Universidade Federal de Campina Grande
Centro de Ciências e Tecnologia
Programa de Pós-Graduação em Matemática
Curso de Mestrado em Matemática

Adenilson Borba Lopes da Silva

A distribuição log-Bilal com zeros e/ou uns ajustados

Campina Grande - PB
2023

Universidade Federal de Campina Grande
Centro de Ciências e Tecnologia
Programa de Pós-Graduação em Matemática
Curso de Mestrado em Matemática

A distribuição log-Bilal com zeros e/ou uns ajustados

por

Adenilson Borba Lopes da Silva †

sob orientação do

Prof. Dr. Manoel Ferreira dos Santos Neto

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, como requisito parcial para obtenção do título de Mestre em Matemática.

† Este trabalho contou com o apoio financeiro parcial da FAPESQ

S586d Silva, Adenilson Borba Lopes da.
A distribuição log-Bilal com zeros e/ou uns ajustados / Adenilson Borba Lopes da Silva. – Campina Grande, 2023.
40 f. : il. color.

Dissertação (Mestrado em Matemática) – Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, 2023.
"Orientação: Prof. Dr. Manoel Ferreira dos Santos Neto".
Referências.

1. Probabilidade. 2. Estatística. 3. Distribuições log-Bilal. 4. Variáveis no Intervalo 0, 1. I. Santos Neto, Manoel Ferreira dos. II. Título.

CDU 519.2(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

A distribuição log-Bilal com zeros e/ou uns ajustados

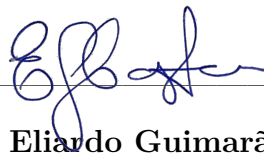
por

Adenilson Borba Lopes da Silva

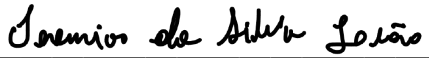
Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, como requisito parcial para obtenção do título de Mestre em Matemática.

Área de Concentração: Probabilidade e Estatística

Aprovada por:



Prof. Dr. Eliardo Guimarães da Costa



Prof. Dr. Jeremias da Silva Leão

Prof. Dr. Manoel Ferreira dos Santos Neto

Orientador

Universidade Federal de Campina Grande

Centro de Ciências e Tecnologia

Programa de Pós-Graduação em Matemática

Curso de Mestrado em Matemática

17 de Fevereiro de 2023

Resumo

No âmbito da modelagem de dados, é comum trabalhar com variáveis limitadas a um intervalo. Apesar de existir várias distribuições definidas especificamente para modelar esses tipos de dados, é comum dados como proporções, taxas e razões conterem zeros e/ou uns e isto inviabiliza a utilização de uma distribuição contínua para a modelagem. Sendo assim, nessa dissertação, foram propostas distribuições construídas a partir da mistura entre uma distribuição log-Bilal e uma distribuição Bernoulli degenerada em zero e/ou um, com o intuito de modelar dados no intervalo $[0, 1]$, $[0, 1)$ ou $(0, 1]$. Também foram realizadas simulações de Monte Carlo para estudar as propriedades dos estimadores de máxima verossimilhança. Por fim, apresentamos uma ilustração com um conjunto de dados reais.

Abstract

When modeling data, it is common to work with variables limited to an interval. Although there are several distributions defined specifically to model these types of data, it is normal data such as proportions, rates and ratios contain zeros and/or ones and this makes it unfeasible to use a continuous distribution for modeling. Thus, in this dissertation, was proposed distributions built from the mixture between a log-Bilal distribution and a Bernoulli distribution degenerated in zero and/or one, in order to model data in the interval $[0, 1]$, $[0, 1)$ or $(0, 1]$. Monte Carlo simulations were also performed to study the properties of the maximum likelihood estimators. Finally, we present an illustration with a real data set.

Agradecimentos

Primeiramente, agradeço a Deus por ter me dado saúde, sabedoria e dedicação para chegar até o final dessa etapa, sem ele nada disso seria possível.

A minha mãe, Denise Borba, aos meus irmãos Anderson e Ardson, e a minha família por sempre estarem ao meu lado dando apoio.

A minha esposa Simone que sempre esteve ao meu lado em toda minha jornada acadêmica.

Ao meu orientador prof. Dr. Manoel Santos Ferreira dos Santos Neto que me auxiliou na germinação das ideias durante todo o processo de desenvolvimento deste presente projeto. Agradeço por me orientar na construção deste trabalho, pelo apoio, dedicação e compreensão em todas as fases da realização desta dissertação.

Aos professores Eliardo Guimarães da Costa e Jeremias da Silva Leão pelos conhecimentos transmitidos e também por aceitarem o convite para participarem da banca examinadora, e por indicar as devidas correções do trabalho, muito obrigado.

Ao colega de Mestrado Iago Renan Valentim da Silva, pela convivência, troca de conhecimento e amizade desde a época de graduação.

Dedicatória

Dedico essa dissertação primeiramente a Deus que me deus saúde, conhecimento e dedicação. Dedico também à minha esposa, Simone que, com muito carinho e apoio, não mediu esforço para que eu chegasse até esta etapa de minha vida.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nos mais diferentes campos de atuação, muitas vezes, faz-se necessário modelar variáveis aleatórias contínuas, que assumem valores nos mais diversos intervalos. A modelagem de variáveis aleatórias contínuas que assumem valores em um intervalo é necessária para descrever muitos fenômenos naturais e artificiais que não podem ser representados por variáveis discretas. Dentro desse contexto, há os profissionais que precisam modelar variáveis que assumem valores no intervalo $(0, 1)$, como taxas, proporções e índices de concentração. Também existem casos em que a variável resposta é uma variável contínua que assume valores em uma abertura limitada ao intervalo (a, b). Para esse caso, uma transformação pode ser proposta, usando a transformação básica $(y - a)/(b - a)$, (sendo y o valor da variável) para que a variável mude seu intervalo para $(0, 1)$.

Para dados limitados a um intervalo, existem várias distribuições que são definidas especificamente para modelar esses tipos de dados. Um dos principais modelos para modelar dados no intervalo $(0, 1)$ é o modelo seguindo a distribuição Beta, uma vez que sua densidade pode assumir diversas formas dependendo dos valores dos parâmetros, no qual, esse modelo pode ser estendido a um modelo de regressão onde a variável resposta segue uma distribuição Beta, detalhes especificados por Ferrari & Cribari-Neto (2004).

Por outro lado, além dos modelos que seguem a distribuição Beta, várias distribuições foram propostas por pesquisadores com o intuito de aumentar a precisão da modelagem, algumas dessas distribuições são a Topp-Leone desenvolvida por Topp &

Leone (1955) e Kumaraswamy por Kumaraswamy (1980). Nesse contexto Cribari-Neto & Santos (2019) afirmaram que a distribuição Kumaraswamy se encaixa melhor que a distribuição Beta em dados nos quais suas observações são hidrológicas de pequena frequência e uma alternativa a essas distribuições é a unit-Birnbaum-Saunders proposta por Mazucheli et al. (2018), a qual é uma distribuição de dois parâmetros com domínio limitado e também apresenta a vantagem de não incluir nenhuma função especial ou parâmetros em sua formulação.

Altun et al. (2021) mencionam que, embora a distribuição Beta seja amplamente utilizada para modelar conjuntos de dados em intervalos limitados, ela tem deficiência para modelar conjuntos de dados extremamente assimétricos à esquerda e leptocúrticos. Assim como a distribuição Beta, as outras distribuições também têm problemas com relação a alguns contextos dentro do intervalo de dados distribuídos no intervalo $(0, 1)$. Para resolver esse problema com relação a dados extremamente assimétricos à esquerda e leptocúrticos, Altun et al. (2021) propuzeram, a partir da distribuição Bilal apresentada em Abd-Elrahman (2013), a distribuição log-Bilal. Adicionalmente, os autores desenvolveram algumas de suas propriedades estatísticas, como também, destacaram quatro motivos que refletem a importância desta distribuição, isto é, a log-Bilal.

Os motivos que resumem a importância da distribuição log-Bilal são: (i) a distribuição log-Bilal tem expressões simples e fechadas para suas funções estatísticas (ii) as propriedades da distribuição log-Bilal são derivadas em formas explícitas sem quaisquer funções matemáticas especiais, (iii) a distribuição proposta fornece mais flexibilidade do que as distribuições existentes para as formas da função de taxa de risco, (iv) graças as suas simples funções matemáticas, foi introduzido um novo modelo de regressão baseado na densidade log-Bilal para modelar as variáveis dependentes extremamente assimétricas com covariáveis associadas.

Neste trabalho, apresentamos uma adaptação ao processo gerador de números pseudoaleatórios para a distribuição log-Bilal. Em seguida, é proposto uma distribuição log-Bilal com zeros e/ou uns ajustados.

O restante do trabalho está dividio da seguinte maneira: No Capítulo 2 é apresentada a distribuição log-Bilal com zeros e/ou uns ajustados. Além disso, algumas propriedades são apresentadas. No Capítulo 3 é realizado um estudo de simulação de

Monte Carlo com o objetivo de estudar as propriedades dos estimadores de máxima verossimilhança. No Capítulo 4, é realizada uma ilustração usando um conjunto de dados reais. Por fim, são apresentadas as conclusões no Capítulo 5.

No que se refere aos aspectos computacionais do trabalho, este trabalho foi escrito usando o **Overleaf**, que é um editor LaTeX colaborativo baseado em nuvem usado para digitar, editar e publicar documentos. Já a estimação, as simulações, a aplicação e os gráficos foram feitos usando a linguagem de programação `R`.

Capítulo 2

Distribuição log-Bilal com zeros e/ou uns ajustados

2.1 A distribuição log-Bilal

Seja X uma variável aleatória (VA) com distribuição Bilal, então sua função função de densidade de probabilidade (FDP) é dada por

$$f(x;\theta) = \frac{6}{\theta}\exp\Big(-\frac{2x}{\theta}\Big)\left[1-\exp\Big(-\frac{x}{\theta}\Big)\right], \quad x>0,$$

em que $\theta > 0$ é o parâmetro de escala. A função de distribuição acumulada (FDA) de X é dada por

$$F(x) = 1-\exp\Big(-\frac{2x}{\theta}\Big)\left[3-2\exp\Big(-\frac{x}{\theta}\Big)\right].$$

Seguindo Altun & Hamedani (2018) e Altun (2021), temos que $Y = \exp(-X)$, então a função densidade de probabilidade (FDP) da distribuição log-Bilal é dada por

$$f(y;\theta) = \frac{6}{\theta}y^{2/\theta-1}\left(1-y^{1/\theta}\right), \quad 0<y<1, \tag{2.1}$$

em que $\theta > 0$. Aqui, o parâmetro θ se comporta como um parâmetro de forma em contraste com a distribuição Bilal. A partir de agora, a variável aleatória Y com densidade (2.1) é indicada como $Y \sim$ log-Bilal (θ). A FDA de Y para $(0<y<1)$ é

$$F(y;\theta) = 3y^{2/\theta} - 2y^{3/\theta}.$$

Na Figura 2.1, apresentamos formas assumidas pela distribuição log-Bilal a partir de diferentes valores do parâmetro θ. A partir desta, nota-se que a distribuição log-Bilal

pode ser usada para modelar diferentes tipos de dados definidos no intervalo $(0, 1)$, sejam estes com assimétria à direita, à esquerda ou simétricos neste intervalo.

