Família Kumaraswamy-G para analisar dados de sobrevivência de longa duração

Amanda Morales Eudes

Família Kumaraswamy-G para analisar dados de sobrevivência de longa duração

Amanda Morales Eudes

Orientadora: **Profa. Dra. Vera Lucia Damasceno Tomazella**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

USP/UFSCar – São Carlos
Abril de 2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

E86fk

Eudes, Amanda Morales.
Família Kumaraswamy-G para analisar dados de sobrevivência de longa duração / Amanda Morales Eudes. -- São Carlos : UFSCar, 2015.
59 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2015.

1. Análise de sobrevivência. 2. Kumaraswamy generalizada. 3. Abordagem bayesiana. 4. Fração de cura. I. Título.

CDD: 519.9 (20ª)

AMANDA MORALES EUDES

Dissertação apresentada ao Programa de Pós-Graduação em Estatística – UFSCar/USP, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 25 de fevereiro de 2015.

COMISSÃO JULGADORA:

Profa. Dra. Vera Lucia Damasceno Tomazella
UFSCar
Presidente

Prof. Dr. Antonio Carlos Pedroso de Lima
IME-USP
Membro

Prof. Dr. Victor Hugo Lachos Dávila
UNICAMP
Membro

Aos meus pais, **Marcia Morales** e **Antonio de Lucena Eudes** e meu noivo **Thiago Lima D'Andrea**, pelo apoio, incentivo e carinho incondicionais.

Agradecimentos

Agradeço primeiramente à Deus, por ter a certeza de que Ele esteve presente em todos os momentos dessa jornada, e me deu força para continuar até nos momentos mais difíceis da vida. Por ter me iluminado nas decisões mais difíceis e por ter me guiado ao longo do curso.

À FAPESP e CAPES pelo apoio financeiro.

Aos meus pais, Antônio de Lucena Eudes e Márcia Morales, pelo amor e dedicação e por terem me proporcionado essa oportunidade de um futuro promissor.

Ao meu noivo, Thiago Lima D'Andrea, pelo amor, incentivo, paciência e compreensão sempre, me deixando mais tranquila nos momentos mais difíceis do mestrado e me dando apoio nas minhas decisões.

À minha orientadora, Vera Lucia Damasceno Tomazella, pela sabedoria na orientação e por sua compreensão e auxílio para alcançarmos mais essa vitória.

Às minhas amigas, por tudo que pudemos compartilhar: a convivência, as alegrias, as frustrações, as descobertas e o que aprendemos. Jamais lhes esquecerei!

Não poderia deixar de agradecer aos professores, funcionários e colegas da UFSCar e USP.

Enfim, à todos que de alguma forma me ajudaram na concretização deste sonho.

Resumo

Em análise de sobrevivência estuda-se o tempo até a ocorrência de um determinado evento de interesse e na literatura, uma abordagem muito utilizada é a paramétrica, em que os dados seguem uma distribuição de probabilidade. Diversas distribuições conhecidas são utilizadas para acomodar dados de tempos de falha, porém, grande parte destas distribuições não é capaz de acomodar funções de risco não monótonas. Kumaraswamy (1980) propôs uma nova distribuição de probabilidade e, baseada nela, mais recentemente Cordeiro e de Castro (2011) propuseram uma nova família de distribuições generalizadas, a Kumaraswamy generalizada (Kum-G). Esta distribuição, além de ser flexível, contém distribuições com funções de risco unimodal e em forma de banheira. O objetivo deste trabalho é apresentar a família de distribuições Kum-G e seus casos particulares para analisar dados de tempo de vida de indivíduos em risco, considerando que uma parcela da população nunca apresentará o evento de interesse, além de considerarmos que covariáveis influenciem na função de sobrevivência e na proporção de curados da população. Algumas propriedades destes modelos serão abordadas, bem como métodos adequados de estimação, tanto na abordagem clássica quanto na bayesiana. Por fim, são apresentadas aplicações de tais modelos a conjuntos de dados existentes na literatura.

Palavras-chave: Análise de Sobrevivência, Kumaraswamy generalizada, Abordagem Bayesiana, Modelo de Fração de Cura, Covariáveis.

Abstract

In survival analysis is studied the time until the occurrence of a particular event of interest and in the literature, the most common approach is parametric, where the data follow a specific probability distribution. Various known distributions maybe used to accommodate failure time data, however, most of these distributions are not able to accommodate non-monotonous hazard functions. Kumaraswamy (1980) proposed a new probability distribution and, based on that, recently Cordeiro and de Castro (2011) proposed a new family of generalized distributions, the so-called Kumaraswamy generalized (Kum-G). In addition to its flexibility, this distribution may also be considered for unimodal and tub shaped hazard functions. The objective of this dissertation is to present the family of Kum-G distributions and their particular cases to analyze lifetime data of individuals at risk, considering that part of the population will never present the event of interest, and considering that covariates may influence the survival function and the cured proportion of the population. Some properties of these models will be discussed as well as appropriate estimation methods, in the classical and Bayesian approaches. Finally, applications of such models are presented to literature data sets.

Keywords: Survival analysis, Kumaraswamy generalized, Bayesian approach, long term model, covariates.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo

1

Introdução

Em análise de sobrevivência o estudo é feito a partir de dados relacionados ao tempo até a ocorrência de um determinado evento de interesse, também conhecido como tempo de falha. Estes dados podem ser provenientes do tempo de duração de um componente eletrônico até que haja falha; do tempo até que um paciente tenha a ocorrência de uma determinada doença; do tempo até um determinado medicamento surtir o efeito desejado, entre outros. O comportamento de tais dados pode ser verificado de maneira empírica, esta abordagem é dita não paramétrica, ou pode-se considerar que os dados seguem uma distribuição de probabilidade, e esta abordagem é dita paramétrica, a mais utilizada neste trabalho.

As funções de sobrevivência e de risco, objetos de maior interesse em análise de sobrevivência, permitem o estudo de tal comportamento. A função de sobrevivência é a probabilidade de um indivíduo ou componente sobreviver após um tempo pré-estabelecido e a função de risco é a taxa de falha instantânea, que pode assumir graficamente diversas formas, tais como constante, crescente, decrescente, unimodal ou em forma de banheira. Porém, quando o comportamento da função de risco é não monótono, grande parte das distribuições usualmente conhecidas, como exponencial e Weibull, não é capaz de acomodar este tipo de comportamento.

Isto ocorre porque uma desvantagem destes modelos é que eles são muito restritos devido à pequena quantidade de parâmetros e, portanto, as conclusões dos modelos podem não ser suficientemente robustas para desvios em relação aos dados. Existem algumas distribuições que acomodam função de risco não monótona, porém elas são geralmente muito complicadas e com muitos parâmetros.

Podemos modelar dados reais de sobrevivência utilizando praticamente qualquer distribuição contínua e de valores positivos, os modelos mais simples e comuns, como a distribuição Exponencial e Weibull, podem não ser apropriados. Assim, encontrar uma

distribuição que acomode funções de risco não monótonas é um conhecido problema em análise de sobrevivência. Portanto, é desejável considerar outras abordagens para alcançar uma maior flexibilidade, e isto é o que vem motivando os estudos para encontrar distribuições que acomodem tais tipos de função.

Kumaraswamy (1980) propôs a distribuição Kumaraswamy, que foi muito utilizada em hidrologia e, baseada nela, Cordeiro e de Castro (2011) propuseram uma nova família de distribuições generalizadas, a Kumaraswamy generalizada (Kum-G). Ela é flexível e contém distribuições com funções de risco unimodal e em forma de banheira, como mostrado por de Pascoa *et al.* (2011), além de ter como casos especiais qualquer distribuição, como a normal, exponencial, Weibull, gama, Gumbel e gaussiana inversa. O domínio da distribuição será o intervalo em que os casos particulares estão definidos.

Em uma população podem existir indivíduos que não experimentaram o evento de interesse até o final do estudo, o que é denominado por censura. Quando houver um grande número de indivíduos censurados, temos um indício de que nesta população existem indivíduos que não estão sujeitos a experimentar o evento, e então eles são considerados imunes, curados ou não suscetíveis ao evento de interesse.

A partir dos modelos tradicionais de sobrevivência não é possível estimar a fração de cura da população, ou seja, a proporção de indivíduos que são considerados curados. Assim, são necessários modelos estatísticos que incorporem tal fração e estes são denominados modelos de longa duração ou modelos de fração de cura. Devido a tal capacidade, diferentes métodos de ajuste têm sido propostos em diversas áreas como em estudos biomédicos, financeiros, criminologia, demografia e confiabilidade industrial, entre outros. Por exemplo, em dados biomédicos um evento de interesse pode ser a morte do paciente devido à recorrência do tumor, mas podem ter pacientes que se curam e não morrem devido ao câncer. Quando se trabalha com dados financeiros, um evento de interesse pode ser o desligamento do cliente de um banco devido à inadimplência, mas podem existir clientes que nunca se tornarão inadimplentes. Já em dados de criminologia, o evento de interesse pode ser a reincidência no crime e pode haver pessoas que não cometam crime novamente. Em confiabilidade industrial, os modelos de longa duração são utilizados para verificar a proporção de componentes que falham, ao serem colocados em teste no tempo zero e expostos a vários regimes de tensão ou uso. Na área de pesquisa de mercado, os imunes são considerados os indivíduos que nunca comprarão um determinado produto. Vide, por exemplo, Anscombe (1961), Farewell (1977), Goldman (1984), Broadhurst e Maller (1991), Meeker e Escobar (1998).

Muitos autores contribuíram para a teoria dos modelos de longa duração, sendo Boag (1949) o pioneiro, em que o método de máxima verossimilhança foi utilizado para estimar a proporção de sobreviventes em uma população de 121 mulheres com câncer de mama, experimento esse que teve a duração de 14 anos. Baseado na ideia de Boag, Berkson e Gage (1952) propuseram um modelo de mistura com o objetivo de estimar a proporção de curados numa população submetida a um tratamento de câncer de estômago. Modelos mais complexos de longa duração, tais como Yakovlev e Tsodikov (1996), Chen *et al.* (1999) entre outros, surgiram com o objetivo de

explicar melhor os efeitos biológicos envolvidos. Mais recentemente, Rodrigues *et al.* (2009) propuseram uma teoria unificada de longa duração, considerando diferentes causas competitivas.

Um ponto bastante importante em análise de sobrevivência é o estudo de covariáveis, pois diversos fatores podem influenciar o tempo de sobrevivência de um indivíduo. Assim, incorporar covariáveis nos permite ter um modelo muito mais completo e repleto de informações valiosas. Por exemplo, se estamos interessados em estudar o tempo de vida de pacientes com uma determinada doença que estão recebendo um certo tratamento, outros fatores podem influenciar na cura do paciente, assim pode-se encontrar novos meios de tratar a doença a partir de covariáveis.

Neste trabalho será estudada a família de distribuições Kum-G, cuja proposta principal é investigar sua aplicação aos tempos de vida dos indivíduos em risco, bem como suas propriedades e um método adequado de estimação para os parâmetros do modelo, com enfoque nos casos particulares, como a distribuição Kumaraswamy exponencial. O mesmo será feito para modelos de longa duração, considerando o modelo unificado de fração de cura para tal família de distribuições. Há a possibilidade de que algumas covariáveis influenciem no tempo de sobrevivência então, consideramos um modelo de regressão para incorporar tais covariáveis.

Na seção 1.1 temos uma breve introdução à análise de sobrevivência e na seção 1.2 temos a descrição da distribuição base do nosso modelo, a distribuição Kumaraswamy.

1.1 Introdução à Análise de Sobrevivência

Análise de sobrevivência estuda o tempo de vida, que pode ser, por exemplo, desde o tempo de duração de um componente eletrônico até o tempo de vida de pacientes com graves doenças. Algumas áreas em que se usa análise de sobrevivência são: medicina, biologia, engenharia, estatística, economia, criminalística, entre outras.

A principal variável analisada é o tempo até a ocorrência de um determinado evento de interesse (tempo de falha). Em grande parte dos estudos, o tempo de falha é o tempo até o reaparecimento de uma doença ou até a morte de um paciente. Para definir o tempo de falha é necessário estabelecer o "instante inicial" a partir do qual os tempos são medidos.

Uma característica específica é um tipo de observação incompleta, a chamada censura. Com a presença de censuras, torna-se impossível a aplicação de técnicas estatísticas convencionais. Portanto, foram desenvolvidos métodos para analisar dados deste tipo.

A seguir definiremos as censuras mais comuns em estudos de análise de sobrevivência.

• Censura tipo I

Este tipo de censura ocorre quando o tempo para o fim do estudo é pré-estabelecido; assim, alguns indivíduos deixam de experimentar o evento de interesse ao fim deste estudo, tendo os seus tempos de vida censurados. Um exemplo para esse tipo de censura é quando um determinado banco deseja verificar o tempo até que os clientes, de determinada carteira, se tor-

nam inadimplentes. Estuda-se, portanto, esta carteira durante um tempo pré-determinado pela instituição e ao fim, alguns desses deixaram de experimentar o evento de interesse (portanto não são inadimplentes), observando assim, a censura do tipo I.

• Censura tipo II

Quando o estudo é terminado após um número pré determinado, r, de indivíduos experimentar o evento de interesse, ou seja, após um número r de ocorrências a pesquisa é finalizada e os indivíduos que deixaram de experimentar o evento de interesse terão seus tempos censurados. Este tipo de estudo economiza tempo e dinheiro, já que pode haver indivíduos que levam muito tempo até falhar.

• Censura aleatória

Diferentemente das outras, este tipo de censura foge ao controle do pesquisador. Geralmente ocorre quando o indivíduo abandona determinado experimento sem ter experimentado o evento de interesse, por exemplo, se o paciente morrer por uma razão diferente da estudada. A censura aleatória é o caso mais comum.

Existem outros tipos de censuras (ver por exemplo Colosimo e Giolo (2006) e Lawless (1982)).

Neste estudo, a representação dos dados e possíveis censuras será:

Cada indivíduo i, $i = 1, \ldots, n$, é representado por (t_i,δ_i), t_i é o tempo de falha ou de censura e δ_i é uma variável indicadora que acusa se há presença de falha ou de censura, ou seja,

$$\delta_i = \begin{cases} 1 \text{ se } t_i \text{ for tempo até a falha,} \\ 0 \text{ se } t_i \text{ for tempo até a censura.} \end{cases}$$

Suponhamos que a variável aleatória T, $T \geq 0$, tenha função de densidade de probabilidade denotada por $f(t)$. Podemos descrever a função de densidade de probabilidade como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t + \Delta t)$ por unidade de tempo e expressar como sendo

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t}. \tag{1.1}$$

Sua função de distribuição acumulada é definida por

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{1.2}$$

A estimação da probabilidade de um indivíduo sobreviver após um tempo t é um dos principais interesses na análise de sobrevivência. Para estimar esta probabilidade definimos a função

de sobrevivência, que é dada por

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t). \tag{1.3}$$

Propriedades de $S(t)$:

1. é não crescente;
2. $S(0) = 1$;
3. $\lim_{t\to\infty} S(t) = 0$.

A função de risco, também chamada de taxa de falha, fornece a taxa de falha instantânea, ou seja, sabendo que um indivíduo sobreviveu até o tempo t, tem-se o risco deste indivíduo falhar no intervalo de tempo $[t, t + \Delta t)$ com $\Delta t \to 0$. É definida por

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{1.4}$$

Graficamente a função de risco pode apresentar comportamento constante, crescente, decrescente e até mesmo formas não monótonas como a "curva em forma de banheira", em que esta pode representar a função de risco do tempo até a morte de um ser humano.

A função de risco acumulada é definida por

$$H(t) = \int_0^t h(u)du. \tag{1.5}$$

Algumas relações importantes entre as funções definidas anteriormente, e que são amplamente utilizadas na prática, são

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log\left(S(t)\right), \tag{1.6}$$

$$H(t) = -\log\left(S(t)\right) \tag{1.7}$$

e

$$S(t) = \exp\left(-H(t)\right). \tag{1.8}$$

1.1.1 Estimação por Máxima Verossimilhança

O foco principal deste trabalho é ajustar modelos paramétricos por geralmente apresentarem uma interpretação natural (possibilidade de prever o risco e estimar a probabilidade de cura).

Baseando-se em resultados obtidos na amostra, o estimador de máxima verossimilhança escolhe o "melhor" conjunto de parâmetros da distribuição dos dados, ou seja, consegue encontrar parâmetros os quais levam a uma distribuição que se encaixe bem aos dados.

O método da máxima verossimilhança é capaz de incorporar censuras e possui ótimas propriedades para amostras grandes. Este é um método muito utilizado em análise de sobrevivência.

Para um conjunto de dados com n observações de tempo $(t_1, t_2, \ldots, t_n)$ sem censura, a função de verossimilhança é definida por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta}), \tag{1.9}$$

em que t_i, $i = 1, 2, \ldots, n$, é uma observação da variável aleatória T que representa o tempo de vida com função densidade de probabilidade $f(t; \boldsymbol{\theta})$ e $\boldsymbol{\theta}$ é o vetor de parâmetros.

Como dados censurados nos trazem informações importantes (ou seja, quando temos uma censura, sabemos que o tempo de falha do indivíduo é maior do que aquele em que foi censurado), não podemos deixá-los de lado. Portanto, a sua contribuição para $L(\boldsymbol{\theta})$ é dada pela função de sobrevivência $S(t)$.

Temos que a função de verossimilhança com dados censurados é

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i; \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i; \boldsymbol{\theta})\right]^{1-\delta_i} = \prod_{i=1}^{n} \left[h(t_i; \boldsymbol{\theta})\right]^{\delta_i} S(t_i; \boldsymbol{\theta}), \tag{1.10}$$

em que δ_i é a variável indicadora de falha. A expressão acima é válida para as censuras independentes do tipo I, II, aleatória e também sob a suposição que o mecanismo de censura é não informativa.

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$, ou de forma equivalente, $\ell(\boldsymbol{\theta}) = \log\left[L(\boldsymbol{\theta})\right]$. Os estimadores podem ser encontrados resolvendo-se o sistema de equações

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{1.11}$$

Normalmente, o estimador de máxima verossimilhança não possui uma expressão fechada por sua complexidade. Então, faz-se necessário o uso de métodos numéricos para fazer a estimação.

1.1.2 Técnicas Não Paramétricas

Na literatura de análise de sobrevivência são encontrados alguns estimadores da função de sobrevivência obtidos por técnicas não paramétricas. Podemos citar, como exemplo, o estimador de Kaplan-Meier proposto por Kaplan e Meier (1958) e o estimador de Nelson-Aalen proposto por Nelson (1972) e retomado por Aalen (1978). O primeiro é um dos mais utilizados e por isso o apresentamos a seguir.

1.1.2.1 Estimador de Kaplan-Meier

Para estimar a função de sobrevivência de um conjunto de dados com a presença de censuras, podemos utilizar o estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958). Este estimador não paramétrico é o mais utilizado em estudos clínicos, já que possui boas propriedades. Ele também é chamado de estimador limite-produto. Sejam

- $t_{(1)} < t_{(2)} < \ldots < t_{(k)}, j = 1, \ldots, k$, os k tempos observados distintos e ordenados de falha,
- d_j o número de falhas em $t_{(j)}, j = 1, \ldots, k$,
- n_j o número de indivíduos em risco em $t_{(j)}$, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a $t_{(j)}$.

Desta maneira, o estimador de Kaplan-Meier (KM) é dado por

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j} = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right). \tag{1.12}$$

Esta expressão é uma função escada com degraus nos tempos observados de falha $t_{(j)}$.

1.1.3 Distribuição Weibull Modificada

Nesta subseção apresentamos um modelo utilizado em análise de sobrevivência para modelar, de forma plausível, dados relacionados a tempos de sobrevivência e que tem como casos particulares algumas das distribuições mais utilizadas nesta área.

Uma variável aleatória contínua e não negativa T tem distribuição Weibull modificada com parâmetros $\alpha > 0$, $c \geq 0$ e $\beta \geq 0$, se sua função densidade de probabilidade é dada por

$$f(t) = \alpha t^{c-1}(c + \beta t)\mathrm{e}^{\beta t - \alpha t^c \mathrm{e}^{\beta t}}, \quad t \geq 0, \tag{1.13}$$

em que α é o parâmetro de escala, c o parâmetro de forma e β é um fator de aceleração na sobrevida do indivíduo. Esta distribuição foi proposta por Lai *et al.* (2003).

Se T é uma variável aleatória com função densidade de probabilidade dada em (1.13), escrevemos $T \sim \mathrm{WM}(\alpha, \beta, c)$.

As funções correspondentes de distribuição acumulada, de sobrevivência e de risco são, respectivamente,

$$F(t) = 1 - \mathrm{e}^{-\alpha t^c \mathrm{e}^{\beta t}}, \tag{1.14}$$

$$S(t) = 1 - F(t) = \mathrm{e}^{-\alpha t^c \mathrm{e}^{\beta t}}, \tag{1.15}$$

e

$$h(t) = \alpha t^{c-1}(c + \beta t)\mathrm{e}^{\beta t}. \tag{1.16}$$

Uma das vantagens desta distribuição está na capacidade de acomodar funções de risco em várias formas, dependendo somente dos parâmetros c e β, com as seguintes propriedades:

- Se $c > 1$, $h(t)$ é crescente em t;
- Se $0 < c < 1$, $h(t)$ tem forma de banheira;
- Quando $\beta = 0$ e $c = 1$, $h(t)$ tem risco constante;
- Para $c = 0$ e $0 < c < 1$, $h(t)$ é decrescente em t.

Para simular valores da distribuição WM, basta resolver a equação não-linear

$$t^c \mathrm{e}^{\beta t} - \frac{1}{\alpha} \log(1-u) = 0,$$

em que u segue distribuição $\mathrm{U}(0,1)$.

A distribuição WM apresenta como casos particulares algumas distribuições conhecidas na literatura, conforme mostra a Tabela 1.1.

Tabela 1.1: Casos particulares da distribuição Weibull modificada.

Distribuição	Parametrização	Função de distribuição acumulada
Exponencial	$\beta = 0$ e $c = 1$	$F(t) = 1 - \mathrm{e}^{-\alpha t}$
Rayleigh	$\beta = 0$ e $c = 2$	$F(t) = 1 - \mathrm{e}^{-\alpha t^2}$
Valor-extremo	$c = 0$	$F(t) = 1 - \mathrm{e}^{-\alpha \mathrm{e}^{\beta t}}$
Weibull	$\beta = 0$	$F(t) = 1 - \mathrm{e}^{-\alpha t^c}$

Na Figura 1.1, apresentamos as curvas de densidade, de sobrevivência e de risco, respectivamente, considerando $\alpha = 1$ e diferentes escolhas de β e c.

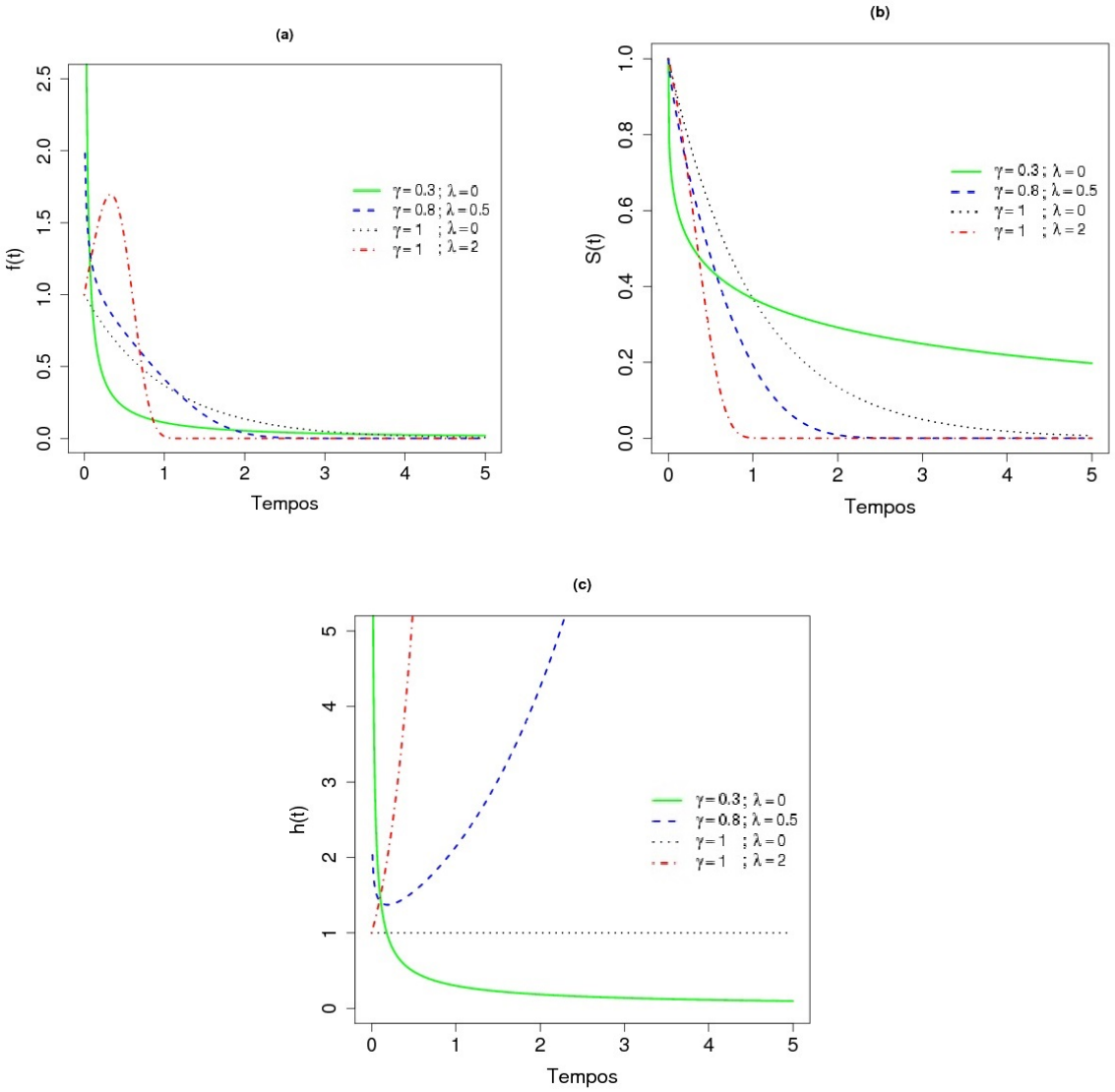

Figura 1.1: Curvas das funções de densidade (a), de sobrevivência (b) e de risco (c).

1.2 Distribuição Kumaraswamy

Em Kumaraswamy (1980) a distribuição Kumaraswamy foi proposta, que modela processos aleatórios aplicados à hidrologia. Seja T a variável aleatória com distribuição Kumaraswamy, cuja notação é $T \sim \mathrm{Kum}(\lambda, \varphi)$, está definida no intervalo $(0, 1)$, com função densidade de probabilidade (fdp) e função de distribuição acumulada (fda) dadas por

$$f(t) = \lambda\varphi t^{\lambda-1}(1 - t^{\lambda})^{\varphi-1}, \tag{1.17}$$

e

$$F(t) = 1 - (1 - t^{\lambda})^{\varphi}, \tag{1.18}$$

sendo $\lambda > 0$ e $\varphi > 0$ os parâmetros de forma da distribuição.

A distribuição Kumaraswamy é relacionada com a distribuição beta, pois $T \sim \mathrm{Kum}(\lambda, \varphi)$ é a λ-ésima raiz de uma variável aleatória $Y \sim \mathrm{beta}(1, \varphi)$. Ou seja,

$$T = \sqrt[\lambda]{Y},$$

com igualdade em distribuição.

Segundo Garg (2009), a distribuição Kum é tão versátil quanto a distribuição beta, mas com uma forma mais simples de uso, especialmente em estudos de simulação já que ela tem uma forma fechada simples para a fdp e para a fda.

Os momentos dessa distribuição são diretamente obtidos por

$$E(T^m) = \varphi B\left(\frac{\lambda + m}{\lambda}, \varphi\right),$$

lembrando que $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ é a função beta.

Assim, sua esperança e variância são dadas por

$$E(T) = \varphi B\left(\frac{\lambda + 1}{\lambda}, \varphi\right)$$

e

$$Var(T) = \varphi B\left(\frac{\lambda + 2}{\lambda}, \varphi\right) - \left[\varphi B\left(\frac{\lambda + 1}{\lambda, \varphi}\right)\right]^2.$$

Relação entre a distribuição Kumaraswamy e outras distribuições:

- Se $X \sim \mathrm{Kum}(1, 1)$, então $X \sim \mathrm{U}(0, 1)$;
- Se $X \sim \mathrm{U}(0, 1)$, então $\left(1 - (1 - X)^{\frac{1}{\varphi}}\right)^{\frac{1}{\lambda}} \sim \mathrm{Kum}(\lambda, \varphi)$;
- Se $X \sim \mathrm{beta}(1, \varphi)$, então $X \sim \mathrm{Kum}(1, \varphi)$;

- Se $X \sim \text{beta}(\lambda, 1)$, então $X \sim \text{Kum}(\lambda, 1)$;
- Se $X \sim \text{Kum}(\lambda, 1)$, então $1 - X \sim \text{Kum}(1, \lambda)$;
- Se $X \sim \text{Kum}(1, \lambda)$, então $1 - X \sim \text{Kum}(\lambda, 1)$;
- Se $X \sim \text{Kum}(\lambda, 1)$, então $-\log(X) \sim \text{exponencial}(\lambda)$;
- Se $X \sim \text{Kum}(1, \varphi)$, então $-\log(1 - X) \sim \text{exponencial}(\varphi)$;
- Se $X \sim \text{Kum}(\lambda, \varphi)$, então $X \sim \text{BG1}(\lambda, 1, 1, \varphi)$, em que $BG1$ é a distribuição beta generalizada do tipo 1, com função densidade

$$f(x; a, b, p, q) = \frac{|a| x^{ap-1} (1 - (x/b)^a)^{q-1}}{b^{ap} B(p, q)},$$

para $0 < x^a < b^a$ em que b, p e q são positivos.

1.3 Organização dos Capítulos

O presente trabalho está dividido em cinco capítulos. No Capítulo 2 será apresentada a família Kum-G, sendo que consideramos um de seus casos particulares, a distribuição Kumaraswamy Weibull modificada e descrevemos detalhadamente a Kumaraswamy exponencial, que é um caso particular dela, inclusive com uma abordagem clássica e bayesiana. O modelo de longa duração será descrito no Capítulo 3, considerando a teoria unificada. No Capítulo 4 temos a família Kum-G de longa duração com covariáveis. E finalmente no Capítulo 5 temos algumas considerações finais e propostas futuras.

Capítulo

2

Distribuição Kumaraswamy Generalizada

Como descrito anteriormente, o tempo até a ocorrência de algum evento de interesse, em geral, pode ser acomodado por uma distribuição de probabilidade. Na literatura, várias distribuições têm sido utilizadas para descrever tempos de sobrevivência mas a maioria das distribuições comumente utilizadas não apresenta flexibilidade para modelar funções de risco não monótonas, tais como, em forma unimodal e banheira, o que são bem comuns em estudos biológicos. Dessa forma, neste capítulo estudaremos a distribuição Kumaraswamy generalizada, pois é uma distribuição flexível porém simples.

A distribuição Kumaraswamy generalizada (Kum-G) apresentada por Cordeiro e de Castro (2011) apresenta flexibilidade em acomodar diversas formas para a função de risco, podendo ser usada em uma variedade de problemas para modelar dados de sobrevivência. Ela é uma generalização da distribuição Kumaraswamy, com o acréscimo de uma função distribuição $G(t)$ de uma família de distribuições contínuas.

Definição: Seja $G(t)$ a função de distribuição acumulada de uma variável aleatória contínua qualquer. A fda da distribuição Kum-G é dada por

$$F(t) = 1 - \left[1 - G(t)^{\lambda}\right]^{\varphi}, \tag{2.1}$$

em que $\lambda > 0$ e $\varphi > 0$. Seja $g(t) = \frac{dG(t)}{dt}$, a fdp correspondente é

$$f(t) = \lambda\varphi g(t) G(t)^{\lambda-1} \left[1 - G(t)^{\lambda}\right]^{\varphi-1}. \tag{2.2}$$

Dessa forma, obtemos a função de sobrevivência e de risco, dadas respectivamente por

$$S(t) = \left[1 - G(t)^{\lambda}\right]^{\varphi} \tag{2.3}$$

e

$$h(t) = \frac{\lambda\varphi g(t)G(t)^{\lambda-1}}{1-G(t)^{\lambda}}. \tag{2.4}$$

Existem na literatura diversas distribuições que foram generalizadas, uma delas é a distribuição beta. A fdp das generalizações da beta utilizam a função beta, que é de difícil manuseio. Por outro lado, a distribuição Kum-G é uma generalização que não apresenta nenhuma função complicada em sua fdp, sendo mais vantajosa do que muitas generalizações.

Como a distribuição Kum-G depende de uma função distribuição $G(t)$, então para cada distribuição contínua, temos um caso da Kum-G com o número de parâmetros da $G(t)$ mais os dois parâmetros λ e φ. Por exemplo, se tivermos como $G(t)$ a função distribuição acumulada da distribuição exponencial, teremos neste caso a distribuição Kumaraswamy exponencial. Na literatura, diversos casos dessa distribuição foram estudados, alguns deles são: Kumaraswamy normal (Correa *et al*, 2012), Kumaraswamy log-logística (de Santana *et al.* , 2012), Kumaraswamy pareto (Bourguignon *et al*, 2013), Kumaraswamy pareto generalizada (Nadarajah e Eljabri, 2013), Kumaraswamy gamma generalizada (de Pascoa *et al*, 2011), Kumaraswamy *half*-normal generalizada (Cordeiro *et al*, 2012), Kumaraswamy Weibull inversa (Shahbaz *et al*, 2012) e Kumaraswamy Rayleigh inversa (Hussian e A Amin, 2014).

2.1 Distribuição Kumaraswamy Weibull Modificada

Como a distribuição Weibull Modificada é uma distribuição muito utilizada em análise de sobrevivência e tem outras distribuições amplamente utilizadas como casos particulares, nesta seção a utilizamos para chegar em um caso especial da Kum-G. A distribuição Kumaraswamy Weibull modificada (Kum-WM) foi inicialmente estudada por Cordeiro *et al.* (2014). Para estabelecer esta distribuição Kum-WM, devemos tomar a função $G(t)$ como a função distribuição da Weibull Modificada, dada em 1.14.

Seja T a variável aleatória contínua que segue distribuição Kumaraswamy Weibull modificada, sua notação é $T \sim$ Kum-WM$(\varphi, \lambda, \alpha, c, \beta)$. Para $t > 0$, a função distribuição acumulada é dada por

$$F(t) = 1 - \left[1 - \left(1 - \mathrm{e}^{-\alpha t^{c}\mathrm{e}^{\beta t}}\right)^{\lambda}\right]^{\varphi}. \tag{2.5}$$

A função de densidade de probabilidade é determinada por

$$f(t) = \lambda\varphi\alpha t^{c-1}(c+\beta t)\mathrm{e}^{\beta t-\alpha t^{c}\mathrm{e}^{\beta t}}\left(1-\mathrm{e}^{-\alpha t^{c}\mathrm{e}^{\beta t}}\right)^{\lambda-1}\left[1-\left(1-\mathrm{e}^{-\alpha t^{c}\mathrm{e}^{\beta t}}\right)^{\lambda}\right]^{\varphi-1}. \tag{2.6}$$

É possível determinar a função de sobrevivência e a função de risco, respectivamente, dadas por

$$S(t) = \left[1 - \left(1 - \mathrm{e}^{-\alpha t^c \mathrm{e}^{\beta t}}\right)^{\lambda}\right]^{\varphi} \tag{2.7}$$

e

$$h(t) = \frac{\lambda\varphi\alpha t^{c-1}(c+\beta t)\mathrm{e}^{\beta t - \alpha t^c \mathrm{e}^{\beta t}}\left(1 - \mathrm{e}^{-\alpha t^c \mathrm{e}^{\beta t}}\right)^{\lambda-1}}{1 - \left(1 - \mathrm{e}^{-\alpha t^c \mathrm{e}^{\beta t}}\right)^{\lambda}}. \tag{2.8}$$

Na Tabela 2.1 temos casos particulares da distribuição Kum-WM.

Tabela 2.1: Casos particulares da distribuição Kumaraswamy Weibull modificada.

Distribuição	Parametrização	Função distribuição acumulada
Kumaraswamy exponencial	$\beta = 0$ e $c = 1$	$F(t) = 1 - [1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}$
Kumaraswamy Rayleigh	$\beta = 0$ e $c = 2$	$F(t) = 1 - [1 - (1 - \mathrm{e}^{-\alpha t^2})^{\lambda}]^{\varphi}$
Kumaraswamy valor extremo	$c = 0$	$F(t) = 1 - [1 - (1 - \mathrm{e}^{-\alpha \mathrm{e}^{\beta t}})^{\lambda}]^{\varphi}$
Kumaraswamy Weibull	$\beta = 0$	$F(t) = 1 - [1 - (1 - \mathrm{e}^{-\alpha t^c})^{\lambda}]^{\varphi}$
Weibull modificada	$\lambda = 1$ e $\varphi = 1$	$F(t) = 1 - \mathrm{e}^{-\alpha t^c \mathrm{e}^{\beta t}}$
Exponencial	$\lambda = 1$, $\varphi = 1$, $\beta = 0$ e $c = 1$	$F(t) = 1 - \mathrm{e}^{-\alpha t}$
Rayleigh	$\lambda = 1$, $\varphi = 1$, $\beta = 0$ e $c = 2$	$F(t) = 1 - \mathrm{e}^{-\alpha t^2}$
Valor-extremo	$\lambda = 1$, $\varphi = 1$, $c = 0$	$F(t) = 1 - \mathrm{e}^{-\alpha \mathrm{e}^{\beta t}}$
Weibull	$\lambda = 1$, $\varphi = 1$, $\beta = 0$	$F(t) = 1 - \mathrm{e}^{-\alpha t^c}$

A seguir, trataremos com mais detalhes a distribuição Kumaraswamy exponencial, que é um caso particular da distribuição Kum-WM e que não foi abordada na literatura com muitos detalhes.

2.1.1 Distribuição Kumaraswamy Exponencial

Nesta seção será apresentado o estudo da distribuição Kumaraswamy exponencial (Kum-Exp) para modelar o tempo de vida de indivíduos em risco. Uma importante característica desta distribuição é a habilidade de modelar funções de risco monótonas e não monótonas.

Para estabelecer a distribuição Kum-Exp como um caso especial da Kum-G, a função $G(t)$ deve ser a função distribuição acumulada da distribuição exponencial com parâmetro α.

Seja T a variável aleatória contínua que segue distribuição Kumaraswamy exponencial, com notação $T \sim$ Kum-Exp$(\varphi, \lambda, \alpha)$. Os parâmetros $\varphi > 0$, $\lambda > 0$ são parâmetros de forma e $\alpha > 0$ é o parâmetro de escala. Para $t > 0$, a função densidade de probabilidade é dada por

$$f(t) = \varphi\lambda\alpha\mathrm{e}^{-\alpha t}(1 - \mathrm{e}^{-\alpha t})^{\lambda-1}[1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi-1}. \tag{2.9}$$

A função distribuição acumulada é determinada por

$$F(t) = 1 - [1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}. \tag{2.10}$$

Quando os parâmetros assumem os valores ($\varphi = 1$,$\lambda = 1$), ($\alpha = 1$,$\lambda = 1$) e ($\lambda = 1$), temos casos particulares da Kum-Exp, em que todos chegam na distribuição exponencial, diferindo apenas nos parâmetros, sendo, respectivamente, exponencial(α), exponencial(φ) e exponencial($\alpha\varphi$).

É possível determinar a função de sobrevivência e a função de risco, respectivamente, dadas por

$$S(t) = [1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi} \tag{2.11}$$

e

$$h(t) = \frac{\varphi\lambda\alpha\mathrm{e}^{-\alpha t}(1 - \mathrm{e}^{-\alpha t})^{\lambda-1}}{1 - (1 - \mathrm{e}^{-\alpha t})^{\lambda}}. \tag{2.12}$$

Conforme será visto, o modelo Kum-Exp pode oferecer grande flexibilidade para modelar a distribuição do tempo de sobrevivência. As representações gráficas da fdp, da função de sobrevivência e de risco, para alguns valores dos parâmetros, são apresentadas na Figura 2.1.

Interpretação física: Seja um sistema formado por φ componentes independentes, cada um destes componentes é formado por λ subcomponentes independentes. O sistema falha se qualquer um dos φ componentes falhar e um dos componentes falha se todos os λ sub-componentes falharem, ou seja, os φ componentes estão em série e os λ componentes estão em paralelo. Também, $T_{j1}, ..., T_{j\lambda}$ representam os tempos de sobrevivência dos sub-componentes do j-ésimo componente, $j = 1, ..., \varphi$, todos eles tendo a mesma fda $G(t)$, que é a fda da distribuição exponencial. Suponha T_k o tempo de sobrevivência do componente k, para $k = 1, ..., \varphi$, e T o tempo de sobrevivência de todo o sistema. Dessa forma, a distribuição Kum-Exp pode ser interpretada como a distribuição do tempo de falha do sistema inteiro.

A esperança e a variância são comumente utilizadas para expressar uma medida de tendência central e de variabilidade dos dados, respectivamente. Cordeiro e de Castro (2011) propuseram o cálculo para os momentos da distribuição Kum-Exp dado pela seguinte fórmula

$$E(T^n) = n\lambda^n \sum_{i,j=1}^{\infty} W_i \frac{-1^{n+j}\binom{\lambda(i+1)-1}{j}}{(j+1)^{n+1}}, \tag{2.13}$$

em que $T > 0$ é a variável aleatória; $W_i = (-1)^i\lambda\varphi\binom{\varphi-1}{i}$; $j = 1, \ldots, \varphi$; $i = 1, \ldots, \lambda$; sendo que φ é um valor positivo natural que indica o número de componentes independentes e λ é um valor positivo natural que indica o número de subcomponentes independentes. Fazendo $n = 1$ teremos a esperança da distribuição ($E(T)$) e, para encontrarmos a variância basta fazer o cálculo de $E(T^2) - [E(T)]^2$.

A função quantílica, utilizada em simulações, é dada por

$$F^{-1}(t) \;=\; G^{-1}\left\{\left[1 - (1-t)^{1/\varphi}\right]^{1/\lambda}\right\} = -\frac{1}{\alpha}\log\left\{\left[1 - (1-t)^{1/\varphi}\right]^{1/\lambda}\right\}. \tag{2.14}$$

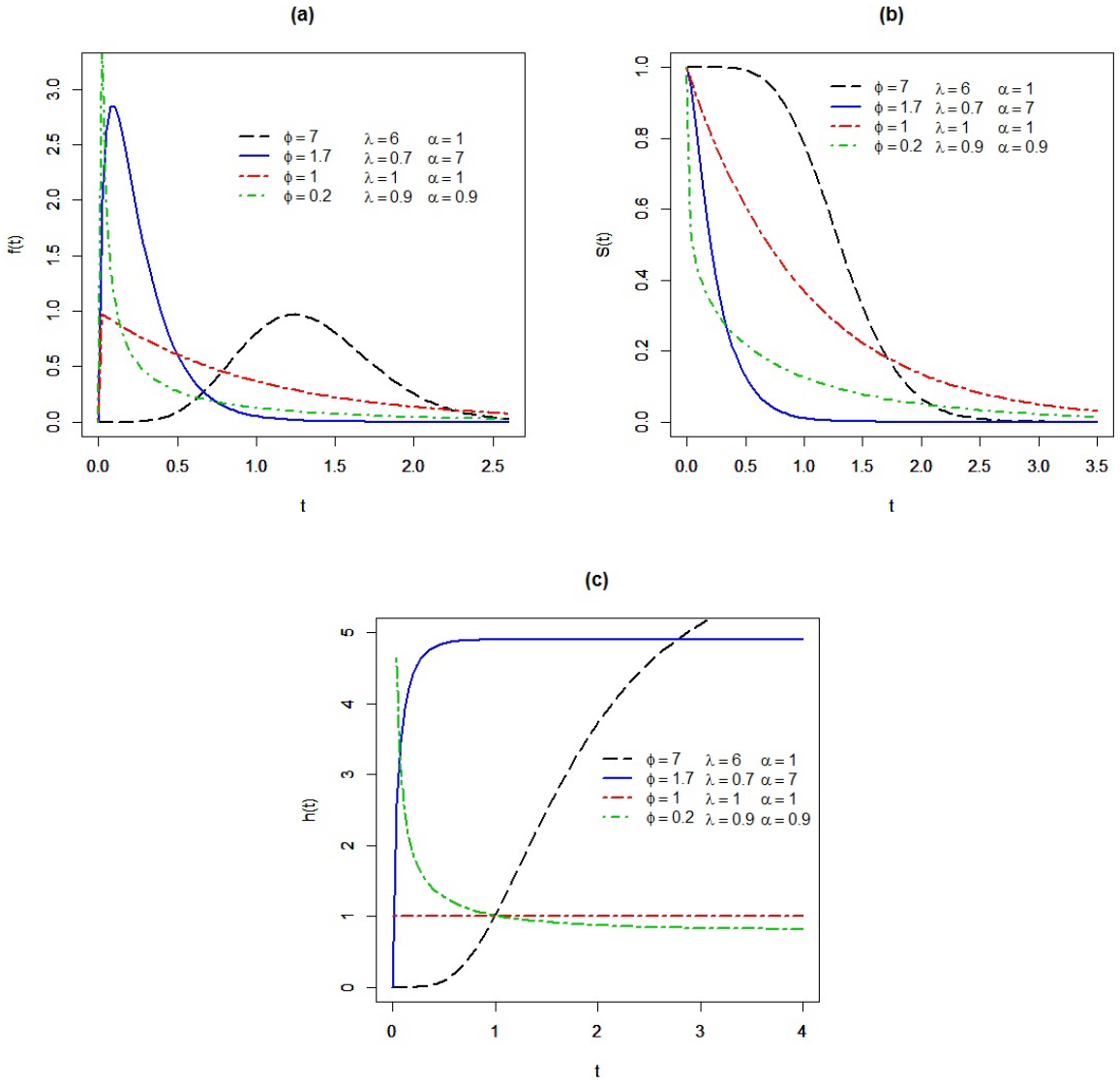

Figura 2.1: Gráficos das funções de densidade (a), sobrevivência (b) e risco (c) da distribuição Kum-Exp.

Assim, basta simular valores de uma variável aleatória U(0,1), substituir em t na função quantílica e teremos valores simulados da distribuição Kum-Exp$(\varphi, \lambda, \alpha)$.

Para modelar o tempo de vida de indivíduos é necessário estimar as funções de sobrevivência e de risco (2.11) e (2.12) da distribuição Kum-Exp, e para tal basta estimar os parâmetros φ, λ e α, devido à propriedade de invariância do estimador de máxima verossimilhança.

Considerando dados de sobrevivência, a partir dos tempos observados $\boldsymbol{t} = (t_1, t_2, ..., t_n)$ e do vetor de parâmetros $\boldsymbol{\theta} = (\varphi, \lambda, \alpha)$, a função de verossimilhança da distribuição Kum-Exp é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}} \right]^{\delta_i} \left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi} \tag{2.15}$$

na qual δ_i é o indicador de falha (1 indica falha e 0 indica censura).

O logaritmo da função de verossimilhança é

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\delta_i \log\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right] \\ &\quad + \sum_{i=1}^{n}\varphi\log\left[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}\right].\end{aligned} \tag{2.16}$$

As derivadas de primeira ordem para os parâmetros λ, φ e α são dadas por

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\varphi} &= \sum_{i=1}^{n}\frac{\delta_i}{\varphi} + \sum_{i=1}^{n}\log(1-(1-e^{-\alpha t_i})^{\lambda}) \\ &= \frac{\sum_{i=1}^{n}\delta_i}{\hat{\varphi}} + \sum_{i=1}^{n}\log(1-(1-e^{-\alpha t_i})^{\lambda}),\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\lambda} &= \sum_{i=1}^{n}\delta_i\left[\frac{1}{\lambda}+\log(1-e^{-\alpha t_i}) + \frac{(1-e^{-\alpha t_i})^{\lambda}\log(1-e^{-\alpha t_i})}{1-(1-e^{-\alpha t_i})^{\lambda}}\right] \\ &\quad +\varphi\sum_{i=1}^{n}\left[-\frac{(1-e^{-\alpha t_i})^{\lambda}\log(1-e^{-\alpha t_i})}{1-(1-e^{-\alpha t_i})^{\lambda}}\right],\end{aligned}$$

e

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\alpha} &= \sum_{i=1}^{n}\delta_i\left[\frac{1}{\alpha}-t_i+\frac{(\lambda-1)t_i e^{-\alpha t_i}}{1-e^{-\alpha t_i}} + \frac{\lambda(1-e^{-\alpha t_i})^{\lambda-1}t_i e^{-\alpha t_i}}{1-(1-e^{-\alpha t_i})^{\lambda}}\right] \\ &\quad +\varphi\sum_{i=1}^{n}\left[\frac{-\lambda(1-e^{-\alpha t_i})^{\lambda-1}t_i e^{-\alpha t_i}}{1-(1-e^{-\alpha t_i})^{\lambda}}\right].\end{aligned}$$

Os estimadores de máxima verossimilhança para os três parâmetros são encontrados a partir da resolução das equações de verossimilhança

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0}. \tag{2.17}$$

Resolvendo a equação de verossimilhança (2.17) para φ, o seu EMV é

$$\hat{\varphi} = -\frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n}\log[1-(1-\mathrm{e}^{-\hat{\alpha}t_i})^{\hat{\lambda}}]}. \tag{2.18}$$

Embora tenha sido possível encontrar uma expressão para o parâmetro φ, ele depende dos demais parâmetros λ e α, e seus respectivos estimadores não puderam ser encontrados de forma analítica, como pode-se ver pelas suas respectivas equações de verossimilhança. Portanto, as estimativas por máxima verossimilhança para os três parâmetros φ, λ e α devem ser obtidas via métodos numéricos, maximizando o logaritmo da verossimilhança.

2.1.1.1 Estudo de Simulação

Um estudo de simulação foi realizado supondo que o tempo de vida dos indivíduos em risco segue uma distribuição Kum-Exp, afim de verificar as propriedades frequentistas (por exemplo, verificar se o valor estimado do parâmetro se aproxima cada vez mais do verdadeiro valor conforme o tamanho amostral aumenta e, por consequência, o erro quadrático médio deve diminuir). Foram realizados três estudos, o primeiro com 25% de censura, o segundo com 50% de censura e o terceiro com 75% de censura. Estamos supondo que os tempos para ambos os estudos seguem a distribuição Kum-Exp com os parâmetros $\varphi = 0,5$, $\lambda = 1$ e $\alpha = 1$. Ambos os conjuntos de parâmetros foram escolhidos ao acaso. Como o campo de variação dos parâmetros é positivo, foi feita uma reparametrização em que temos $\varphi = \exp(\varphi_1)$, $\lambda = \exp(\lambda_1)$ e $\alpha = \exp(\alpha_1)$. Essa reparametrização é válida devido à propriedade de invariância dos estimadores de máxima verossimilhança.

Utilizamos quatro tamanhos amostrais: $n = 30$, $n = 50$, $n = 200$ e $n = 500$. Foram geradas 1.000 réplicas, sendo que em cada réplica foram obtidas as estimativas de máxima verossimilhança de cada um dos parâmetros, a partir da rotina "optim" da linguagem "R" (R Core Team, 2013). Foram atribuídos como valores iniciais para o processo de otimização os valores dos parâmetros fixados para a simulação dos tempos.

Em cada simulação foi obtido o erro padrão, erro quadrático médio (EQM) de cada estimativa e a probabilidade de cobertura (PC) através do intervalo de confiança assintótico de 95%. O EQM foi calculado da seguinte maneira:

$$EQM(\widehat{\theta_i}) = \sum_{j=1}^{d} \frac{(\theta_i - \widehat{\theta_{ij}})^2}{d},$$

em que d é o número de réplicas, θ_i é o valor fixado do parâmetro i, $i = 1, 2, 3$, e $\widehat{\theta_{ij}}$ é a estimativa do parâmetro i na j-ésima réplica, no qual $\boldsymbol{\theta} = (\varphi, \lambda, \alpha)$.

O intervalo com 95% de confiança (IC(95)) foi calculado da seguinte forma:

$$IC(1-\gamma) = \widehat{\theta_i} \ \pm \ \hat{\sigma_i} \, Z_{\gamma/2},$$

em que $Z_{\gamma/2}$ é o quantil $\gamma/2$ da distribuição normal padrão e $\hat{\sigma_i}$ é o erro padrão estimado da estimativa de cada parâmetro. Vale ressaltar que devido à reparametrização dos parâmetros, o cálculo do erro padrão foi realizado pelo método delta, ver Casella e Berger (2002). O valor de γ utilizado foi de 0,05.

Os tempos foram gerados utilizando as seguintes etapas:

- Fixar os valores dos parâmetros;

- Gerar $u_i \sim U(0, 1)$;

- Gerar o tempo de falha y_i tal que $y_i = -\dfrac{\log(1-(1-(1-u_i)^{(1/\lambda)})^{(1/\varphi)})}{\alpha}$, utilizando os parâmetros do tempo de falha.
- Gerar o tempo de censura $x_i \sim U(0, \mu_c)$;
- Fazer $t_i = \min(y_i, x_i)$;
- Se $y_i < x_i$, então $\delta_i = 1$, caso contrário, $\delta_i = 0$, $i = 1, ..., n$.

em que μ_c é calculado de acordo proporção de censuras k de cada estudo, da seguinte forma

$$P(Y > X) = \int_0^\infty S_Y(y) f_X(y) dy,$$

sendo $S_Y(y)$ é a função de sobrevivência da distribuição Kum-Exp e $f_X(x) = 1/\mu_c$. Assim,

$$P(Y > X) = \frac{1}{\mu_c} \int_0^\infty [1-(1-\mathrm{e}^{-1y})^1]^1 dy = \frac{1}{\mu_c} \int_0^\infty \mathrm{e}^{-y} dy = \frac{1}{\mu_c} = k.$$

Dessa forma,

$$\mu_c = \frac{1}{k}.$$

Então, no estudo com 25% de censura, temos $\mu_c = \frac{1}{0{,}25} = 4$; no caso em que temos 50% de censura, $\mu_c = \frac{1}{0{,}50} = 2$ e no estudo em que temos 75% de censura, $\mu_c = \frac{1}{0{,}75} = 1,33$.

As estimativas obtidas através dos dados com 25% de censuras estão apresentadas na Tabela 2.2, as estimativas obtidas através dos dados com 50% de censuras estão apresentadas na Tabela 2.3 e as estimativas obtidas através dos dados com 75% de censuras estão apresentadas na Tabela 2.4. O interesse é saber se, conforme o tamanho da amostra aumenta, as estimativas se aproximam dos parâmetros fixados em $\varphi = 0.5$, $\lambda = 1$ e $\alpha = 1$,

Tabela 2.2: Estimativas da Simulação - Dados com 25% de Censura

n	Parâmetros	Estimativas	Erro Padrão	EQM	PC
30	φ	1,370	14,446	15,399	0,936
	λ	1,135	0,632	0,515	0,959
	α	1,611	9,201	5,614	0,963
100	φ	1,265	9,800	1,019	0,981
	λ	1,043	0,267	0,052	0,985
	α	1,108	8,243	1,292	0,994
200	φ	1,027	9,344	0,022	0,998
	λ	1,008	0,177	0,014	0,992
	α	0,994	9,163	0,086	0,999
500	φ	1,051	9,862	0,054	0,998
	λ	1,006	0,104	0,005	0,991
	α	0,985	9,559	0,088	0,999

Tabela 2.3: Estimativas da Simulação - Dados com 50% de Censura

n	Parâmetros	Estimativas	Erro Padrão	EQM	PC
30	φ	1,324	13,897	1,936	0,903
	λ	1,401	1,235	4,629	0,954
	α	2,262	12,500	19,494	0,960
100	φ	1,124	10,779	0,313	0,972
	λ	1,015	0,498	0,049	0,988
	α	1,258	12,252	2,317	0,998
200	φ	1,027	10,372	0,015	0,996
	λ	1,005	0,175	0,016	0,986
	α	1,021	10,286	0,406	1,000
500	φ	1,014	10,945	0,006	1,000
	λ	1,000	0,112	0,006	0,992
	α	0,992	10,858	0,017	1,000

Tabela 2.4: Estimativas da Simulação - Dados com 75% de Censura

n	Parâmetros	Estimativas	Erro Padrão	EQM	PC
30	φ	1,633	19,701	50,151	0,904
	λ	1,766	1,855	49,771	0,968
	α	2,514	12,434	30,759	0,966
100	φ	1,284	13,512	1,850	0,969
	λ	1,024	0,320	0,077	0,980
	α	1,460	12,186	7,526	0,993
200	φ	1,058	13,872	0,465	0,988
	λ	1,008	0,196	0,021	0,993
	α	1,136	13,409	2,521	0,997
500	φ	1,103	12,611	0,469	0,973
	λ	0,999	0,130	0,008	0,991
	α	1,351	13,042	4,764	0,997

As Figuras 2.2 a 2.4 mostra o comportamento da estimativa de máxima verossimilhança, erro padrão, erro quadrático médio e probabilidade de cobertura das simulações com 25% de censura, 50% de censura e 75% de censura.

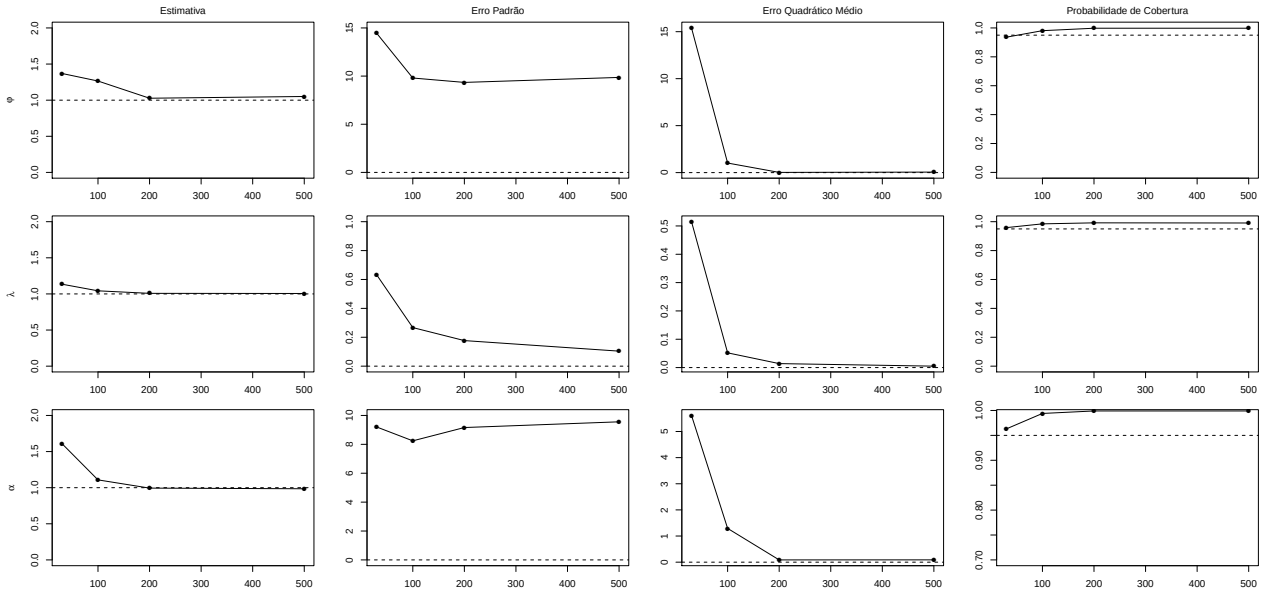

Figura 2.2: Comportamento da simulação com 25% de censura.

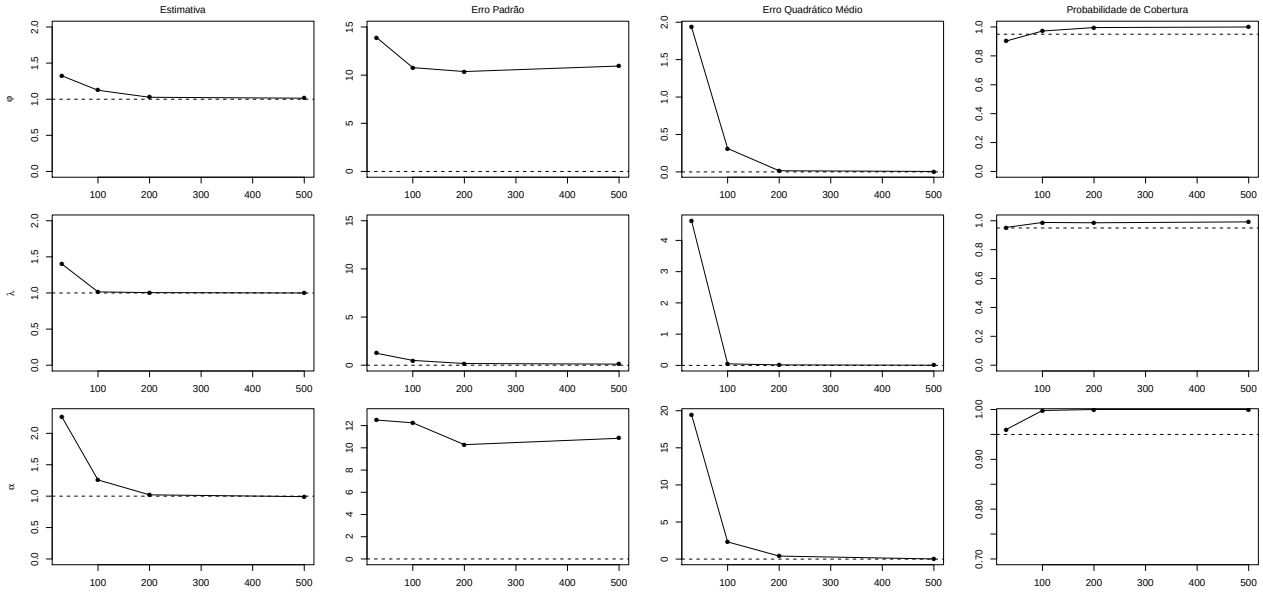

Figura 2.3: Comportamento da simulação com 50% de censura.

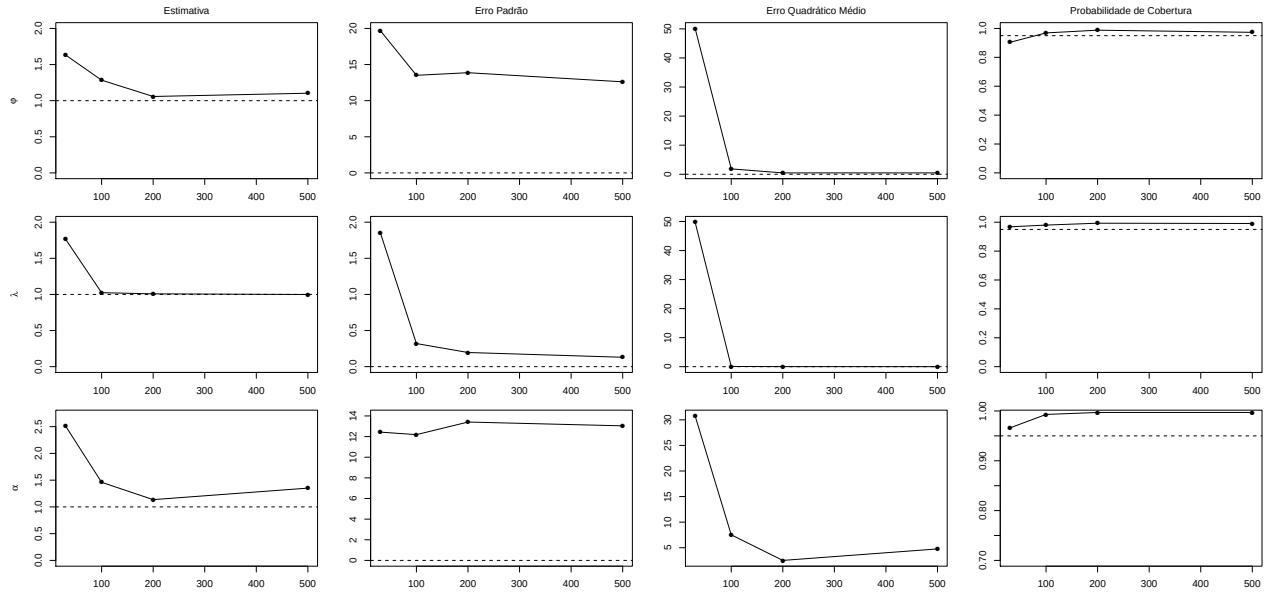

Figura 2.4: Comportamento da simulação com 75% de censura.

Os resultados da simulação mostram, de uma maneira geral, que conforme o tamanho amostral aumenta, as médias das estimativas para cada parâmetro ficaram próximas dos valores fixados, os EQMs decrescem e a probabilidade de cobertura aumenta, atendendo as propriedades frequentistas. Também, podemos notar que quanto maior a porcentagem de censura nos dados, precisa-se de mais dados para obtermos uma boa precisão.

2.1.1.2 Aplicação

Nesta seção foi considerado um conjunto de dados reais para ilustrar a metodologia proposta. Todos os cálculos e gráficos foram feitos utilizando a linguagem R.

Este conjunto de dados refere-se à um grupo de ratas que foi exposto ao câncer, e foi registrado o tempo, em anos, até a morte por câncer vaginal. Temos 19 observações e há a presença de duas censuras, isto é, o tempo que foi registrado ocorreu quando a rata saiu do experimento por outras razões e não pela morte devido à este tipo de câncer. O tempo máximo observado foi 0,668 anos, ou 244 dias. Em Pike (1966) encontram-se os dados de dois grupos de ratas, mas foi analisado apenas o primeiro grupo. Da mesma forma, o interesse é modelar a função de sobrevivência destes tempos.

Afim de conseguirmos supor uma distribuição adequada aos dados, o gráfico Tempo Total em Teste (TTT Plot) é de grande utilidade.

O TTT Plot é útil para a análise de dados não-negativos, ajudando na escolha de um modelo estatístico para os dados e fornece informações sobre a função de risco. Também, dados incompletos podem ser analisados. Existe uma base teórica para tal análise, ver Barlow e Campo (1975). Assim, sabendo sua forma, basta supor distribuições que tenham função de risco da mesma forma.

O gráfico TTT plot descrito por Aarset (1985) é calculado utilizando apenas tempos de falha, supondo que os dados são completos. Porém, em análise de sobrevivência geralmente temos a presença de censuras, o que torna equivocada a utilização desse método. Algumas vezes, podemos ter dados de longa duração, que são caracterizados por um grande número de censuras, principalmente no final do estudo. Dessa forma, o TTT plot poderia indicar uma forma da função de risco equivocada. Com o intuito de contornar este problema, Sun e Kececioglu (1999) propuseram uma modificação do TTT plot, que incorpora as censuras em seu cálculo. Assim, temos um método gráfico muito mais confiável para verificarmos a forma da função de risco para dados de análise de sobrevivência na presença de censuras. No Apêndice A apresentamos uma descrição de como é feito o TTT plot com e sem censuras.

Na Figura 2.5 temos o TTT plot do conjunto de dados com censuras. Neste caso temos um indicativo de que a função de risco é crescente, nos indicando que a distribuição Kumaraswamy Weibull modificada (Kum-WeiMod) é adequada aos dados, bem como seus casos particulares Kumaraswamy exponencial (Kum-Exp) e exponencial (Exp). Sendo assim, aplicamos os três modelos aos dados.

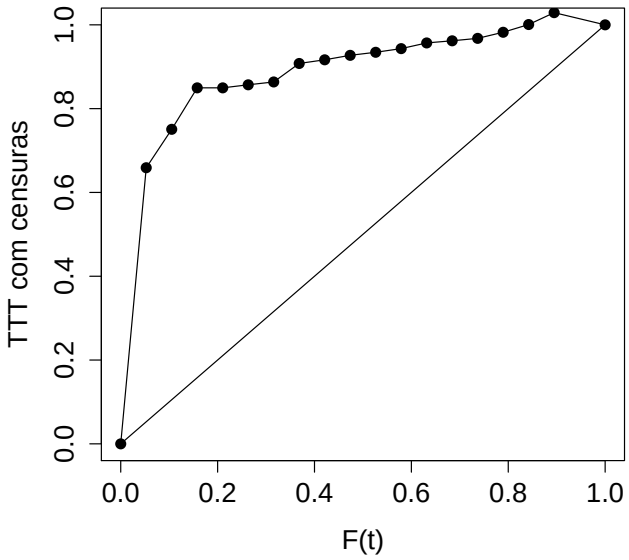

Figura 2.5: TTT plot com censuras dos dados.

Os resultados das estimativas de máxima verossimilhança dos parâmetros dos três modelos são apresentados na Tabela 2.5, bem como os erros padrões, o intervalo de confiança de 95% para cada parâmetro.

Na Tabela 2.5, vemos que alguns intervalos de confiança do modelo Kum-WM incluem valores negativos, mas isso provavelmente está acontecendo porque temos a presença de muitos parâmetros, e seus erros padrões estão um pouco altos. Esse problema poderia ser contornado utilizando intervalo de confiança via Bootstrap ou intervalo de credibilidade bayesiano. Também vemos que o modelo Exp tem intervalo de confiança menor comparado ao modelo

Tabela 2.5: Estimativas dos parâmetros do modelos.

Modelo	Parâmetros	Estimativas	EP	IC 95%
Kum-WM	λ	$12,298$	$4,934$	$(2,627; 21,968)$
	φ	$2,108$	$5,774$	$(-9,209; 13,425)$
	α	$3,429$	$2,768$	$(-1,996; 8,854)$
	β	$0,607$	$7,460$	$(-14,014; 15,228)$
	c	$1,445$	$3,281$	$(-4,986; 7,876)$
Kum-Exp	λ	$40,107$	$1,606$	$(36,959; 43,255)$
	φ	$4,681$	$1,899$	$(0,959; 8,403)$
	α	$5,114$	$0,805$	$(3,536; 6,691)$
Exp	α	$1,5153$	$0,3675$	$(0,7949; 2,2355)$

Kum-Exp e que o modelo Kum-Exp têm intervalos de confiança menores do que o modelo Kum-WM.

A Figura 2.6 apresenta os gráficos do ajuste da curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelos modelos Kum-WM, Kum-Exp e exponencial. Podemos observar que a curva do modelo Kum-WM e Kum-Exp estão bem mais próximas da curva de KM do que o modelo exponencial, sendo assim um indicativo de que estes modelos se adequaram bem aos dados. Neste caso, iremos utilizar critérios de seleção de modelos para verificar qual modelo, Kum-WM ou Kum-Exp, será selecionado para a aplicação.

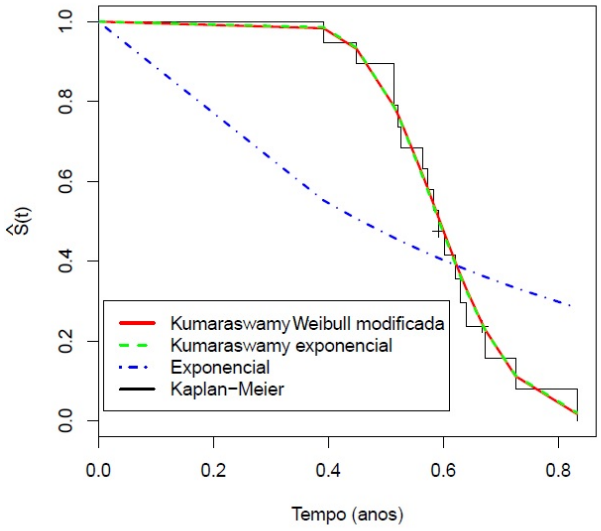

Figura 2.6: Curva de Kaplan-Meier juntamente com a função de sobrevivência estimada dos modelos Kum-WeiMod, Kum-Exp e do modelo exponencial.

Afim de selecionar modelos, algumas técnicas estatísticas são comumente utilizadas. Uma técnica tradicional utilizada para a seleção de modelos é o AIC (*Akaike Information Criterion*). A estatística AIC é dada por $AIC = -2l(L) + 2d$, em que $l(L)$ denota o máximo da função de log-verossimilhança e d é o número de parâmetros do modelo.

Schwarz propôs uma pequena alteração ao AIC, o BIC (*Bayesian Information Criterion*), que é definido como $BIC = -2l(L) + d\log(n)$, em que n representa o tamanho da amostra.

Tanto para o AIC quanto para o BIC, menores valores correspondem aos melhores modelos. É interessante notar que estes critérios comparam modelos que não são encaixados ou mesmo com números diferentes de parâmetros, pois consideram o número de parâmetros e penalizam a verossimilhança de modelos mais complexos.

Outra opção é analisar o logarítmo da função de verossimilhança. Os melhores modelos são com a maior $l(L)$.

Na Tabela 2.6 temos os valores dos critérios AIC, BIC e a log-verossimilhança de cada modelo. E, segundo os critérios, o modelo escolhido é o modelo Kum-Exp.

Tabela 2.6: Critérios AIC, BIC e Log-verossimilhança dos modelos

Modelo	AIC	BIC	Log-verossimilhança
Kumaraswamy Weibull modificada	$36,026$	$40,748$	$-13,013$
Kumaraswamy exponencial	$32,059$	$34,893$	$-13,029$

2.1.1.3 Abordagem Bayesiana

Sob o enfoque bayesiano, podemos expressar a incerteza a respeito dos parâmetros antes de observar os dados, utilizando uma distribuição *a priori* para os parâmetros, enquanto que a distribuição *a posteriori* une a informação contida na verossimilhança com a distribuição *a priori* e, basicamente, as estimativas bayesianas são construídas a partir dessa distribuição *a posteriori* .

Para se estimar de forma bayesiana, devemos construir a distribuição *a posteriori* que, pelo teorema de Bayes, é dada por

$$\pi(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{\delta}) \propto L(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{\delta})\pi(\boldsymbol{\theta}), \tag{2.19}$$

em que $\boldsymbol{\theta} = (\varphi, \lambda, \alpha)$ é o conjunto de parâmetros do modelo, $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ os tempos observados, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ o indicador de falha e $L(\boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{\delta})$ é a função de verossimilhança do modelo.

Considerando que os parâmetros são independentes, então a distribuição *a priori* de $\boldsymbol{\theta}$ é dada por

$$\pi(\boldsymbol{\theta}) = \pi(\lambda|\alpha_1, \beta_1)\pi(\varphi|\alpha_2, \beta_2)\pi(\alpha|\alpha_3, \beta_3), \tag{2.20}$$

em que α_1, β_1, α_2, β_2, α_3 e β_3 são os hiperparâmetros relacionados a $\boldsymbol{\theta}$.

Como o parâmetro λ assume valores positivos, admitimos que λ tem distribuição *a priori* $\text{gama}(\alpha_1, \beta_1)$ e, da mesma forma, admitimos que φ tem distribuição *a priori* $\text{gama}(\alpha_2, \beta_2)$ e que α tem distribuição *a priori* $\text{gama}(\alpha_3, \beta_3)$.

Combinando a função de verossimilhança (3.37) e a densidade *a priori* de $\boldsymbol{\theta}$ (2.20), obtemos a densidade *a posteriori* , dada por

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) &\propto L(\boldsymbol{\theta};\boldsymbol{t},\boldsymbol{\delta})\pi(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\quad \pi(\lambda|\alpha_1,\beta_1)\pi(\varphi|\alpha_2,\beta_2)\pi(\alpha|\alpha_3,\beta_3). \end{aligned} \tag{2.21}$$

Para se obter as densidades marginais *a posteriori* de cada um dos parâmetros, deve-se integrar (2.21) com relação aos parâmetros, mas estas integrais não são analiticamente calculáveis. Uma alternativa é fazer o uso das densidades condicionais completas de todos os parâmetros, dadas por

$$\begin{aligned}\pi(\lambda|\varphi,\alpha,\boldsymbol{t},\boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\quad \lambda^{\alpha_1-1}e^{-\beta_1\lambda}\end{aligned}$$

$$\begin{aligned}\pi(\varphi|\lambda,\alpha,\boldsymbol{t},\boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\quad \varphi^{\alpha_2-1}e^{-\beta_2\varphi}\end{aligned}$$

$$\begin{aligned}\pi(\alpha|\lambda,\varphi,\boldsymbol{t},\boldsymbol{\delta}) &\propto \prod_{i=1}^{n}\left[\frac{\varphi\lambda\alpha \mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}}{1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}}\right]^{\delta_i}\left[\left(1-\left(1-\mathrm{e}^{-\alpha t_i}\right)^{\lambda}\right)^{\varphi}\right] \\ &\quad \alpha^{\alpha_3-1}e^{-\beta_3\alpha}\end{aligned}$$

As densidades condicionais não apresentam nenhuma distribuição conhecida, então pode ser feito o uso do algoritmo de Metropolis-Hastings para gerar valores de λ, φ e α. Tal algoritmo permite simular amostras de distribuições conjuntas complexas, utilizando as distribuições condicionais completas dos parâmetros desconhecidos.

A convergência do algoritmo de Metropolis-Hastings pode ser verificada a partir de técnicas gráficas e de algum método numérico, sendo que o utilizado foi de Gelman-Rubin (Gelman e Rubin, 1992), que está implementado no sistema R.

2.1.1.4 Aplicação

Nesta seção consideramos o mesmo um conjunto de dados para ilustrar a metodologia bayesiana. Os cálculos foram feitos no software R e WinBUGS.

Consideramos para as distribuições *a priori* a distribuição gama, sendo a média o valor estimado dos parâmetros pelo método de máxima verossimilhança, da seguinte forma: $\pi(\lambda) \sim \mathrm{gama}(4;0,1)$, $\pi(\varphi) \sim \mathrm{gama}(2,2;0,5)$ e $\pi(\alpha) \sim \mathrm{gama}(2,6;\,0,5)$. Realizamos um estudo de

sensibilidade em que observamos que os resultados não se alteram com diferentes valores dos parâmetros das prioris.

Foram simulados valores de cada um dos parâmetros através do WnBUGS, com um burn-in de 1000000 de valores e depois 1000000 de valores foram simulados, com um salto de 40 em 40 valores afim de não se ter correlação entre os valores simulados.

Os resultados das estimativas de máxima verossimilhança dos parâmetros da Kum-Exp e a média *a posteriori* são apresentados na Tabela 2.7, bem como os erros padrões, o intervalo de confiança (e de credibilidade) de 95% e o método de Gelman-Rubin para cada parâmetro.

Tabela 2.7: Estimativas dos parâmetros do modelo Kum-Exp

Abordagem	Parâmetros	Estimativas	EP	IC 95%	R
Clássica	λ	$40,107$	$1,606$	$(36,959; 43,255)$	
	φ	$4,681$	$1,899$	$(0,959; 8,403)$	
	α	$5,114$	$0,805$	$(3,536; 6,691)$	
Bayesiana	λ	$39,990$	$20,01$	$(10,880; 87,690)$	1
	φ	$4,397$	$2,962$	$(0,616; 11,810)$	1
	α	$5,203$	$3,225$	$(0,906; 13,160)$	1

A Figura 2.7 apresenta os gráficos do ajuste da curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelos modelos Kum-Exp clássico e bayesiano. Podemos observar que a curva do modelo Kum-Exp clássico e do bayesiano estão bem próximas.

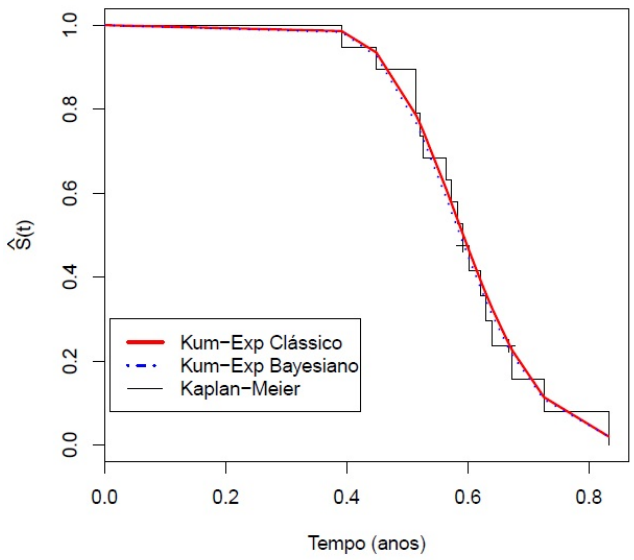

Figura 2.7: Curva de Kaplan-Meier juntamente com a função de sobrevivência estimada dos modelos Kum-Exp.

Nas Figuras 2.8 - 2.10, temos três métodos gráficos que foram utilizados para verificar a convergência do algoritmo bayesiano. O primeiro método gráfico é através de histogramas, basta dividir os dados simulados de cada um dos parâmetros em três partes iguais e representar

graficamente a primeira e a última parte. Na Figura 2.8 os histogramas dos três parâmetros são apresentados, em que a cor azul é da primeira parte dos valores simulados, a rosa é da segunda parte e a lilás é o encontro de ambos, e nos três casos os histogramas ficaram muito próximos. O segundo método é através do gráfico das estimativas das funções de densidade da primeira e da última parte dos valores. A Figura 2.9 mostra que estes gráficos da estimativa das funções de densidade dos três parâmetros estão muito próximos. O terceiro método para verificar a convergência é através da sequência dos valores simulados da primeira e da última parte dos valores, uma das partes está em vermelho e a outra em preto. A Figura 2.10 mostra que em todos os casos os gráficos novamente estão muito próximos e sem nenhuma tendência. Dessa forma, temos um indicativo de convergência do algoritmo.

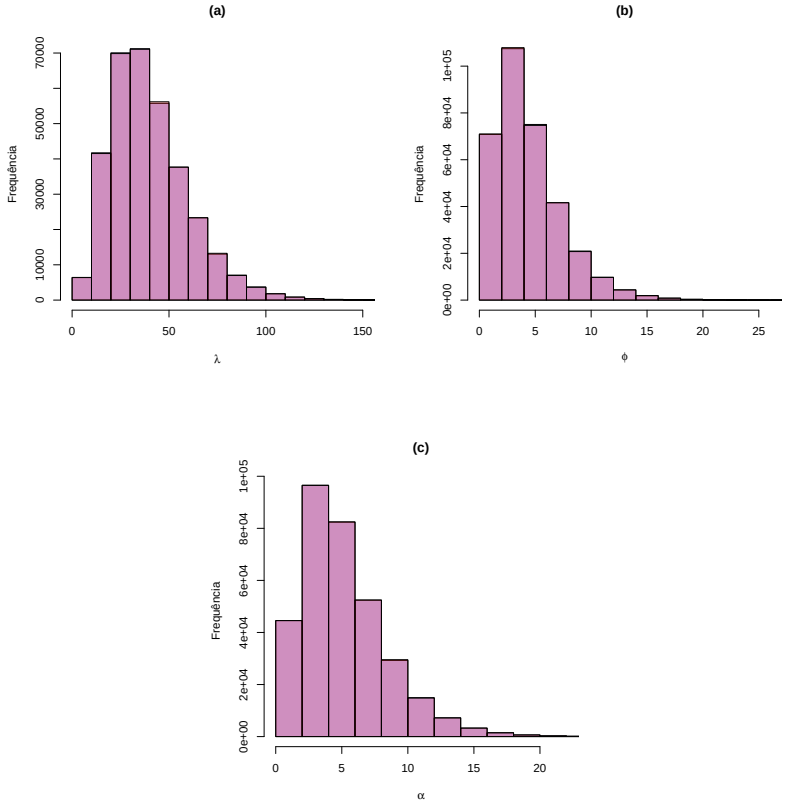

Figura 2.8: Gráficos dos histogramas do parâmetro λ (a), φ (b) e α (c).

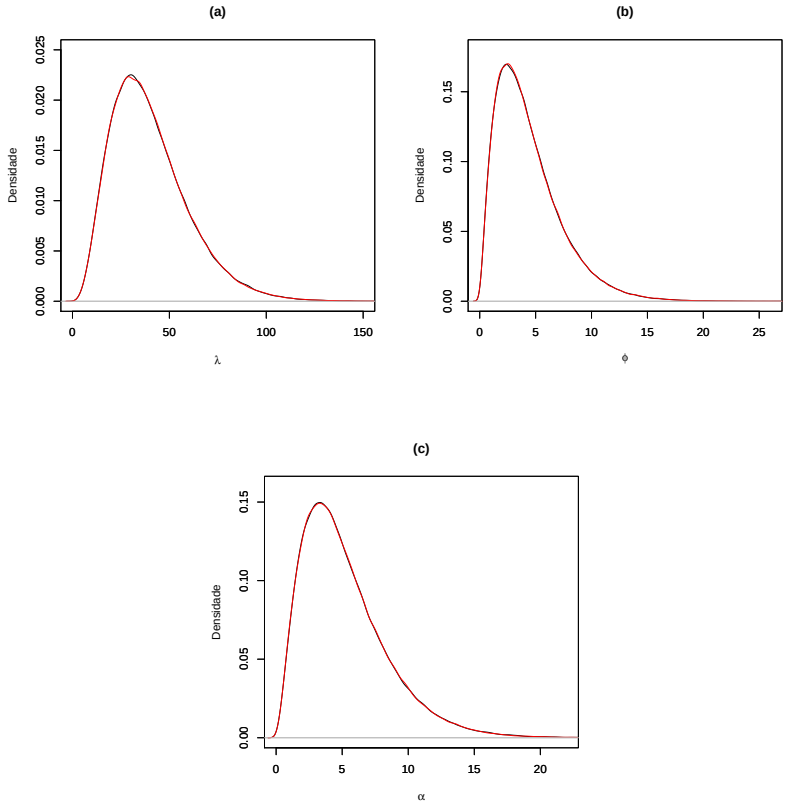

Figura 2.9: Gráficos de densidade do parâmetro λ (a), φ (b) e α (c).

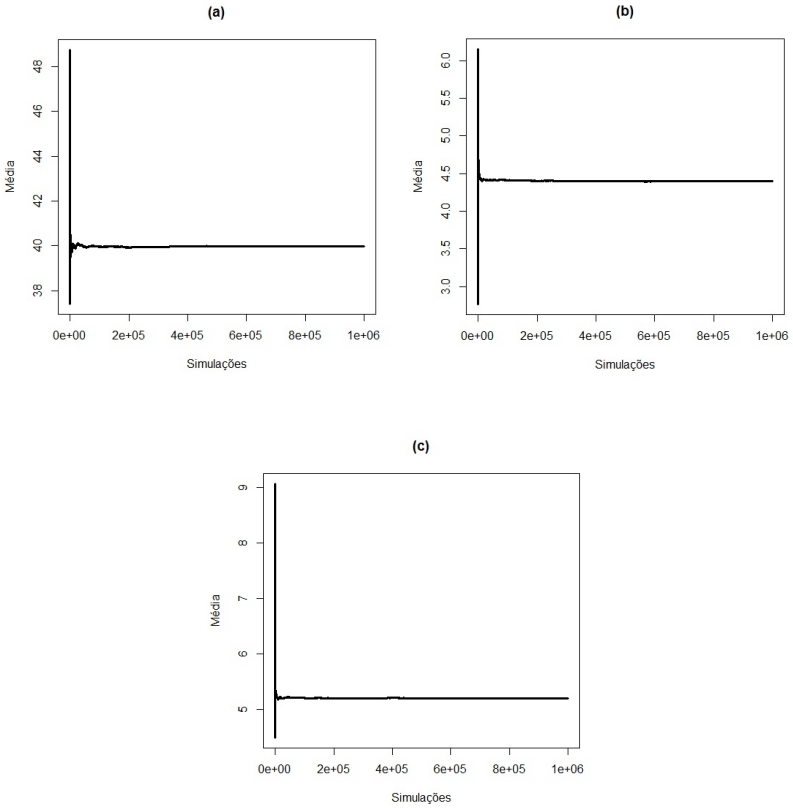

Figura 2.10: Gráfico das médias ergóticas dos parâmetro λ (a), φ (b) e α (c).

2.2 Considerações Finais

Neste capítulo apresentamos a distribuição Kumaraswamy generalizada. Vimos um caso dessa distribuição, a Kumaraswamy Weibull modificada e estudamos com mais detalhes um de seus casos particulares, a Kumaraswamy exponencial.

No estudo de simulação, observamos que embora tivemos um erro padrão um pouco alto, as propriedades frequentistas foram atendidas e que quanto maior a porcentagem de censura nos dados, precisa-se de mais dados para obtermos uma precisão maior. