyer (1983), Dantas (2004), James (2002) e Hazzan (2013).

Definição 2.1. (Definição Clássica) Seja um espaço amostral finito com N elementos, $N \neq 0$, e considere que todos eles tem a mesma possibilidade de ocorrer (espaço amostral equiprovável). Então a probabilidade de ocorrência de um evento A com n elementos é definida como

$$P(A) = \frac{n}{N}.$$

—

Definição 2.2. (Definição Frequentista) Suponha que um experimento é repetido N vezes nas mesmas condições de modo que as repetições sucessivas não dependam dos resultados anteriores. Se o evento A associado a esse experimento ocorre n_A vezes, então

$$P(A) = \frac{n_A}{N}.$$

Esta definição equivale a calcular a proporção de vezes que um evento aleatório ocorre em N repetições independentes de um experimento e é equivalente ao conceito de frequência relativa. —

Definição 2.3. (Definição Axiomática) Seja um experimento aleatório com espaço amostral Ω. Define-se probabilidade como uma função denotada por P que satisfaz os seguintes axiomas de Kolmogorov:

1. $0 \leq P(A) \leq 1$, para todo $A \subset \Omega$;
2. $P(\Omega) = 1$;
3. Para qualquer sequência de eventos mutuamente exclusivos $A_1, A_2, \cdots \subset \Omega$, isto é, eventos para os quais $A_i \bigcap A_j = \emptyset$ quando $i \neq j$, tem-se:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

—

A partir dos três axiomas, deduzem-se as seguintes propriedades para a probabilidade:

P1. $P(\emptyset) = 0$

P2. $P(A^c) = 1 - P(A)$

P3. $P(A - B) = P(A) - P(A \cap B)$

P4. $P(B - A) = P(B) - P(A \cap B)$

P5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

P6. $A \subset B \Rightarrow P(A) \leq P(B)$.

2.1 Variáveis aleatórias

Muitas vezes, para modelar estatisticamente os fenômenos aleatórios, é necessário "transformar" os resultados em resultados numéricos.

Definição 2.4. Uma variável aleatória X é uma função que associa a cada elemento ω_i do espaço amostral Ω um número real $X(\omega_i)$. —

Uma variável aleatória X tem domínio no espaço amostral Ω e imagem 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 as variáveis singulares que não serão detalhadas aqui, pois foge ao objetivo proposto, mas que podem ser encontradas, por exemplo, em James (2002).

Definição 2.5. Uma variável aleatória X será considerada discreta se ela assume valores em um conjunto enumerável. Quando ela assume um único valor x, ela é dita degenerada. Neste caso, $P(X = x) = 1$. Se o conjunto dos valores possíveis da variável for a reta $\mathbb{R}$ ou parte dela, X é uma variável aleatória contínua. —

Definição 2.6. Uma variável aleatória X será dita mista se o conjunto dos valores que ela assume tiver ao mesmo tempo uma parte finita enumerável e outra contínua, ou seja, assume valores em um conjunto enumerável e na reta real ou parte dela. —

Segundo Dantas (2004), como se associa valores de probabilidade aos eventos, é possível também atribuir probabilidade aos valores da variável aleatória.

Definição 2.7. Uma função de probabilidade de uma variável aleatória X discreta, estabelecida em um espaço de probabilidade Ω, é uma função que associa a cada valor x de X as suas probabilidades de ocorrência. Se X assume valores em $\{x_1, x_2, \dots\}$ e definindo $A_i = \{\omega \in \Omega : X(\omega) = x_i\}$. , então

$$P[X = x_i] = p(x_i) = P(A_i).$$

—

A função de probabilidade, por construção, satisfaz as seguintes propriedades:

fp1. $0 \leq p(x_i) \leq 1, \forall i = 1, 2, \dots$ e

fp2. $\sum_i p(x_i) = 1$.

Definição 2.8. Uma função $f(\cdot)$ é função densidade de probabilidade da variável aleatória contínua X, se

1. $f(x) \geq 0, \forall x \in \mathbb{R}$;
2. $\int_{-\infty}^{\infty} f(x)dx = 1$;
3. $P[a \leq X \leq b] = \int_a^b f(x)dx$.

—

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, a qual será definida a seguir.

Definição 2.9. Define-se função de distribuição acumulada ou simplesmente função de distribuição de uma variável aleatória X como

$$F(x) = P[X \leq x], \forall x \in \mathbb{R}.$$

—

A função de distribuição acumulada possui as seguintes propriedades, as quais não serão demonstradas aqui (DANTAS, 2004).

F1. $0 \leq F(x) \leq 1$,

F2. $F(x)$ é não decrescente e contínua à direita,

F3. $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to \infty} F(x) = 1$.

Quando X é uma variável aleatória discreta assumindo valores em $S = \{x_1, x_2, \dots\}$, sua função de distribuição é

$$F(x) = \sum_{x_i \in S : x_i \leq x} P[X = x_i].$$

Neste caso, a função é do tipo escada, os saltos ocorrem nos pontos $x_i \in S$ e a altura do salto no ponto x_i é igual a $p(x_i)$.

Se X é uma variável aleatória contínua,

$$F(a) = \int_{-\infty}^{a} f(x)dx, \forall a \in \mathbb{R}.$$

A função de distribuição de uma variável aleatória mista é composta por partes distintas: uma discreta e outra contínua. De maneira mais geral, seria incluída ainda uma parte singular, mas a variável aleatória singular foge do objetivo proposto, além de raramente ocorrer em situações práticas (JAMES, 2002). Se X é uma variável aleatória mista, sua função de distribuição pode ser especificada através da média ponderada entre uma função de distribuição discreta e uma contínua , isto é,

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

onde $\alpha > 0$ será o peso dado a cada componente da mistura e F^d e F^c são suas respectivas distribuições do tipo discreta e contínua.

A seguir, serão definidas duas quantidades que representam o centro de massa da distribuição de probabilidade da variável aleatória X e de sua dispersão em torno deste centro de massa.

Definição 2.10. O **valor esperado**, também chamado média ou esperança, de uma variável aleatória X, que será denotado por $E(X)$ é

- $E(X) = \sum_{i=1}^{\infty} x_i p(x_i)$, se X for discreta ou
- $E(X) = \int_{-\infty}^{\infty} x f(x) dx$, se X for contínua.

Diz-se que tal esperança existe quando a soma acima - ou a integral no caso contínuo - converge. —

Sejam X e Y variáveis aleatórias, tais que $E(X)$ e $E(Y)$ existem, e $c \in \mathbb{R}$ uma constante. Então valem as seguintes propriedades:

E1 $E(c) = c$

E2 $E(X + Y) = E(X) + E(Y)$

E3 $E(cX) = cE(X)$

Definição 2.11. A **variância** de uma variável aleatória X, denotada por $Var(X)$, é dada por

$$Var(X) = E[(X - E(X))^2].$$

—

A variância indica o grau de dispersão de probabilidade em torno da esperança, ou seja, quanto menor a variabilidade em torno da esperança mais ela será precisa. Sabendo que $E(X)$ é uma constante, tem-se

$$\begin{aligned} Var(X) = E[(X - E(X))^2] &= E\left(X^2 - 2XE(X) + E^2(X)\right) \\ &= E(X^2) - 2E(X)E(X) + E^2(X) \\ &= E(X^2) - E^2(X), \end{aligned}$$

sendo que

- $E(X^2) = \sum_{i=1}^{\infty} x_i^2 p(x_i)$, se X for discreta ou

- $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$, se X for contínua.

Considerando $c \in \mathbb{R}$ uma constante, são propriedades da variância:

V1 $Var(c) = 0$;

V2 $Var(cX) = c^2 Var(X)$;

V3 $Var(c + X) = Var(X)$;

V4 $Var(X + Y) = Var(X) + Var(Y)$, para X e Y variáveis independentes.

Para encontrar esperança e variância de uma **variável aleatória mista** X, será usado o seguinte resultado, o qual não será demonstrado aqui. (ROSS, 2010)

Teorema 2.1. *$E(X) = E[(X|Y)]$ e $Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$.* —

Considere a variável

$$Y = \begin{cases} 0, & \text{se } X = X_c \\ 1, & \text{se } X = X_d. \end{cases}$$

Deste modo, $P(Y = 0) = 1 - \alpha$ e $P(Y = 1) = \alpha$. Daí,

$$\begin{aligned} E(X) &= E[E(X|Y)] = E(X|Y = 1) \cdot P(Y = 1) + E(X|Y = 0) \cdot P(Y = 0) \\ &= \alpha\, E^d(X) + (1 - \alpha)\, E^c(X), \end{aligned}$$

sendo $E^d(X)$ a esperança da parte discreta, $E^c(X)$ a esperança da parte contínua e α o peso da mistura. No caso da variância da **variável aleatória mista**

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)].$$

O primeiro termo da soma é

$$\begin{aligned} E[Var(X|Y)] &= Var(X|Y = 1) \cdot P(Y = 1) + Var(X|Y = 0) \cdot P(Y = 0) \\ &= \alpha Var^d(X) + (1 - \alpha)Var^c(X). \end{aligned}$$

Usando a definição de variância,

$$\begin{aligned} Var[E(X|Y)] &= E\left[E(X|Y) - E(E(X|Y))\right]^2 \\ &= E\left[E(X|Y) - E(X)\right]^2 \\ &= [E(X|Y = 0) - E(X)]^2 \cdot P(Y = 0) + [E(X|Y = 1) - E(X)]^2 \cdot P(Y = 1) \\ &= [E^c(X) - E(X)]^2(1 - \alpha) + [E^d(X) - E(X)]^2\alpha \\ &= \alpha^2(1 - \alpha)[E^c(X) - E^d(X)]^2 + \alpha(1 - \alpha)^2[E^d(X) - E^c(X)]^2 \\ &= \alpha(1 - \alpha)[E^c(X) - E^d(X)]^2. \end{aligned}$$

Portanto,

$$\begin{aligned} Var(X) &= E[Var(X|Y)] + Var[E(X|Y)] \\ &= \alpha Var^d(X) + (1-\alpha)Var^c(X) + \alpha(1-\alpha)[E^c(X) - E^d(X)]^2. \end{aligned}$$

Há outra forma também de identificar uma variável aleatória X além das funções de probabilidade e função de distribuição, ou seja, a função geradora de momentos também determina outras distribuições de funções de variáveis aleatórias.

Definição 2.12. Chame-se a esperança de X^k de k-ésimo momento ou momento de ordem k da variável aleatória X, $\forall k = 1, 2, 3, \cdots$. Deste modo, o valor da esperança é o momento de ordem 1. Define-se ainda a esperança de $(X - E(X))^k$ como sendo o k-ésimo momento central de X. Assim, a variância é o momento central de ordem 2 de X. —

Definição 2.13. A função geradora de momentos de uma variável aleatória X é definida por

$$\phi_X(t) = E(e^{tX}).$$

—

Sendo assim, são propriedades da função geradora de momentos (MEYER, 1983)

GM1. Desde que a esperança exista e seja finita e que para cada número real t em que $-\infty < t < \infty$ e que a função seja definida em uma vizinhança do ponto zero, sendo os momentos obtidos em uma sucessão diferenciável aplicada em zero,

$$\phi^k(0) = E(X^k).$$

GM2. $\phi(0) = 1$

GM3. A função geradora de momentos define por completo a distribuição da variável aleatória e ela é única.

No caso para **variável aleatória discreta** a função geradora de momentos é calculada da seguinte maneira

$$\phi_X(t) = \sum_{i=1}^{\infty} e^{tx_i}\ P[X = x_i],$$

para **variável aleatória contínua** a função geradora de momentos é descrita como

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

e para **variável aleatória mista** sua função geradora de momentos é da seguinte maneira

$$\phi_X(t) = \alpha\phi_d(x) + (1-\alpha)\phi_c(x).$$

2.2 Distribuição Bernoulli

Muitos experimentos aleatórios são de tal forma que quando observados há apenas a ocorrência ou não de uma certa característica, ou seja, só haverá dois resultados aleatórios possíveis. Por exemplo, quando se lança uma moeda honesta e observa-se a face que ficou voltada para cima, os resultados possíveis são cara ou coroa. Esse tipo de experimento é chamado ensaio de Bernoulli.

Geralmente, há interesse especial na ocorrência de um dos resultados, o qual será chamada de sucesso, ficando o outro resultado possível caracterizado como fracasso. Neste caso, pode-se definir uma variável aleatória X como

$$X = \begin{cases} 0, \text{se o experimento resultou em fracasso,} \\ 1, \text{se o experimento resultou em sucesso.} \end{cases}$$

Assumindo que a probabilidade de sucesso seja p, a distribuição da variável aleatória X será

$$P[X = x] = \begin{cases} 1-p, & \text{se} \quad x = 0, \\ p, & \text{se} \quad x = 1, \\ 0, & \text{caso contrário,} \end{cases}$$

sendo que neste caso diz-se que a variável aleatória tem **Distribuição Bernoulli**, o que será denotado por $X \sim$ Bernoulli (p). Alternativamente, pode-se escrever a função de probabilidade da seguinte maneira

$$P(X = x) = p^x(1-p)^{1-x},$$

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

Como consequência, segue que a função de distribuição acumulada aqui será dada por

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

Se $X \sim Bernoulli(p)$, então valor esperado será $E(X) = p$, pois

$$\begin{aligned} E(X) &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot (1-p) + 1 \cdot p = p. \end{aligned}$$

Além disso,

$$\begin{aligned} E(X^2) &= 0^2 \cdot P(X = 0) + 1^2 \cdot P(X = 1) \\ &= 0 \cdot (1-p) + 1 \cdot p = p. \end{aligned}$$

o que implica que variância de X é dada por

$$Var(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p).$$

A função geradora de momentos da distribuição Bernoulli é

$$\phi_X(t) = E(e^{tX}) = e^{t\cdot 0}(1-p) + e^{t\cdot 1}p = 1 - p + pe^t.$$

2.3 Distribuição Exponencial

Definição 2.14. Uma variável aleatória X tem distribuição exponencial com parâmetro $\lambda > 0$, denotado $X \sim exp(\lambda)$, se sua função densidade de probabilidade é dada por

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

Se uma variável aleatória X é exponencialmente distribuída, então, para $x > 0$,

$$\begin{aligned} F(x) = P(X \leq x) &= \int_0^x f(x)dx \\ &= \int_0^x \lambda e^{-\lambda x}dx = 1 - e^{-\lambda x}. \end{aligned}$$

Em outras palavras, se $X \sim exp(\lambda)$,

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

Na Figura 1 é possível ver os gráficos da função densidade de probabilidade e da função de distribuição acumulada da exponencial para um valor particular do parâmetro.

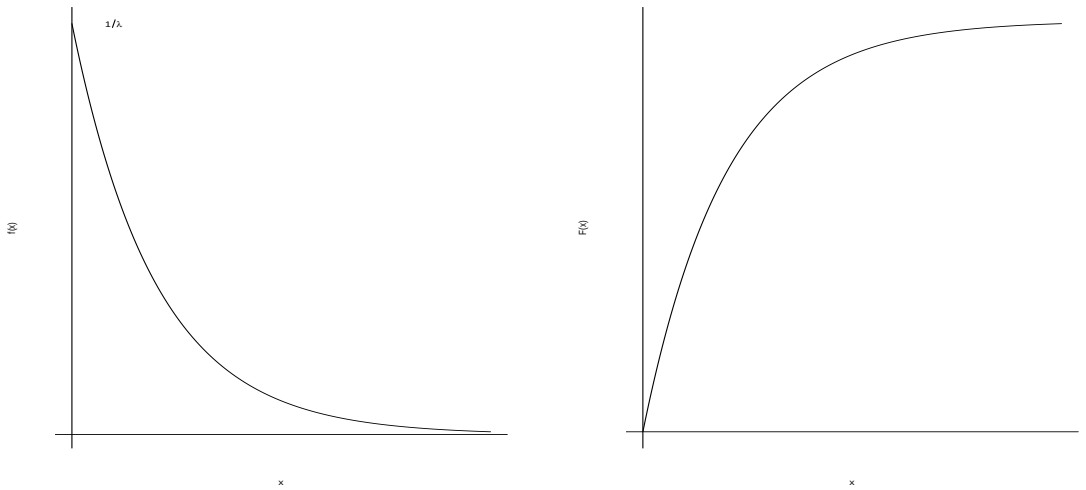

Figura 1 – Gráficos da função densidade de probabilidade (esquerda) e da função de distribuição acumulada de uma variável aleatória X exponencialmente distribuída com $\lambda = \frac{1}{2}$. (Fonte: própria)

O valor esperado de $X \sim exp(\lambda)$ é dado por

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{\infty} x\ \lambda e^{-\lambda x} dx$$

integrando por partes, obtém-se

$$\begin{aligned} E(X) &= \lambda \left[-\frac{x\ e^{-\lambda x}}{\lambda} \Big|_0^{\infty} - \int_{-\infty}^{\infty} -\frac{e^{-\lambda x}}{\lambda} dx \right] \\ &= \lambda \left[-\frac{x\ e^{-\lambda x}}{\lambda} \Big|_0^{\infty} + \frac{1}{\lambda} \left(-\frac{e^{-\lambda x}}{\lambda} \right) \Big|_0^{\infty} \right] \\ &= \lambda \left[-\frac{x\ e^{-\lambda x}}{\lambda} - \frac{e^{-\lambda x}}{\lambda^2} \Big|_0^{\infty} \right] \\ &= \lambda \left[\frac{1}{\lambda^2} \right] = \frac{1}{\lambda}. \end{aligned}$$

Também, usando integração por partes, pode-se obter o segundo momento da variável

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{0}^{\infty} x^2 \lambda e^{-\lambda x} dx \\ &= \lambda \left[-\frac{x^2 e^{-\lambda x}}{\lambda} \Big|_0^{\infty} + 2 \int_0^{\infty} \frac{x e^{-\lambda x}}{\lambda} dx \right] \\ &= \lambda \left[-\frac{x^2 e^{-\lambda x}}{\lambda} \Big|_0^{\infty} + \frac{2}{\lambda} \left(-\frac{x e^{-\lambda x}}{\lambda} \right) \Big|_0^{\infty} + \frac{1}{\lambda} \left(-\frac{e^{-\lambda x}}{\lambda} \right) \Big|_0^{\infty} \right] \\ &= \lambda \left[-\frac{x^2 e^{-\lambda x}}{\lambda} - \frac{2x e^{-\lambda x}}{\lambda^2} - \frac{2e^{-\lambda x}}{\lambda^3} \Big|_0^{\infty} \right] \\ &= \lambda \left[\frac{2}{\lambda^3} \right] = \frac{2}{\lambda^2}. \end{aligned}$$

E assim, substituindo na fórmula da variância tem-se

$$Var(X) = \left[E(X^2) - (E(X))^2 \right] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Para encontrar a função geradora de momentos da distribuição exponencial, usa-se a definição:

$$\phi_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx.$$

Se $t = \lambda$,

$$\phi_X(t) = \int_0^{\infty} \lambda dx \to \infty.$$

Se $t \neq \lambda$,

$$\phi_X(t) = \lambda \left[-\frac{e^{-(\lambda - t)x}}{(\lambda - t)} \Big|_0^{\infty} \right].$$

Neste caso, quando $t > \lambda$, então $(\lambda - t) < 0$ e consequentemente

$$\left[-\frac{e^{-(\lambda - t)x}}{(\lambda - t)} \right] \to \infty.$$

Por fim, para $t < \lambda$,

$$\phi_X(t) = \lambda \left[0 + \frac{1}{(\lambda - t)} \right] = \frac{\lambda}{(\lambda - t)}.$$

2.4 Estimador de máxima verossimilhança

Para caracterizar completamente a distribuição da variável aleatória, é essencial conhecer além de sua distribuição e os valores dos parâmetros envolvidos. Quando se decide usar um determinado modelo probabilístico para modelar uma amostra de alguma variável aleatória, deve-se buscar um estimador para o(s) parâmetro(s), ou seja, uma função da amostra que represente adequadamente o valor desconhecido do parâmetro. Um dos métodos mais populares para encontrar estimadores é o **método da máxima verossimilhança**. Tal método consiste em encontrar uma função da amostra que maximize a função de verossimilhança

$$L(\theta) = L(\theta; X_1, ..., X_n) = f(X_1; \theta) \times ... \times f(X_i; \theta) = \prod_{i=1}^{n} f(X_i; \theta),$$

sendo $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $f(x; \theta)$.

Quando $L(\theta)$ é derivável, os candidatos a estimador de máxima verossimilhança serão $\hat{\theta}$ que tornam a primeira derivada da função igual a zero. No entanto, calcular a derivada dessa função pode não ser uma tarefa muito fácil, pois ela é o produto de n funções $f(X_i, \theta)$. Desta maneira, usa-se um artifício para facilitar a derivação: em vez de maximizar a função $L(\theta)$, maximiza-se a função $l(\theta) = ln(L(\theta))$. Dado que a função $\ln(\cdot)$ é crescente, o mesmo valor de θ maximizará as duas funções. E ao aplicar a função ln, o produto transforma-se em soma, o que é bem mais fácil de derivar.

A função de verossimilhança para o parâmetro p de uma distribuição Bernoulli é

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

Daí,

$$\begin{aligned} l(p) &= \left(\sum_{i=1}^{n} X_i \right) \ln(p) + \left(n - \sum_{i=1}^{n} X_i \right) \ln(1-p) \\ &= n[\bar{X} \ln(p) + (1 - \bar{X}) \ln(1-p)]. \end{aligned}$$

Derivando a função $l(p)$ acima e igualando a zero. encontra-se o "candidato" a ponto de máximo da função $(\hat{p})$

$$\begin{aligned} \frac{dl}{dp} = 0 \Rightarrow n \left[\frac{\bar{X}}{\hat{p}} - \frac{(1-\bar{X})}{(1-\hat{p})} \right] &= 0 \\ \Rightarrow n\bar{X}(1-\hat{p}) &= \hat{p}(n - n\bar{X}) \\ \Rightarrow n\bar{X} - n\bar{X}\hat{p} &= \hat{p}n - n\bar{X}\hat{p} \\ \Rightarrow \hat{p} &= \bar{X}. \end{aligned}$$

No caso da distribuição exponencial, a função de verossimilhança é

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i},$$

de onde segue que

$$l(\lambda) = n \log \lambda + (-\lambda \sum_{i=1}^{n} x_i).$$

Derivando $l(\lambda)$ e igualando a zero tem-se que o estimador de máxima verossimilhança de λ será

$$\frac{dl}{d\hat{\lambda}} = 0 \Rightarrow n \left[\frac{1}{\hat{\lambda}} - \bar{X} \right] = 0 \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}.$$

3 Aplicação

Os dados usados aqui são referentes a pagamentos realizadas por usuários de internet da empresa Brejo Net, do mês de março do ano de 2017. A empresa, localizada na cidade de Alagoa Grande-PB, fornece internet banda larga (em cabos coaxis, fibras ópticas ou cabos metálicos), sendo o preço estabelecido conforme a quantidade de megabytes que o cliente deseja usar. A amostra é composta de 302 observações, sendo que 82 desses não pagaram e 220 pagaram pelo o uso dos serviços prestados pela empresa. Os dados podem ser vistos na Tabela 1.

Tabela 1 – Dados de valores pagos pelo uso de internet referente ao mês de março do ano de 2017 para a Brejo Net.

Val. = 0	Obs.	Val. > 0	Obs.	Val. > 0	Obs.	Val. > 0	Obs.	Val. > 0	Obs.
0,00	82	28,00	1	175,00	2	293,00	1	448,00	1
		40,00	14	180,00	1	295,00	1	450,00	2
		45,00	32	196,00	2	300,00	1	455,00	1
		50,00	2	200,00	2	310,00	1	485,00	1
		60,00	5	201,20	1	315,00	2	488,00	1
		65,00	2	205,00	3	325,00	2	490,00	3
		80,00	8	210,00	1	330,00	2	493,00	1
		85,00	8	215,00	2	331,00	1	495,00	1
		88,00	1	216,00	1	335,00	1	528,00	1
		90,00	9	230,00	1	340,00	2	533,00	1
		100,00	1	235,00	2	345,00	1	535,00	1
		105,00	1	236,00	2	346,00	1	545,00	1
		115,50	1	238,00	1	350,00	3	550,00	1
		120,00	4	241,00	1	355,00	1	551,00	1
		121,00	1	242,00	1	366,00	1	553,00	1
		125,00	6	243,00	1	370,00	2	569,00	1
		130,00	6	245,00	1	373,00	1	571,00	1
		135,00	8	246,00	1	375,00	1	575,00	3
		145,00	1	248,00	2	380,00	1	590,00	1
		156,00	1	255,00	2	400,00	1	591,00	1
		138,00	1	260,00	1	408,00	1	593,00	1
		161,00	1	275,00	3	420,00	1	610,00	1
		163,00	1	278,00	1	423,00	1	615,00	1
		165,00	3	280,00	1	428,00	1	635,00	1
		170,00	3	286,00	1	446,00	1	743,40	1

Na Figura 2, pode-se observar um histograma dos dados, sendo que, sua forma indica que os dados parecem seguir uma distribuição exponencial. A empresa presta serviços de internet e, ao longo do mês de março do ano de 2017, 82 clientes não pagaram,

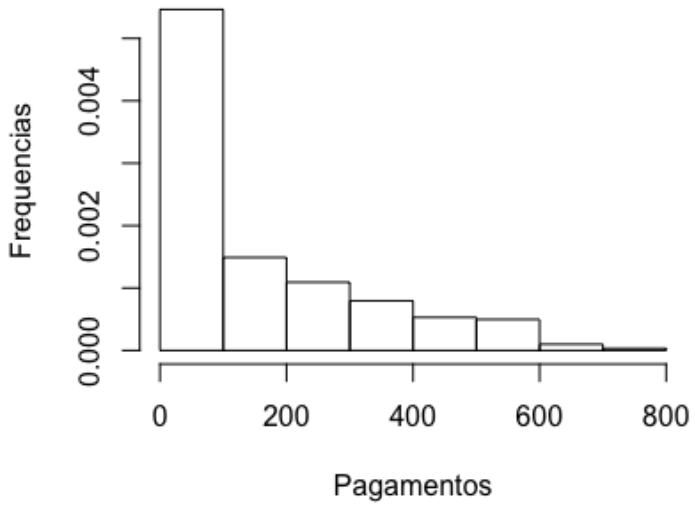

Figura 2 – Histograma dos dados referentes a pagamentos de clientes da Brejo Net.

o que representa $27,15\%$ de inadimplência. Ou seja, apesar do comportamento exponencial, há uma quantidade considerável de zeros, sendo que a distribuição exponencial é definida apenar para valores positivos. A Figura 3 apresenta um histograma dos dados sem os valores nulos, que reforça a ideia de se usar uma distribuição exponencial. Apesar de ambos os gráficos apresentarem valores que decrescem exponencialmente, quando se compara a primeira barra (valores menores que 100) nos dois gráficos, vê-se que a frequência no primeiro gráfico é praticamente o dobro da frequência para o mesmo intervalo de variação no segundo histograma.

Deste modo, percebe-se que há duas variáveis distintas: uma discreta, degenerada no valor 0, e outra contínua, referente aos clientes que não ficaram inadimplentes. Em outra palavras, definindo-se a variável X como sendo a quantidade de dinheiro paga pelo cliente à empresa, tem-se a seguinte variável mista

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

sendo $\alpha \in [0,1]$. A proposta é considerar $X^d \sim Bernoulli(0)$ - $X \equiv 0$ - e $X_c \sim exp(\lambda)$.

Uma vez proposto o modelo, o passo seguinte será estimar os parâmetros envolvidos, neste caso α e λ. A estimativa de α é a frequência de zeros, o que neste caso implica

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

O parâmetro da distribuição exponencial (λ) será estimado por máxima verossimilhança. Conforme visto anteriormente, $\widehat{\lambda} = (\bar{X}^c)^{-1}$, sendo $\bar{X}^c$ a média amostral dos dados sem os

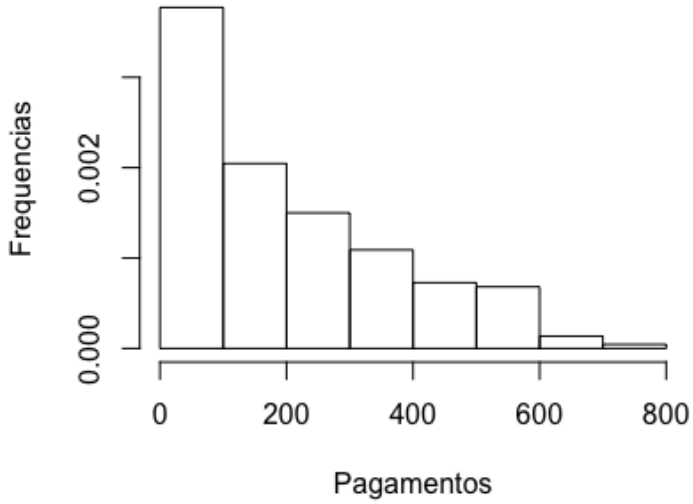

Figura 3 – Histograma dos dados referentes a pagamentos de clientes da Brejo Net, excluindo-se os que não pagaram.

zeros. Daí,

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

O componente discreto da variável mista tem apenas um valor, que é o 0. Daí,

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

Para a parte contínua,

$$F^c(x) = \begin{cases} 0, & \text{se} \quad x < 0, \\ 0,7285 \ \left(1 - e^{-\frac{x}{210,2823}}\right), & \text{se} \quad x \geq 0. \end{cases}$$

Assim, a função distribuição de X é

$$\begin{aligned} F(x) &= \alpha \ F^d(x) + (1-\alpha) \ F^c(x) \\ &= 0,2715 \cdot 1 + 0,7285 \cdot \left(1 - e^{-\frac{x}{210,2823}}\right) \\ &= 0,2715 + 0,7285 \left(1 - e^{-\frac{x}{210,2823}}\right), \end{aligned}$$

ou seja,

$$F(x) = \begin{cases} 0,2715 + 0,7285 \left(1 - e^{-\frac{x}{210,2823}}\right), \text{ se } x \geq 0, \\ 0, \text{ caso contrário.} \end{cases} \tag{3.1}$$

Nas Figuras 4 e 5, são apresentados dois gráficos funções de distribuição acumulada, sendo que a primeira representa a parte discreta e o segundo representa a parte contínua. Na

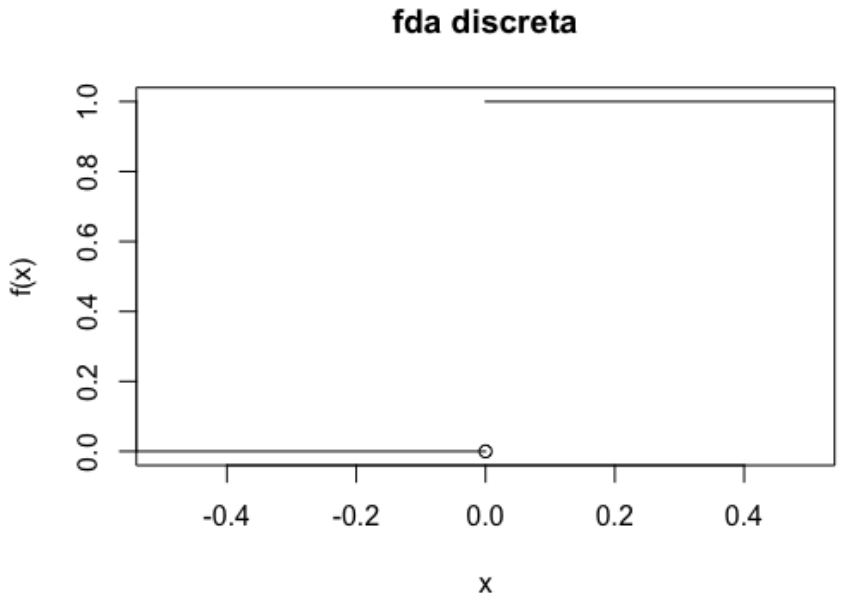

Figura 4 – Gráfico da função de distribuição acumulada do componente discreto.

Figura 5, a curva da função de distribuição acumulada da distribuição exponencial foi sobreposta aos dados excluindo-se os valores nulos. Pode-se perceber que a maior pontos está próxima da curva.

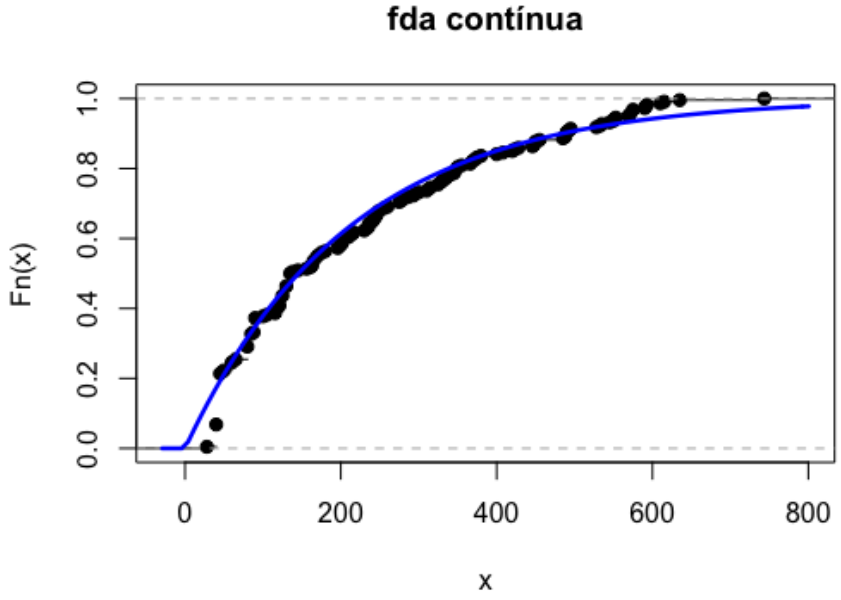

Figura 5 – Gráfico da função de distribuição acumulada do componentes contínuo.

Já na Figura 6, é possível visualizar a função de distribuição acumulada da variável mista, juntamente com o gráfico da função de distribuição acumulada, cuja função com os

parâmetros estimados pode ser vista na Equação 3.1. Neste caso, vê-se que há um salto no ponto 0 (representando a parte discreta) e uma curva (representando o componente contínuo). Visualmente, os pontos amostrais parecem próximos da curva. Para tal gráfico foram usados todos os dados, incluindo os de clientes inadimplentes.

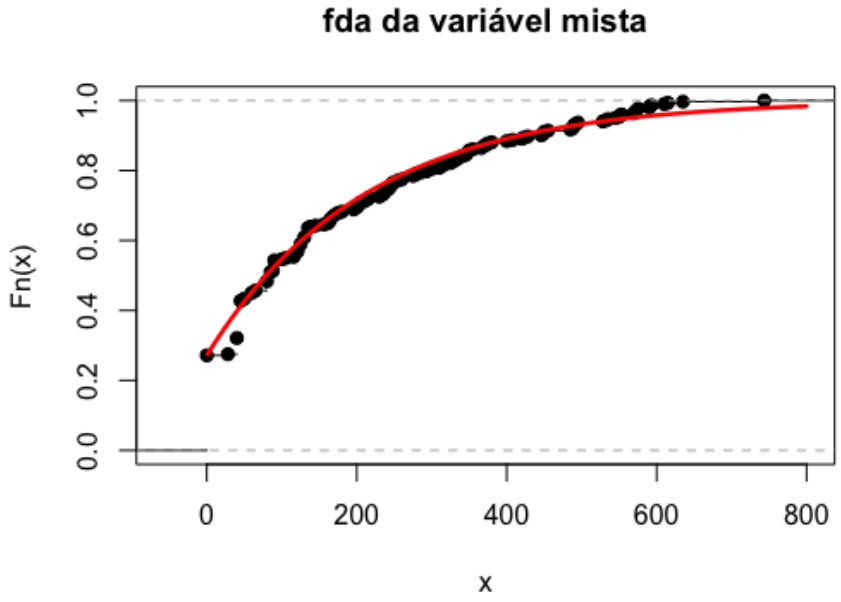

Figura 6 – Gráfico da função de distribuição acumulada da variável aleatória mista.

Para encontrar a esperança e variância da distribuição da variável aleatória mista é só calcular da seguinte maneira

$$\begin{aligned} E(X) &= \alpha E^d(X) + (1-\alpha)\, E^c(X) \\ &= 0,2715\ .\ 0 + 0,7285\ .\ 210,2823 \\ &= 153,1906. \end{aligned}$$

e

$$\begin{aligned} Var(X) &= \alpha Var^d(X) + (1-\alpha)Var^c(X) + \alpha(1-\alpha)\ (E^d(X) - E^c(X))^2 \\ &= (0,2715\ .\ 0 + 0,7285\ .\ 44218,6457) + 0,2715\ .\ 0,7285\ (0 - 210,2823)^2 \\ &= 32213,2834 + 8745,9064 \\ &= 40959,1898. \end{aligned}$$

4 Conclusão

Por serem muito frequentes as variáveis aleatórias discreta e contínua encontradas em problemas práticos no dia a dia, também devemos considerar que ambas as variáveis juntas são muito frequentes em um conjunto de dados e que conforme esses dados também pode-se perceber com base na análise gráfica dos dados de pagamentos de provedor de internet que há uma distorção à direita indicando a ocorrência de altos valores com baixa frequência e parece haver uma indicação aproximada de que é adequado a modelagem com a distribuição Exponencial , no entanto, quando existem uma proporção de valores nulos, a mesma torna-se inadequada, então recomenda-se o uso de mistura de variáveis entre a distribuição exponencial e a distribuição degenerada em zero (Bernoulli), com isso, provavelmente há indícios de que o modelo parece estar bem ajustado.

Referências

DANTAS, C. A. B. *Probabilidade: Um Curso Introdutório*. [S.l.]: Edusp, 2004. Citado 3 vezes nas páginas 13, 15 e 16.

HAZZAN, S. *Fundamentos de Matemática Elementar*. [S.l.]: Atual Editora, 2013. v. 5. Citado na página 13.

JAMES, B. R. *Probabilidade: um curso em nível intermediário*. [S.l.]: IMPA, 2002. (Projeto Euclides). Citado 3 vezes nas páginas 13, 15 e 16.

MEYER, P. L. *Probabilidade - Aplicações à Estatística*. [S.l.]: Livros Técnicos e Científicos Editora, 1983. Citado 2 vezes nas páginas 13 e 19.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <http://www.R-project.org/>. Citado na página 12.

ROSS, S. *Probabilidade: um curso moderno com aplicações*. [S.l.]: Bookman, 2010. Citado na página 18.