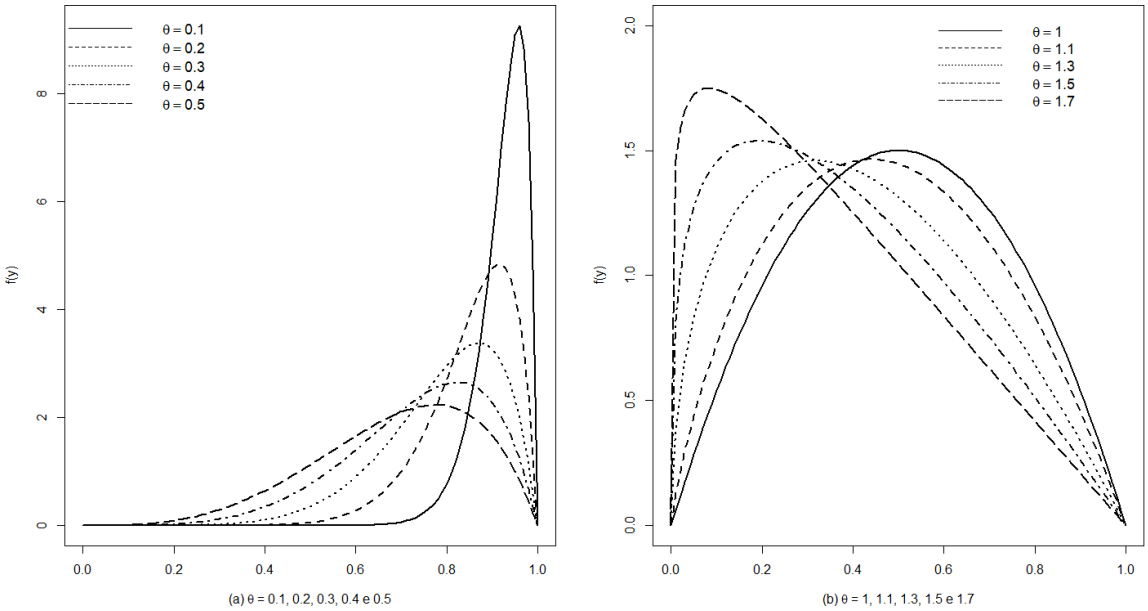


Figura 2.1: FDP de uma log-Bilal para diferentes valores de θ

Para gerar números aleatórios de uma distribuição log-Bilal, é apresentado em Altun et al. (2021) um procedimento baseado na função quantílica de Y. No entanto, não é uma solução dentro do conjunto dos números reais. Para gerar números pseudoaleatórios de uma distribuição log-Bilal numericamente, encontrar a raiz real y_0 $(0 < y_0 < 1)$ da equação

$$3y_0^{2/\theta} - 2y_0^{3/\theta} - U = 0, \tag{2.2}$$

é necessário, em que $U \sim \mathcal{U}(0, 1)$ e o valor do parâmetro θ sejam conhecidos. Podemos resolver a equação (2.2) usando o comando `uniroot()` do `R`.

```
rlogbilal<- function(Nsim,theta){

X <- NULL

for (i in 1:Nsim) {
    u <- runif(1)
    X[i] <- uniroot(function(x) 3*x^(2/theta) - 2*x^(3/theta) - u,
       lower = 0, upper = 1, tol = 1e-9 )$root
}
```

```
  return(X)
}
}
```

Neste trabalho, apresentamos também um procedimento utilizando o método de aceitação e rejeição para gerar números pseudoaleatórios da distribuição log-Bilal.

Algoritmo 1 Algoritmo para gerar valores pseudo-aleatórios da distribuição log-Bilal.

1: Defina um valor para o parâmetro θ;
2: Gerar um valor uniforme $y = u_1 \sim U(0,1)$ e fazer $f(u_1)$ e $g(u_1)$ nos quais f e g seguem uma distribuição log-Bilal;
3: Defina C como sendo Max da função (2.1);
4: Gerar um valor uniforme $u_2 \sim U(0,1)$;
5: Se $u_2 < \frac{f(u_1)}{Cg(u_1)}$ faça $x = y$, se não volte para o passo 2;
6: Repetir os passos 2, 3, 4 e 5 n vezes.

O Algoritmo 1 foi implementado na linguagem R. Foi criada uma função chamada de `rlogbilal()`, que recebe como argumentos o tamanho da amostra e o valor do parâmetro θ a ser considerado.

```
rlogbilal<- function(n,theta){

  fx1 <- function(x,theta){
    (6/theta)*(xˆ(2/theta-1))*(1-xˆ(1/theta))
  }

  max.bilal <- optimize(f = function(x) {fx1(x,theta)},
                        interval = c(0, 1), maximum =
                           TRUE);max.bilal
  # Agora podemos entao determinar o valor de C, como sendo
  c <- max.bilal$objective/1;c

  ## Define funcoes
  f<- function(x) {fx1(x,theta)}
  g <- function(x) 1 + 0 * x

  ## Amostra da proposta U(0,1)
  y <- runif(n)
  ## Amostra u tambem de U(0,1)
  u <- runif(n)
```

```
  ## Calcula a razao
  r <- f(y)/(c * g(y))
  ## x serao os valores de y onde u < r
  x <- y[u < r]
  return(x)
}
```

Na Figura 2.2, são apresentados três casos o quais são comparados a FDP da log-Bilal com os valores pseudoaleatórios gerados pelo método de aceitação e rejeição; os valores escolhidos para θ foram $\theta = 0.1, 0.5$ e 1. Observando os gráficos, é evidente que os valores gerados provêm da FDP da log-Bilal em questão.

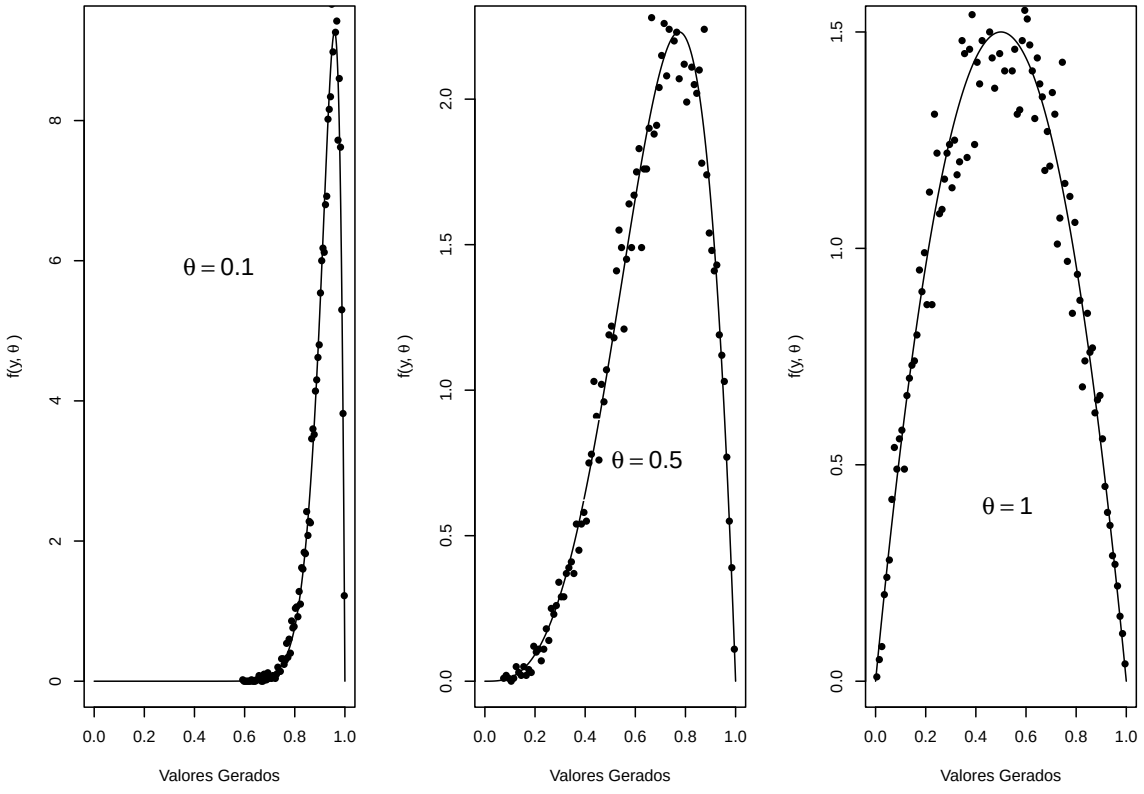


Figura 2.2: Geração de números pseudos aleatórios de uma log-Bilal

Se $Y \sim LB(\theta)$ então a média e a variância são dadas, respectivamente, por

$$E(Y) = \frac{6}{(\theta+2)(\theta+3)} \quad \text{e} \quad Var(Y) = \frac{3\theta^2(\theta^2+10\theta+13)}{(\theta^2+5\theta+6)^2(2\theta^2+5\theta+3)}.$$

Podemos reescrever a distribuição log-Bilal em termos de sua média. Basta fazer $\theta = \mu_1 = {}^1\!/_2\left(\sqrt{\frac{24+\mu}{\mu}} - 5\right)$ e consequentemente $E(Y) = \mu$ e $Var(Y) = \frac{3\mu_1^2(\mu_1^2+10\mu_1+13)}{(\mu_1^2+5\mu_1+6)^2(2\mu_1^2+5\mu_1+3)}$.

Além disso, temos que (2.1) pode ser reescrita como

$$f(y;\mu) = \frac{6}{1/2\left(\sqrt{\frac{24+\mu}{\mu}}-5\right)} y^{2/\left[1/2\left(\sqrt{\frac{24+\mu}{\mu}}-5\right)\right]-1}\left(1-y^{1/\left[1/2\left(\sqrt{\frac{24+\mu}{\mu}}-5\right)\right]}\right), \tag{2.3}$$

em que $0 < y < 1$. O logaritmo da FDP é dado por

$$\ell(\mu) = \log(12) - \log\left(\sqrt{\frac{\mu+24}{\mu}}-5\right) + \left(\frac{2}{\frac{1}{2}\left(\sqrt{\frac{\mu+24}{\mu}}-5\right)}-1\right)\log(y) + \log\left(1-y^{\frac{1}{\frac{1}{2}\left(\sqrt{\frac{\mu+24}{\mu}}-5\right)}}\right).$$

2.2 A distribuição log-Bilal com zeros ou uns ajustados

É normal ocorrer o feito de, dados como proporções, taxas e razões conterem zeros e/ou uns. Isto pode estar relacionado com alguma intervenção ou até mesmo censura nos dados. Tendo em vista situações como essas a distribuição log-Bilal não é uma distribuição que se adequa a esse tipo de dados. Se os dados contêm zeros ou uns (mas não ambos), um modelo natural consiste em adicionar à distribuição Log-Bilal um ponto de massa em zero ou um. Logo, podemos obter modelos para frações observadas nos intervalos $[0,1)$ ou $(0,1]$. Nessa circunstância, vamos supor que a base de dados seja modelada pela distribuição log-Bilal (2.3) já que essa distribuição, é bastante flexível quando se trata de ajustar dados no intervalo $(0,1)$. Dessa forma, o componente discreto, i.e., o ponto de massa, será modelado através de uma distribuição degenerada no valor conhecido c, onde c é igual a zero ou um dependendo do caso (ver Ospina 2008).

A FDA da distribuição log-Bilal com zeros ou uns ajustados é dada por

$$F(y \mid \mu, \alpha) = \alpha \mathbb{I}_{\{c\}}(y) + (1-\alpha) F(y \mid \mu), \tag{2.4}$$

em que $\mathbb{I}_A(y)$ representa uma função indicadora, com valor 1 se $y \in A$ e 0 caso contrário. O parâmetro $0 < \alpha < 1$ é conhecido como parâmetro de mistura e a função $F(\cdot \mid \mu)$ é a FDA da distribuição log-Bilal com média $0 < \mu < 1$. Como destacado em Ospina (2008) a função $F(y \mid \mu, \alpha)$ não é absolutamente contínua, uma vez que tem um ponto de massa em $Y = c$.

A FDP da distribuição log-Bilal com zero ou um ajustados pode ser escrita da seguinte forma

$$f(y \mid \mu, \alpha) = \left\{\alpha^{\mathbb{I}_{\{c\}}(y)}(1-\alpha)^{1-\mathbb{I}_{\{c\}}(y)}\right\}\left\{f(y \mid \mu)^{1-\mathbb{I}_{\{c\}}(y)}\right\}. \tag{2.5}$$

em que $f(y \mid \mu)$ é a FDP dada em (2.3). Observe que o parâmetro de mistura α representa exatamente a probabilidade de observação do zero ($c = 0$) ou do um ($c = 1$). Note que (2.5) pode ser fatorada em termos apenas de uma função que envolve apenas α e outra que depende apenas de μ.

Considere os seguites casos: se $c = 0$ obteremos a distribuição log-Bilal com zeros ajustados (LBZA); e se $c = 1$ temos a distribuição log-Bilal com uns ajustados (LBUA). Desta forma, a distribuição apresentada em (2.5) permite modelar os exessos de zeros ou uns, de acordo com as características dos dados.

Temos que a média e a variância da variável aleatória Y são respectivamente

$$\begin{aligned} E(Y) &= \alpha c + (1-\alpha)\,\mu, \quad \text{e} \\ Var(Y) &= (1-\alpha)\left(\frac{3\mu_1^2(\mu_1^2+10\mu_1+13)}{(\mu_1^2+5\mu_1+6)^2(2\mu_1^2+5\mu_1+3)}\right) + \alpha(1-\alpha)(c-\mu)^2, \end{aligned} \tag{2.6}$$

em que o valor de c dependerá da natureza dos dados estudados, podendo assumir os valores 0 ou 1. Além disso, é possível notar em (2.6) que a média de Y é uma média ponderada entre a média da variável degenerada em c e a média da distribuição absolutamente contínua.

Na Figura 2.3, estão os gráficos referentes aos valores gerados a partir de uma LBZA. Nele estão quatro diferentes gráficos, no qual foi fixado o valor de $\alpha = 0.4$ e testado diferentes valores para μ, os valores testados foram $\mu = 0.4,\ 0.5,\ 0.7$ e 0.9. Observando o gráfico, é nítida a flexibilidade que tem a distribuição LBZA, no qual a distribuição pode assumir tanto a forma simétrica quanto a forma assimétrica, seja assimetria a esquerda ou à direita.

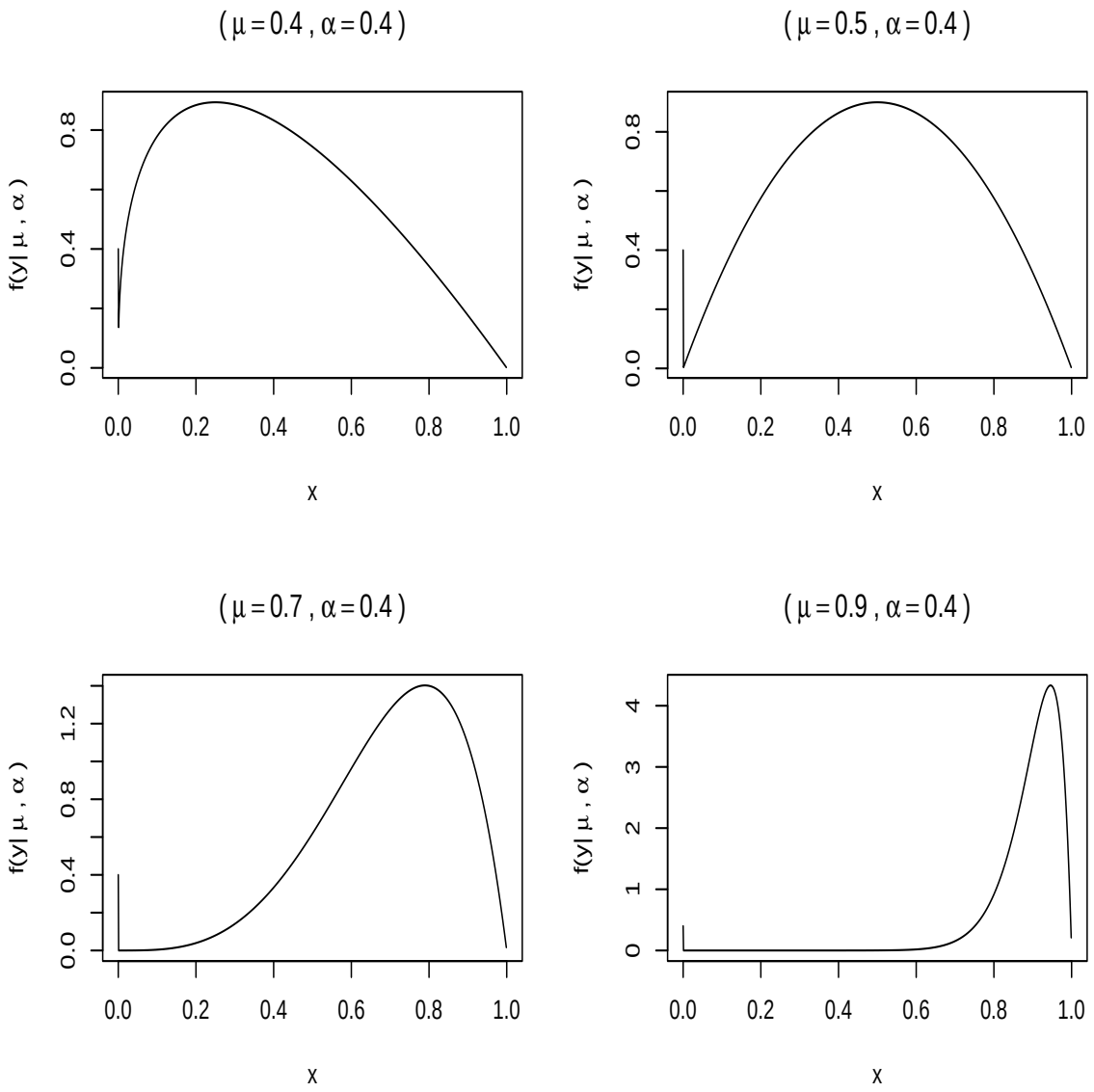


Figura 2.3: FDP de uma LBZA para diferentes valores de μ

Sejam $Y_1, \ldots, Y_n$ uma amostra aleatória de uma distribuição log-Bilal com zeros ou uns ajustados com parâmetros μ e α. A função de verossimilhança, $L(\mu, \alpha)$, pode ser escrita da seguinte maneira

$$L(\mu, \alpha) = L(\alpha)L(\mu),$$

em que $L(\alpha)$ e $L(\mu)$ são as funções de verossimilhança da variável degenerada em c e

da variável absolutamente contínua, respectivamente, ou seja,

$$L(\alpha) = \prod_{i=1}^{n} \alpha^{\mathbb{I}_{\{c\}}(y_i)} (1-\alpha)^{1-\mathbb{I}_{\{c\}}(y_i)} = \alpha^{n_c}(1-\alpha)^{n-n_c},$$
$$L(\mu) = \prod_{i=1}^{n} f(y_i \mid \mu)^{1-\mathbb{I}_{\{c\}}(y_i)},$$

em que $n_c = \sum_{i=1}^{n} \mathbb{I}_{\{c\}}(y_i)$.

Ospina (2008) destaca que quando a função de verossimilhança é escrita desta forma os parâmetros são separáveis e a estimações de α e μ podem ser realizadas de forma independente uma da outra. Para mais detalhes ver também (Pace & Salvan 1997, p. 128).

O logaritmo da função de verossimilhança assume a seguinte forma

$$\ell(\mu, \alpha) = \ell(\alpha) + \sum_{i=1}^{n} [1 - \mathbb{I}_{\{c\}}(y_i)] \ell(\mu),$$

em que

$$\ell(\alpha) = n_c \log \alpha + (n - n_c) \log(1-\alpha)$$
$$\ell(\mu) = \log(12) - \log\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right) + \left(\frac{2}{\frac{1}{2}\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right)} - 1\right) \log(y_i) + \log\left(1 - y_i^{\frac{1}{\frac{1}{2}\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right)}}\right).$$

Derivando $\ell(\mu, \alpha)$ em relação aos parâmetros α e μ é possível obter o vetor escore. Desta maneira, o vetor escore é igual a

$$U(\mu, \alpha) = (U(\alpha), U(\mu)),$$

em que,

$$U(\alpha) = \frac{n_c}{\alpha} - \frac{n - n_c}{1-\alpha},$$
$$U(\mu) = \frac{12(n - n_c)}{\mu\left(\mu + 24 - 5\mu\sqrt{\frac{\mu+24}{\mu}}\right)} + \frac{48\sum_{i=1}^{n}[1 - \mathbb{I}_{\{c\}}(y_i)] y_i}{\mu^2\sqrt{\frac{\mu+24}{\mu}}\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right)}$$
$$+ \frac{24}{\mu^2\sqrt{\frac{24+\mu}{\mu}}\left(\sqrt{\frac{24+\mu}{\mu}} - 5\right)^2} \sum_{i=1}^{n} [1 - \mathbb{I}_{\{c\}}(y_i)] \frac{y_i^{\frac{1}{\frac{1}{2}\sqrt{\frac{24+\mu}{\mu}} - 5}} \log y_i}{1 - y_i^{\frac{1}{\frac{1}{2}\sqrt{\frac{24+\mu}{\mu}} - 5}}}.$$

Note que a solução do sistema $U(\alpha) = 0$ tem forma fechada. O estimador de máxima verossimilhança de α é $\hat{\alpha} = n_c/n$, ou seja, a proporção de zeros ou uns na amostra. O estimador de máxima verossimilhança para μ deve ser encontrado utilizando algum método iterativo, pois o sistema $U(\mu) = 0$ não tem uma solução analítica. Neste trabalho, optamos por colocar a distribuição na estrutura da família de distribuições do pacote `gamlss` (Rigby & Stasinopoulos 2005). O código está presente no Apêndice A. Para obter as estimativas de máxima verossimilhança basta usar a função `gamlssInf0to1()` do pacote `gamlss.inf` (Enea et al. 2019). As estimativas de máxima verossimilhança podem ser obtidas através dos algoritmos RS, CG ou uma mistura dos dois (Rigby & Stasinopoulos 2005).

2.3 A distribuição log-Bilal com zeros e uns ajustados

Considere agora a situação no qual exista uma probabilidade positiva de ocorrer 0 e 1. Desta forma, para modelar dados dessa natureza, pode-se utilizar uma mistura entre uma distribuição Bernoulli e uma distribuição log-Bilal. No entanto, diferentemente da Seção anterior, a distribuição Bernoulli atribui probabilidades não-negativas aos valores 0 e 1. Portanto, a FDA da distribuição log-Bilal com zero e uns ajustados (LBZUA) é dada por

$$F(y \mid \mu, \alpha, \lambda) = \alpha F(y \mid \lambda) + (1 - \alpha) F(y \mid \mu), \tag{2.7}$$

no qual $F(y \mid \lambda)$ é a FDA da distribuição Bernoulli com parâmetro λ. Como na Seção anterior, $0 < \alpha < 1$ é o parâmetro de mistura e a função $F(\cdot \mid \mu)$ é a FDA da distribuição log-Bilal com média $0 < \mu < 1$. Como destacado em Ospina (2008) a função $F(y \mid \mu, \alpha, \lambda)$ não é absolutamente contínua, uma vez que tem pontos de massa em $Y = 0$ e $Y = 1$.

A FDP da distribuição LBZUA pode ser escrita da seguinte forma

$$f(y \mid \mu, \alpha, \lambda) = \begin{cases} \alpha(1 - \lambda), & \text{se } y = 0, \\ \alpha\lambda, & \text{se } y = 1, \\ (1 - \alpha) f(y \mid \mu), & \text{se } y \in (0, 1), \end{cases} \tag{2.8}$$

em que $f(y \mid \mu)$ é a FDP dada em (2.3). Temos que a média e a variância da variável aleatória Y, são respectivamente

$$E(Y) = \alpha\lambda + (1-\alpha)\,\mu, \quad \text{e}$$

$$Var(Y) = \alpha\lambda(1-\lambda) + (1-\alpha)\left(\frac{3\mu_1^2(\mu_1^2+10\mu_1+13)}{(\mu_1^2+5\mu_1+6)^2(2\mu_1^2+5\mu_1+3)}\right) + \alpha(1-\alpha)(\lambda-\mu)^2.$$

Assim como na Figura 2.3, a Figura 2.4 estão os gráficos referentes aos valores gerados a partir de uma LBZUA, nela estão quatro diferentes gráficos, no qual foi fixado o valor de $\alpha = 0.2$ e $\lambda = 0.2$, para assim ser testado diferentes valores para μ, os valores testados foram $\mu = 0.4,\ 0.5,\ 0.7$ e 0.9. Observando os gráficos, é evidente que o formato dos gráficos não se alteraram comparando com os gráficos provenientes de uma LBZA, a diferença está nas proporções associadas aos valores extremos (0 e 1).

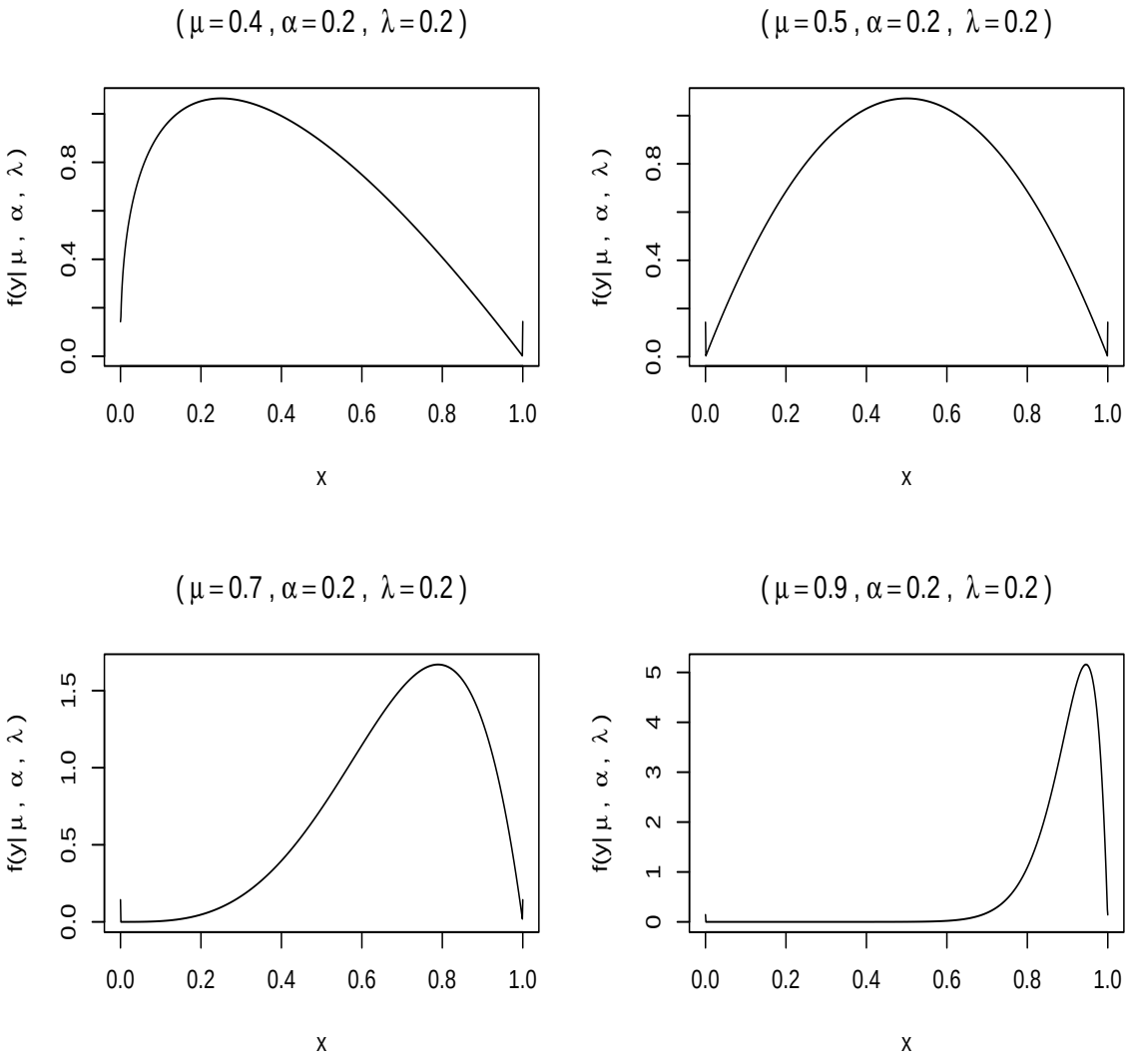


Figura 2.4: FDP de uma LBZUA para dferentes valores de μ

Sejam $Y_1, \ldots, Y_n$ uma amostra aleatória de uma distribuição LBZUA com parâmetros μ, α e λ. A função de verossimilhança, $L(\mu, \alpha, \lambda)$, pode ser escrita da seguinte maneira

$$L(\mu, \alpha, \lambda) = L(\alpha)L(\lambda)L(\mu),$$

em que $L(\alpha)$, $L(\lambda)$ e $L(\mu)$ são dadas por

$$\begin{aligned} L(\alpha) &= \alpha^{n_0+n_1}(1-\alpha)^{n-n_{01}}, \\ L(\lambda) &= \lambda^{n_1}(1-\lambda)^{n_{01}-n_1} = \lambda^{n_1}(1-\lambda)^{n_0}, \\ L(\mu) &= \prod_{i=1}^{n} f\left(y_i \mid \mu\right)^{1-\mathbb{I}_{\{0,1\}}(y_i)}, \end{aligned}$$

em que $n_{01} = \sum_{i=1}^{n} \mathbb{I}_{\{0,1\}}\left(y_i\right)$, $n_1 = \sum_{i=1}^{n} \mathbb{I}_{\{1\}}\left(y_i\right)$ e $n_0 = \sum_{i=1}^{n} \mathbb{I}_{\{0\}}\left(y_i\right)$, que são, respectivamente, os números de zeros e de uns na amostra, o número de uns na amostra e o número de zeros na amostra.

De maneira semelhante a Seção anterior, temos que a função de verossimilhanças é obtida como apresentado na expressão anterior, e os parâmetros são obtidos de maneira independente considerando cada uma das partes que compõem a função de verossimilhanças. O logaritmo da função de verosimilhanças é dado por

$$\ell\left(\mu, \alpha, \lambda\right) = \ell\left(\alpha\right) + \ell\left(\lambda\right) + \sum_{i=1}^{n}[1 - \mathbb{I}_{\{0,1\}}\left(y_i\right)]\ell(\mu),$$

em que

$$\begin{aligned} \ell\left(\alpha\right) &= \left(n_0 + n_1\right)\log\alpha + \left(n - n_{01}\right)\log\left(1-\alpha\right), \\ \ell\left(\lambda\right) &= n_1\log(\lambda) + n_0\log\left(1-\lambda\right), \\ \ell\left(\mu\right) &= \log\left(12\right) - \log\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right) + \left(\frac{2}{\frac{1}{2}\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right)} - 1\right)\log\left(y_i\right) + \\ &\quad \log\left(1 - y_i^{\frac{1}{\frac{1}{2}\left(\sqrt{\frac{\mu+24}{\mu}}-5\right)}}\right). \end{aligned}$$

Obtendo-se a primeira derivada de $\ell\left(\mu, \alpha, \lambda\right)$ em relação aos parâmetros α, λ e μ é possível obter o vetor escore. Desta maneira, o vetor escore é igual a

$$U\left(\mu, \alpha\right) = \left(U\left(\alpha\right), U\left(\lambda\right), U\left(\mu\right)\right),$$

em que,

$$
\begin{aligned}
U(\alpha) =& \frac{(n_0 + n_1)}{\alpha} - \frac{(n - n_{01})}{1-\alpha}, \\
U(\lambda) =& \frac{n_1}{\lambda} - \frac{n_0}{1-\lambda}, \\
U(\mu) =& \frac{12\,(n - n_{01})}{\mu\left(\mu + 24 - 5\mu\sqrt{\frac{\mu+24}{\mu}}\right)} + \frac{48\sum_{i=1}^{n}[1 - \mathbb{I}_{\{0,1\}}(y_i)]y_i}{\mu^2\sqrt{\frac{\mu+24}{\mu}}\left(\sqrt{\frac{\mu+24}{\mu}} - 5\right)} \\
&+ \frac{24}{\mu^2\sqrt{\frac{24+\mu}{\mu}}\left(\sqrt{\frac{24+\mu}{\mu}} - 5\right)^2}\sum_{i=1}^{n}[1 - \mathbb{I}_{\{0,1\}}(y_i)]\frac{y_i^{\frac{1}{\frac{1}{2}\sqrt{\frac{24+\mu}{\mu}-5}}}\log y_i}{1 - y_i^{\frac{1}{\frac{1}{2}\sqrt{\frac{24+\mu}{\mu}-5}}}}.
\end{aligned}
$$

Note que as soluções dos sistemas $U(\alpha) = 0$ e $U(\lambda) = 0$ têm forma fechada e portanto, os estimadores de máxima verossimilhança de α e λ, são dados, respectivamente, por $\hat{\alpha} = n_{01}/n$ e $\hat{\lambda} = n_1/(n_0 + n_1)$, isto é, a proporção de zeros e uns na amostra e a proporção de uns no grupo de observações que são 0 ou 1. O estimador de máxima verossimilhança para μ deve ser encontrado utilizando algum método iterativo, pois o sistema $U(\mu) = 0$. Neste trabalho, optamos por colocar a distribuição na estrutura da família de distribuições do pacote `gamlss` (Rigby & Stasinopoulos 2005). O código é apresentado no Apêndice A. Para gerar as funções relacionadas à distribuição log-Bilal com zeros e uns ajustados, basta usar a função `gen.Inf0to1()` com o argurmento `type.of.Inflation = "Zero&One"`. Para obter as estimativas de máxima verossimilhança, basta usar a função `gamlssInf0to1()` do pacote `gamlss.inf` (Enea et al. 2019). As estimativas de máxima verossimilhança podem ser obtidas através dos algoritmos RS, CG ou uma mistura dos dois (Rigby & Stasinopoulos 2005).

Capítulo 3

Simulações

A eficiência dos estimadores de máxima verosimilhança dos parâmetros do modelo proposto foi analisada através de simulação de Monte Carlo. Para isso, utilizamos o pacote `R` chamado `gamlss.inf` (Enea et al. 2019). Neste pacote, existe uma função chamada `gen.Inf0to1()`, a qual permite gerar distribuições com zero e/ou uns ajustados. No entanto, é necessário ter a distribuição que é absolutamente contínua implementada na estrutura da família de distribuições presentes no pacote `gamlss` (Rigby & Stasinopoulos 2005). Isto foi feito com a distribuição log-Bilal indexada pela média. Desta forma, é possível gerar facilmente as distribuições LBZA, LBUA e LBZUA. Veja nos códigos abaixo

Listagem 3.1: Gerando novas distribuições.

```
library(gamlss.inf)

gen.Inf0to1(family  = "LB", type.of.Inflation = "Zero")
A  0 inflated LB distribution has been generated
 and saved under the names:
 dLBInf0 pLBInf0 qLBInf0 rLBInf0
 plotLBInf0

 gen.Inf0to1(family  = "LB", type.of.Inflation = "One")
A  1 inflated LB distribution has been generated
 and saved under the names:
 dLBInf1 pLBInf1 qLBInf1 rLBInf1
 plotLBInf1
```

```
14
15  gen.Inf0to1(family  = "LB", type.of.Inflation = "Zero&One")
16 A  0to1 inflated LB distribution has been generated
17  and saved under the names:
18  dLBInf0to1 pLBInf0to1 qLBInf0to1 rLBInf0to1
19  plotLBInf0to1
```

Nas simulações, iremos nos concentrar apenas nas distribuições LBZA e LBUA. Os resultados da simulação são interpretados com base nas seguintes quantidades: viés e a raiz do erro quadrático médio (REQM).

$$\text{Viés} = \sum_{i=1}^{N} \frac{\hat{\theta}_i - \theta}{N},$$

$$\text{REQM} = \sqrt{\sum_{i=1}^{N} \frac{\left(\hat{\theta}_i - \theta\right)^2}{N}}.$$

Para obter as estimativas de máxima verossimilhança, foi utilizada a função `gamlssInf0to1()`, do pacote `gamlss.inf`. Os valores escolhidos para o parâmetro μ foram $\mu = (0.2, 0.5, 0.8)$, e para α foram $\alpha = (0.1, 0.5, 0.9)$. O número de réplicas de Monte Carlo foi $N = 5000$ e os tamanhos da amostra foram $n = 20, 30, 40, \ldots, 200$. Para a realização das simulações de Monte Carlo utilizamos a função `MonteCarlo()` do pacote `MonteCarlo` (Leschinski 2019). O viés e a REQM foram obtidos usando a função `calc_absolute()` do pacote `simhelpers` (Joshi & Pustejovsky 2022). Uma parte dos códigos das simulações é apresentado a seguir.

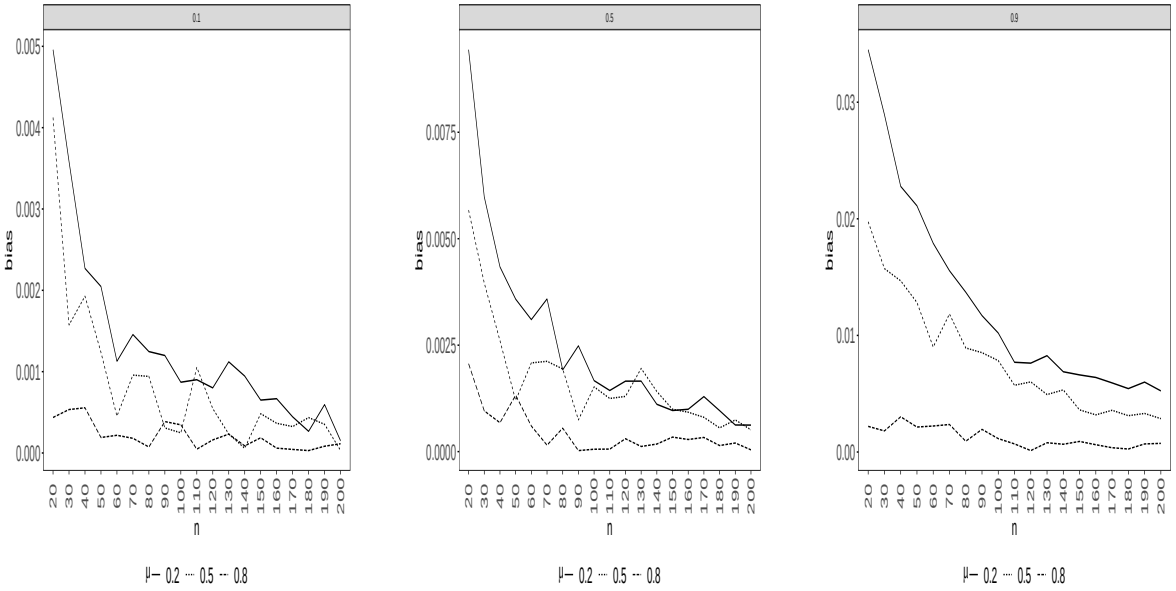
Listagem 3.2: Simulações de Monte Carlo.

```
1  library(MonteCarlo)
2  library(tidyverse)
3  library(ggpubr)
4  library(simhelpers)
5  library(gamlss.inf)
6  library(boot)
7  # Funcoes auxiliares
8  my.gamlssInf0to1 <- function(...) tryCatch(expr =
       gamlssInf0to1(...), error = function(e) NA)
9
10 #Criando a familia com zeros ajustados
11 gen.Inf0to1("LB","Zero") #zero ajusted log-bilal distribution
12
```

```
13
14 est <- function(n, mu, alpha){
15
16   repeat{
17     y <- rLBInf0(n, mu = mu, xi0 = alpha)
18     fit <- my.gamlssInf0to1(y = y, mu.formula =~1, xi0.formula
          =~1, family = LB(mu.link = "identity"))
19
20     if (all(is.na(fit)) == FALSE) break
21
22   }
23
24   coef_mle <-  unname(c(fit$mu.coefficients, fit$xi0.coefficients))
25   #bias <- coef_mle - v_theta
26
27   return(list("mean" = coef_mle[1], "prop0" =
       inv.logit(coef_mle[2]) ))
28 }
29
30 #Simulacao zeros ajustados
31
32 n_grid <- seq(20, 200, by = 10)
33 mu_grid <- c(0.2, 0.5, 0.8)
34 alpha_grid <- c(0.1, 0.5, 0.9)
35 param_list <- list("n" = n_grid, "mu" = mu_grid, "alpha" =
     alpha_grid)
36 MC_result <- MonteCarlo(func = est, nrep = 5000, param_list =
     param_list)
37
38 df <- MakeFrame(MC_result)
39 df$alpha<- df$p0
40 resul_mu <- df %>% group_by(n, mu, alpha) %>%
41   do(calc_absolute(., estimates = mean, true_param = mu,
       perfm_criteria = c("bias", "rmse")))  %>%
42   mutate(mu = as.character(mu), alpha = as.character(alpha), bias
       = abs(bias))
```

A Figura 3.1, apresenta o viés do estimador de máxima verossimilhança do parâmetro μ de uma LBZA para diferentes tamanhos amostrais, como também os valores de μ e α definidos anteriormente. Nota-se que a magnitude do víes do estimador $\hat{\mu}$ aumenta com o incremento da proporção de zeros na amostra. Este comportamento

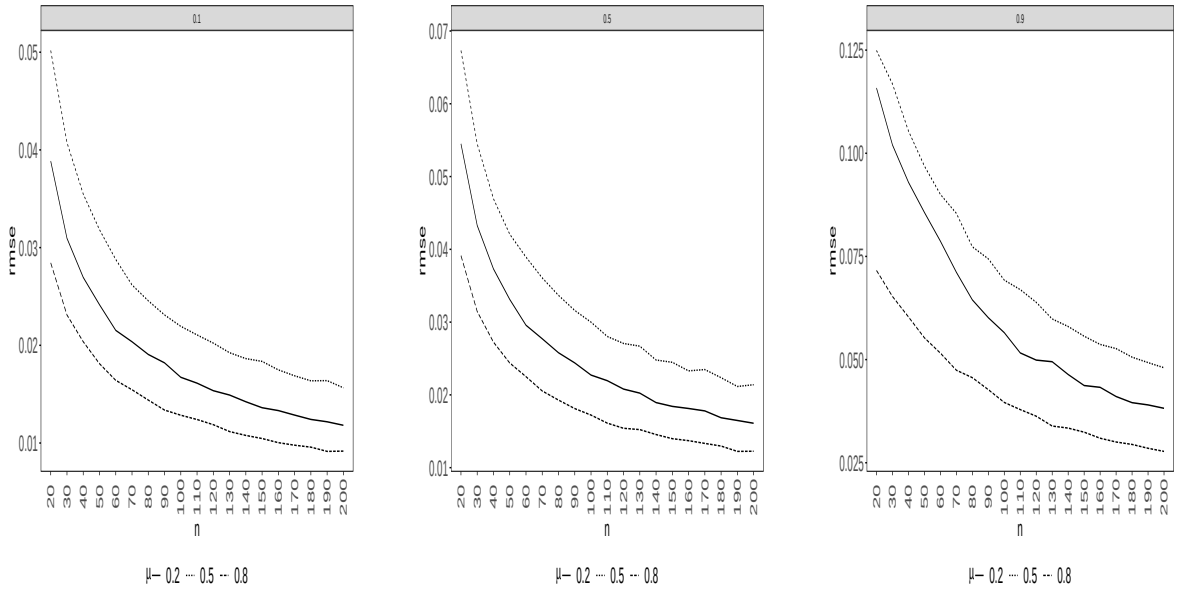
é esperado, uma vez que o aumento da quantidade de zeros, reduz a quantidade de observações positivas na amostra. Para α fixo, de uma maneira geral o víes de $\hat{\mu}$ é maior quando $\mu = 0.2$. Além disso, com o aumento do tamanho da amostra, notamos que, para todos os cenários, o viés é reduzido. Importante destacar que a velocidade dessa redução é menor a medida que a proporção de zeros aumenta.



(a) $\alpha = 0.1$. *(b)* $\alpha = 0.5$. *(c)* $\alpha = 0.9$.

Figura 3.1: Viés de $\hat{\mu}$ para diferentes valores de n, μ e α de uma LBZA.

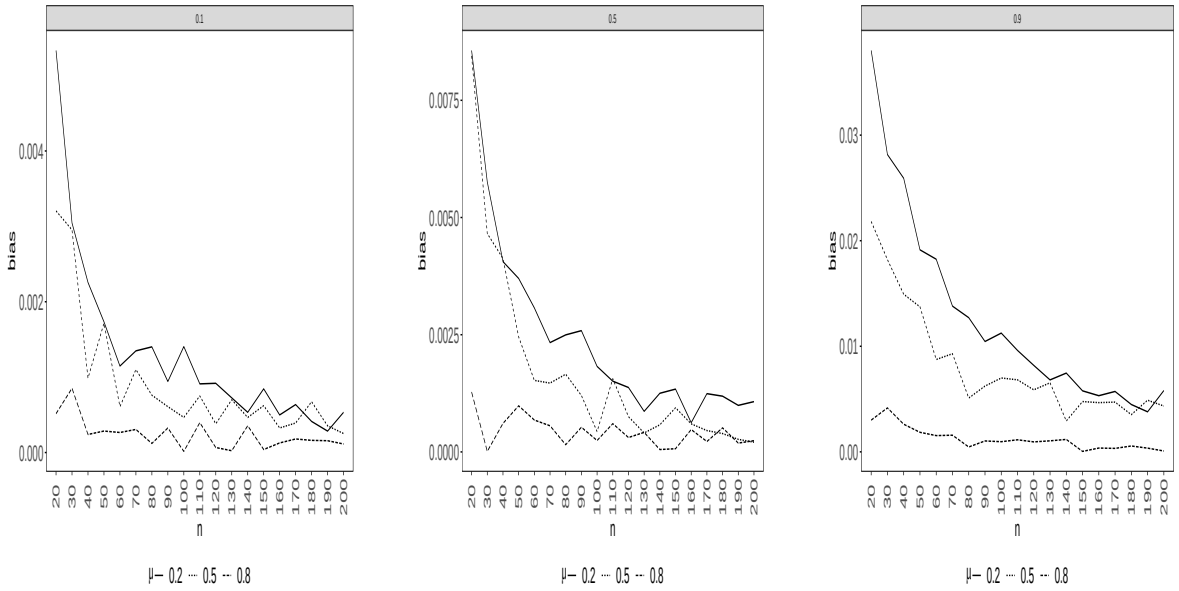
A Figura 3.2 apresenta o REQM do estimador de máxima verossimilhança do parâmetro μ para os cenários considerados inicialmente. A partir desta figura podemos notar um comportamento de acordo com o esperado. Ao fixar o valor de α o maior valor do REQM foi para $\mu = 0.5$ e menor quando $\mu = 0.8$. Para concluir, o REQM, assim como o Viés, diminui em todos os cenários com o aumentar do tamanho amostral.



(a) $\alpha = 0.1$. *(b)* $\alpha = 0.5$. *(c)* $\alpha = 0.9$.

Figura 3.2: REQM de $\hat{\mu}$ para diferentes valores de n, μ e α de uma LBZA.

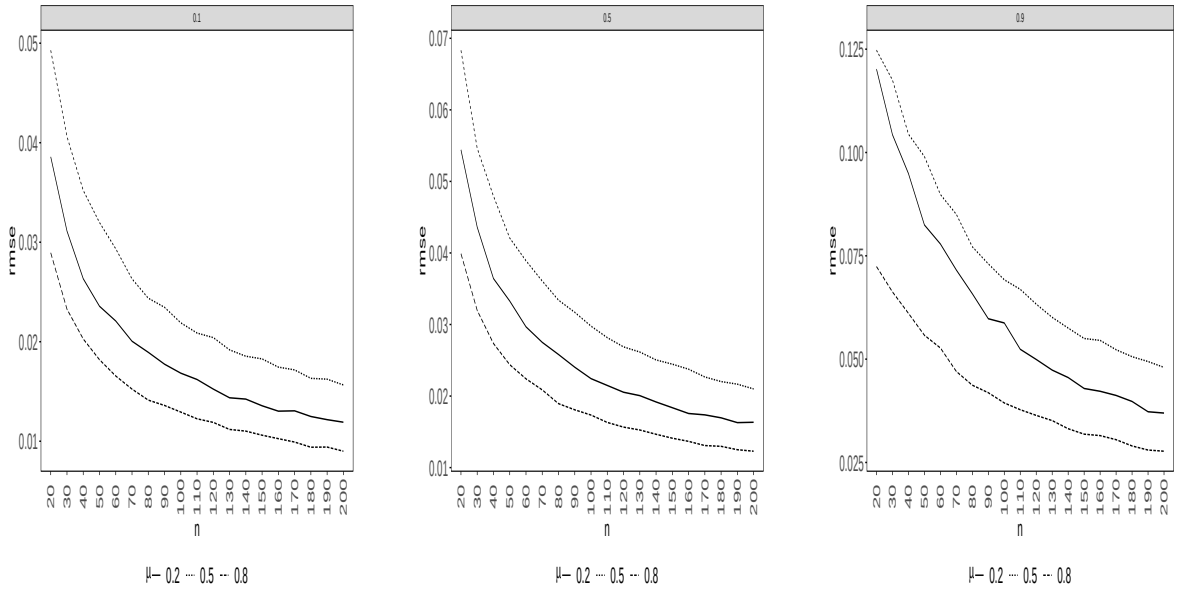
Da mesma forma que foi estimado o parâmetro μ da LBZA, também foi utilizado o estimador de máxima verossimilhança para estimar o parâmetro μ de uma LBUA. As Figuras 3.3 e 3.4 apresentam os resultados obtidos para a estimativa de μ da LBUA. A Figura 3.3 mostra os resultados referente ao Viés de $\hat{\mu}$ para diferentes valores de μ, α e tamanhos amostrais, enquanto a Figura 3.4 nos mostra os resultados referentes ao REQM de $\hat{\mu}$ utilizando os mesmos valores de μ, α e tamanhos amostrais usados para cálculo do viés.



(a) $\alpha = 0.1$. *(b)* $\alpha = 0.5$. *(c)* $\alpha = 0.9$.

Figura 3.3: Viés de $\hat{\mu}$ *para diferentes valores de* n, μ *e* α *de uma LBUA.*

Assim como no $\hat{\mu}$ da LBZA, o viés do estimado de máxima verossimilhança de μ de uma LBUA aumenta de magnitude conforme aumenta a proporção de uns na amostra. Fixando o α, o viés de $\hat{\mu}$ também é maior quando $\mu = 0.2$. Além do mais, para todos os cenários testados, houve redução do viés a médida que o tamanho amostral aumentou.

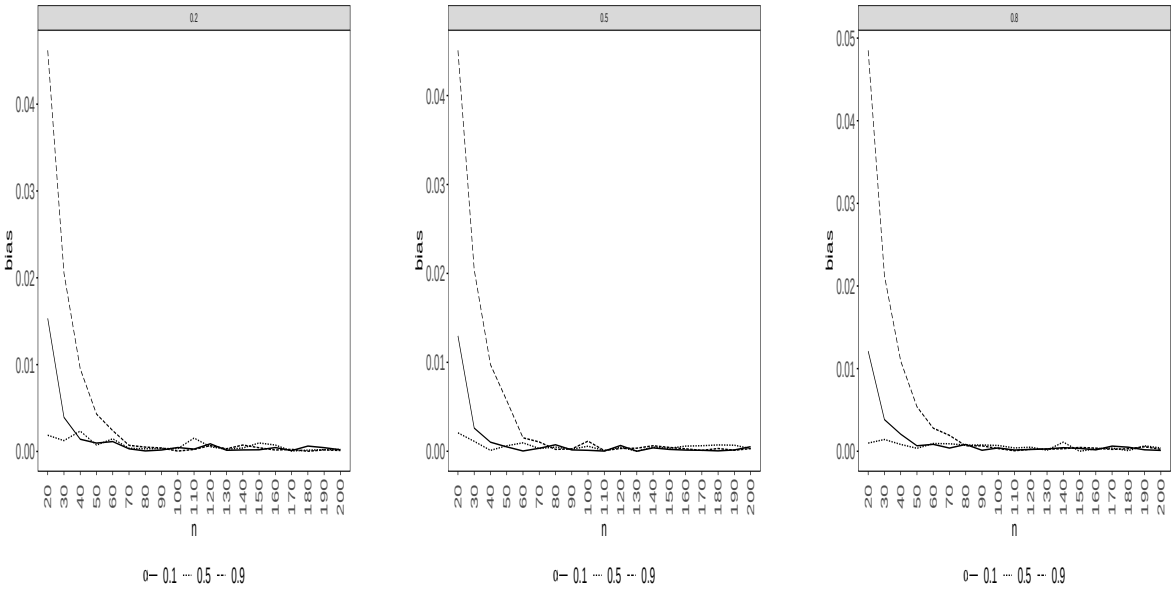


(a) $\alpha = 0.1$. *(b)* $\alpha = 0.5$. *(c)* $\alpha = 0.9$.

Figura 3.4: REQM de $\hat{\mu}$ *para diferentes valores de* n, μ *e* α *de uma LBUA.*

Na Figura 3.4, se trata do REQM de $\hat{\mu}$ da LBUA para diferentes valores de μ, α e amostral. De maneira similar a LBZA, o REQM de $\hat{\mu}$ diminui com o aumentar do tamanho amostral, como também tende a aumentar a médida que a proporção de uns aumenta na amostra. Ao fixar o valor de α, assim como na LBZA, o maior valor do REQM foi quando $\mu = 0.5$ e menor valor para $\mu = 0.8$.

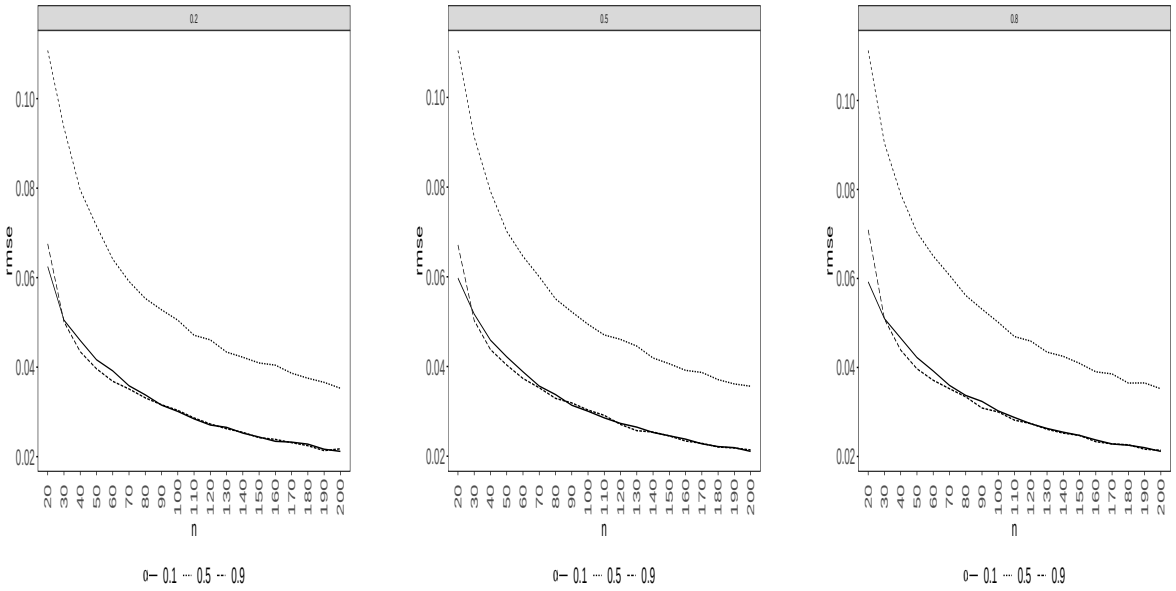
Se tratando do parâmetro α, na Figura 3.5, é apresentado o viés do estimador de máxima verossimilhança do parâmetro α de uma LBZA para diferentes valores amostrais, de μ e de α. Note que, a magnitude do viés do estimador $\hat{\alpha}$ se mantem estável com a variação de μ. Para μ fixo, de uma maneira geral, o víes de $\hat{\alpha}$ é maior quando $\alpha = 0.9$. Além disso, assim como ocorreu com $\hat{\mu}$, percebe-se que para todos os cenários testados o viés diminui a medida que o tamanho amostral aumenta.



(a) $\mu = 0.2$. *(b)* $\mu = 0.5$. *(c)* $\mu = 0.8$.

Figura 3.5: Viés de $\hat{\alpha}$ para diferentes valores de n, μ e α de uma LBZA.

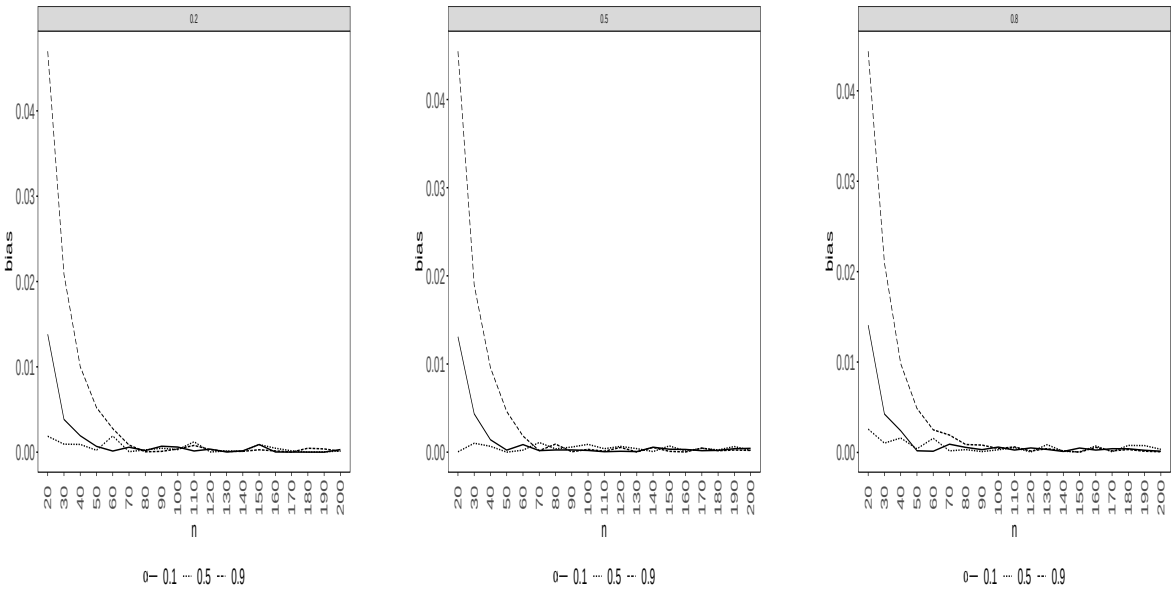
Na Figura 3.6, mostra o REQM do estimador de máxima verossimilhança do parâmetro α de uma LBZA para diferentes valores de amostras, de μ e α. De maneira similar ao viés, a significância do REQM se mantém estável para os diferentes valores de μ. Ao fixar o valor de μ, assim como no viés, o maior valor do REQM foi para $\alpha = 0.9$. Para concluir, o REQM também diminui a medida que aumenta o tamanho amostral.



(a) $\mu = 0.2$. *(b)* $\mu = 0.5$. *(c)* $\mu = 0.8$.

Figura 3.6: REQM de $\hat{\alpha}$ *para diferentes valores de* n, μ *e* α *de uma LBZA.*

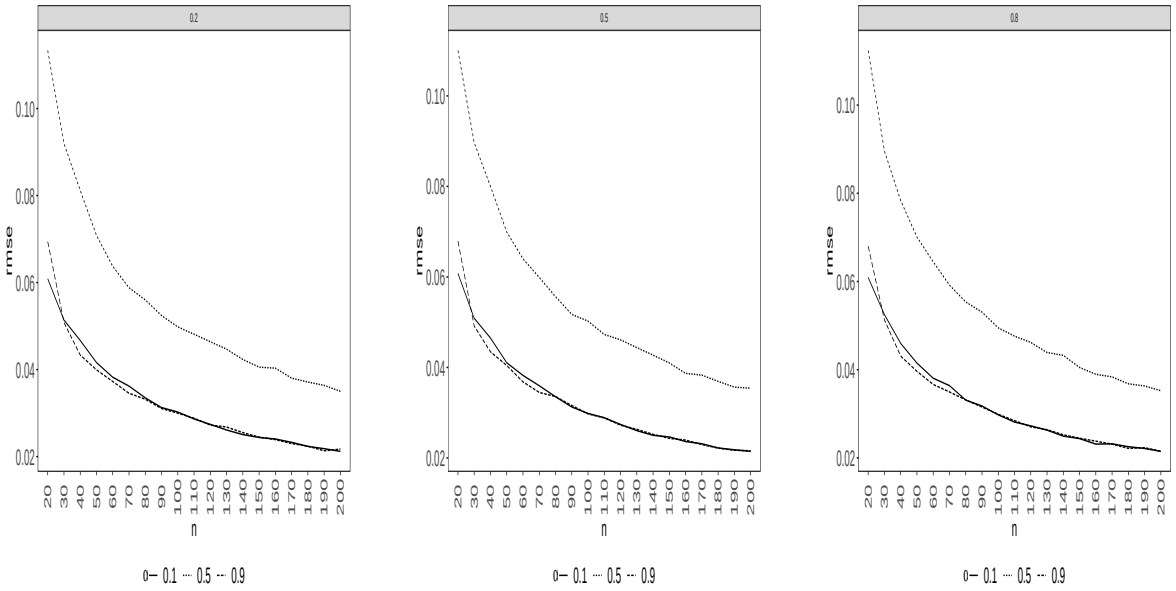
As figuras 3.7 e 3.8, apresentam o viés e REQM de $\hat{\alpha}$ de uma LBUA. A Figura 3.7, mostra os resultados referente ao viés de $\hat{\alpha}$ para diferentes valores de μ, α e tamanhos amostrais. Enquanto a Figura 3.8 nos mostra os resultados referentes ao REQM de $\hat{\alpha}$ utilizando os mesmos valores de μ, α e tamanhos amostrais usados para cálculo do viés.



(a) $\mu = 0.2$. *(b)* $\mu = 0.5$. *(c)* $\mu = 0.8$.

Figura 3.7: Viés de $\hat{\alpha}$ *para diferentes valores de* n, μ *e* α *de uma LBUA.*

De acordo com a figura 3.7, nota-se que, o viés do estimador de máxima verossimilhança de α da LBUA, permanece consistente a medida que o μ aumenta. Quando fixado o μ, o viés de $\hat{\alpha}$ é maior quando $\alpha = 0.9$ e menor quando $\alpha = 0.5$. Além do mais, para todos os cenários testados, houve redução do viés a medida que o tamanho amostral aumentou, também vale destacar que o viés não ultrapassou 0.05, mesmo quando o tamanho amostral foi 20.



(a) $\mu = 0.2$. *(b)* $\mu = 0.5$. *(c)* $\mu = 0.8$.

Figura 3.8: REQM de $\hat{\alpha}$ para diferentes valores de n, μ e α de uma LBUA.

A Figura 3.8, se trata do REQM de $\hat{\alpha}$ da LBUA para diferentes valores de μ, α e amostral. Assim como na LBZA, o REQM de $\hat{\alpha}$ diminui com o aumentar do tamanho amostral. Dos valores testados, $\alpha = 0.9$ foi o valor que obteve o maior REQM ao fixar o valor de μ.

Em resumo, podemos concluir que, o método máxima verossimilhança é apropriado para estimar tanto o parâmetro μ, quanto o parâmetro α da distribuição log-Bibal tanto com zeros, quanto com uns ajustados. Dado que, para os casos simulados, é notório a redução que o viés e REQM sofrem a medida que o tamanho da amostra aumenta, como também, aumentam a medida que a proporção de zeros ou uns crescem na amostra. Também vale ressaltar que viés e REQM de $\hat{\mu}$ e $\hat{\alpha}$ não se apresentam muito alto para tamanhos de amostras pequenos, como também são bem próximos a zero, quando o tamanho da amostra é suficientemente grande.

Capítulo 4

Aplicação

Neste Capítulo, será apresentada uma ilustração com um conjunto de dados reais. Este conjunto de dados é referente uma nova abordagem para conceituar e medir a democracia. O projeto *Varieties of Democracy* (V-Dem) distingue entre cinco princípios de democracia de alto nível: eleitoral, liberal, participativo, deliberativo e igualitário, e coleta dados para medir esses princípios. O projeto é composto por uma equipe de mais de 50 cientistas sociais em seis continentes. O trabalho é desenvolvido por mais de 3.000 especialistas em diversos países e um Conselho Consultivo Internacional verdadeiramente global (Maerz et al. 2021). Para maiores detalhes, ver https://www.v-dem.net/.

São considerados os dados do ano de 2019 e foi usada como base uma análise realizada em https://www.andrewheiss.com/blog/2021/11/08/beta-regression-guide/, feita pelo professor Andrew Heiss do Departamento de Gestão e Política Pública da *Georgia State University*. No entanto, o principal interesse estará em ajustar a variável **proporção de mulheres no parlamento**. O conjunto de dados é composto por 27.013 observações e 4.108 variáveis, foi acessado através do pacote `R` chamado de `vdemdata` desenvolvido por Maerz et al. (2021). Para usar a variável de interesse é necessária a transformação da variável `v2lgfemleg` que representa a porcentagem de mulheres parlamentares. Portanto, será crianda uma nova variável a qual chamaremos de `prop_mulher` é obtida dividindo por 100 a variável `v2lgfemleg`. Também iremos considerar a variável `v2lgqugen` que pode assumir os valores 0, 1, 2, 3 e 4. No caso do valor ser maior do que 0, isso indicar que o país tem cota baseada em gênero. Desta

forma também será criada a variável `cotas` que é igual a ”SIM”se `v2lgqugen` é maior do que 0 e ”NÃO”caso contrário. Além disso, foram eliminadas as linhas com dados faltantes. Abaixo, é apresentado esse tratamento dos dados.

Listagem 4.1: Tratamento dos dados.

```
1 library(tidyverse)
2 library(vdemdata)
3 vdem_novo <- vdem %>%  select(year, v2lgfemleg, v2lgqugen) %>%
     filter(year == 2019) %>%
4 drop_na(v2lgqugen, v2lgfemleg) %>% mutate(cotas =
     if_else(v2lgqugen > 0, "SIM", "NÃO"),
5 prop_mulher = v2lgfemleg / 100)
```

A Figura 4.1 apresenta a distribuição da proporção de mulheres no parlamento entre os dois valores diferentes de cotas. Note que aparentemente os países que não possuem uma cota baseada em gênero apresentam menos mulheres parlamentares, o que é esperado visto que as cotas foram projetadas principalmente para aumentar a quantidade de mulheres parlamentares.

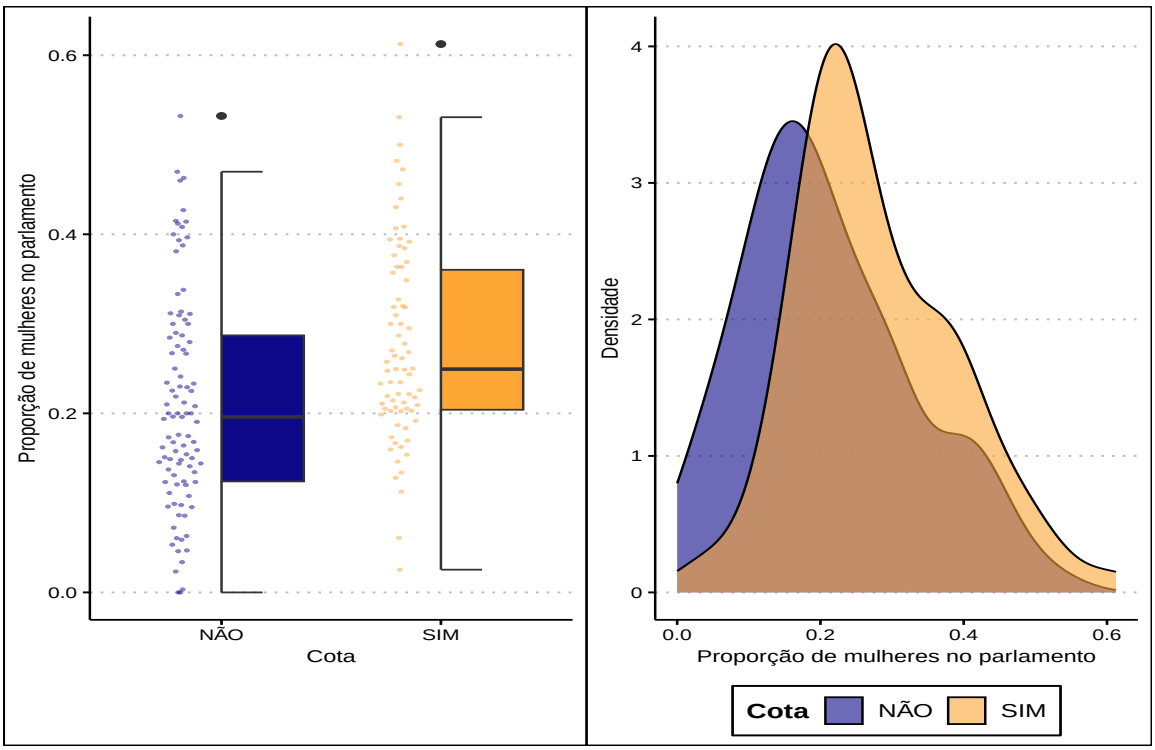


Figura 4.1: Distribuição da proporção de mulhere parlamentares.

Através da Tabela 4.1 percebe-se que em paises que não possuem cotas a proporção

de mulheres no parlamento varia no intervalo $[0, 0.532]$, enquanto que para os países com cotas essa proporção varia entre $[0.0254, 0.612]$. A análise poderia ser conduzida considerando ambos os grupos de cotas, porém, iremos considerar apenas os países que possuem algum tipo de cota baseado em gênero para ajustar a distribuição log-Bilal com zeros ajustados.

Tabela 4.1: Distribuição da proporção de mulheres parlamentares na presença (ou não) de cotas.

Cota	n	mínimo	máximo
SIM	97	0	0.532
NÃO	75	0.0254	0.612

Os resultados obtidos usando o `R`, apresentados abaixo, para a distribuição log-Bilal com zeros ajustados, temos que as estimativas de máxima verossimilhança de μ e α são respectivamente, $0,2859$ e $0,0206$. A função `gamlssInf0to1()` do pacote `R` chamado de `gamlss.inf` considera apenas a ligação logit para modelar a proporção de zeros. Desta forma, utilizamos também a função `inv.logit()` do pacote `boot` para obter a proporção estimada.

Listagem 4.2: Ajuste dos dados.

```
#########
#Distribuicao log-Bilal com zeros ajustados
########
library(gamlss.inf)
library(boot)
gen.Inf0to1("LB","Zero")
modelo_logBilal <- gamlssInf0to1(y = prop_mulher, mu.formula = ~
   1, xi0.formula =~1,
+                                  family = LB(mu.link =
   "identity"), data = vem_sem_cota)
> summary(modelo_logBilal)
******************************************************************
Family:  "InfLB"

Call:
gamlssInf0to1(y = prop_mulher, mu.formula = ~1, xi0.formula = ~1,
    data = vem_sem_cota, family = LB(mu.link = "identity"))

```

```
17
18 Fitting method: RS()
19
20 ------------------------------------------------------------------
21 Mu link function:  identity
22 Mu Coefficients:
23             Estimate Std. Error t value Pr(>|t|)
24 (Intercept)  0.28591    0.01952   14.64   <2e-16 ***
25 ---
26 Signif. codes:
27 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1
28
29 ------------------------------------------------------------------
30 xi0 link function:  logit
31 xi0 Coefficients:
32             Estimate Std. Error t value Pr(>|t|)
33 (Intercept)  -3.8607     0.7145  -5.403 4.84e-07 ***
34 ---
35 Signif. codes:
36 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1
```

Capítulo 5

Conclusões

A distribuição log-Bilal é uma distribuição indexada por apenas um parâmetro e pode ser usada para ajustar dados no intervalo $(0, 1)$. Apesar da sua simplicidade, é uma distribuição flexível. Por exemplo, esta distribuição pode se usada para modelar dados com assimetria negativa. Neste trabalho, foi proposta a distribuição log-Bilal com zeros e/ou uns ajustados. A partir das simulações, foi possível detectar que o estimador do parâmetro possui boas propriedades, dado que, ao usarmos amostras suficientemente grandes, o estimador tende a ser pouco viesado. Por fim, foi utilizado um conjunto de dados que tenta conceituar e medir a democracia para apresentar uma ilustração de como podemos fazer o ajuste de dados usando a distribuição proposta neste trabalho.

Como trabalhos futuros, é possível propor um modelo de regressão baseado na distribuição proposta. Além disso, é possível usar uma abordagem bayesiana para realizar a estimação dos parâmetros. Também pode-se propor modelos para ajustar dados com a presença de censura.

Apêndice A

Códigos R

Listagem A.1: Códigos R - modelo de regressão quantílica.

```
1  library(gamlss.inf)
2  library(boot)
3
4  #Funcoes
5
6  dLB <-function(x, mu = 0.5, log = FALSE){
7    if (any(mu <= 0) | any(mu >= 1) )  stop(paste("mu must be
       between 0 and 1", "\n", ""))
8    if (any(x <= 0) | any(x >= 1))  stop(paste("x must be between 0
       and 1", "\n", ""))
9    theta <- 0.5*(sqrt((24+mu)/mu) - 5)
10
11   log_pdf <- log(6) - log(theta) + (2/theta - 1)*log(x) + log(1 -
       x^(1/theta))
12
13   if(log == FALSE) fy <- exp(log_pdf) else fy <- log_pdf
14
15   fy
16 } #ok!
17
18 pLB <- function(q, mu = 0.5, lower.tail = TRUE, log.p = FALSE)
19 {
20   if (any(mu <= 0) | any(mu >= 1) )  stop(paste("mu must be
       between 0 and 1", "\n", ""))
```

```
21   if (any(q <= 0) | any(q >= 1))  stop(paste("y must be between 0 and 1", "\n", ""))
22
23   theta <- 0.5*(sqrt((24+mu)/mu) - 5)
24
25   if(lower.tail == TRUE) cdf <- 3*q^(2/theta) - 2*q^(3/theta) else cdf <- 1 - (3*q^(2/theta) - 2*q^(3/theta))
26
27   if(log.p == FALSE) return(cdf) else return(log(cdf))
28 } #ok
29
30
31 qLB <- function(p, mu = 0.5,  lower.tail = TRUE, log.p = FALSE){
32   if (any(mu <= 0) | any(mu >= 1) )  stop(paste("mu must be between 0 and 1", "\n", ""))
33   if (any(p <= 0) | any(p >= 1))  stop(paste("p must be between 0 and 1", "\n", ""))
34
35   theta <- 0.5*(sqrt((24+mu)/mu) - 5)
36   x0 <- NULL
37
38   for(i in 1:length(p)){
39     if(lower.tail == TRUE){
40       x0[i] <- uniroot(function(x) 3*x^(2/theta) - 2*x^(3/theta) - p[i], lower = 0, upper = 1, tol = 1e-9 )$root
41     } else x0[i] <- uniroot(function(x) 3*x^(2/theta) - 2*x^(3/theta) - (1-p[i]), lower = 0, upper = 1, tol = 1e-9 )$root
42   }
43
44   if(log.p == FALSE) return(x0) else return(log(x0))
45
46 } #ok
47
48
49 rLB <- function(n, mu = 0.5){
50   if (any(mu <= 0) | any(mu >= 1) )  stop(paste("mu must be between 0 and 1", "\n", ""))
51   if (any(n <= 0))  stop(paste("n must be a positive integer", "\n", ""))
52
53   r <- NULL
```

```
54    theta <- 0.5*(sqrt((24+mu)/mu) - 5)
55
56    for (i in 1:n) {
57      u <- runif(1)
58      r[i] <- uniroot(function(x) 3*x^(2/theta) - 2*x^(3/theta) - u,
          lower = 0, upper = 1, tol = 1e-9 )$root
59    }
60
61    r
62  } #ok
63
64
65  LB <- function (mu.link ="logit")
66  {
67    mstats <- checklink("mu.link", "LogBilal", substitute(mu.link),
        c("logit", "probit", "cloglog", "cauchit", "log", "own"))
68    structure(
69      list(family = c("LB","LogBilal"),
70           parameters = list(mu = TRUE),
71           nopar = 1,
72           type = "Continuous",
73           mu.link = as.character(substitute(mu.link)),
74           mu.linkfun = mstats$linkfun,
75           mu.linkinv = mstats$linkinv,
76           mu.dr = mstats$mu.eta,
77           dldm = function(y,mu){
78             nd = gamlss:::numeric.deriv(dLB(y, mu, log = TRUE),
                 "mu", delta = 1e-04)
79             dldm = as.vector(attr(nd, "gradient"))
80             dldm
81           }, #ok!
82           d2ldm2 = function(y, mu){
83             nd = gamlss:::numeric.deriv(dLB(y, mu, log = TRUE),
                 "mu", delta = 1e-04)
84             dldm = as.vector(attr(nd, "gradient"))
85             d2ldm2 = -dldm * dldm
86             d2ldm2
87           }, #ok!
88           G.dev.incr  = function(y,mu,...)  -2*dLB(x = y, mu = mu,
               log = TRUE) ,
89           rqres = expression(rqres(pfun = "pLB", type =
               "Continuous", y = y, mu = mu)),
```

```
90          mu.initial = expression({mu <- (y+mean(y))/2}), #ok!
91          mu.valid = function(mu) all(mu > 0 & mu < 1) ,  #ok!
92          y.valid = function(y)  all(y > 0 & y < 1), #ok!
93          mean = function(mu) mu, #ok!
94          variance = function(mu) mu^2
95      ),
96      class = c("gamlss.family","family"))
97  }
```

Referências Bibliográficas

Abd-Elrahman, A. M. (2013), 'Utilizing ordered statistics in lifetime distributions production: a new lifetime distribution and applications', *Journal of Probability and Statistical Science* **11**(2), 153–164.

Altun, E. (2021), 'The log-weighted exponential regression model: alternative to the beta regression model', *Communications in Statistics-Theory and Methods* **50**(10), 2306–2321.

Altun, E., El-Morshedy, M. & Eliwa, M. (2021), 'A new regression model for bounded response variable: An alternative to the beta and unit-lindley regression models', *Plos One* **16**(1), e0245627.

Altun, E. & Hamedani, G. (2018), 'The log-xgamma distribution with inference and application', *Journal de la Société Française de Statistique* **159**(3), 40–55.

Cribari-Neto, F. & Santos, J. (2019), 'Inflated kumaraswamy distributions', *Anais da Academia Brasileira de Ciências* **91**.

Enea, M., Stasinopoulos, M., Rigby, B. & Hossain, A. (2019), *gamlss.inf: Fitting Mixed (Inflated and Adjusted) Distributions*.
R package version 1.0-1.
URL: *https://CRAN.R-project.org/package=gamlss.inf*

Ferrari, S. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**(7), 799–815.

Joshi, M. & Pustejovsky, J. (2022), *simhelpers: Helper Functions for Simulation Studies*.

R package version 0.1.2.
URL: *https://CRAN.R-project.org/package=simhelpers*

Kumaraswamy, P. (1980), 'A generalized probability density function for double-bounded random processes', *Journal of Hydrology* **46**(1-2), 79–88.

Leschinski, C. H. (2019), *MonteCarlo: Automatic Parallelized Monte Carlo Simulations*. R package version 1.0.6.
URL: *https://CRAN.R-project.org/package=MonteCarlo*

Maerz, S., Edgell, A., Hellmeier, S. & Ilchenko, N. (2021), 'Vdemdata - an R package to load, explore and work with the most recent V-Dem (Varieties of Democracy) and V-Party datasets'.
URL: *https://www.v-dem.net/en/*

Mazucheli, J., Menezes, A. F. & Dey, S. (2018), 'The unit-birnbaum-saunders distribution with applications', *Chilean Journal of Statistics* **9**(1), 47–57.

Ospina, R. M. (2008), Modelos de regressão beta inflacionados, PhD thesis, Universidade de São Paulo.

Pace, L. & Salvan, A. (1997), *Principles of statistical inference: from a Neo-Fisherian perspective*, Vol. 4, World Scientific.

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape,(with discussion)', *Applied Statistics* **54**, 507–554.

Topp, C. W. & Leone, F. C. (1955), 'A family of j-shaped frequency functions', *Journal of the American Statistical Association* **50**(269), 209–219.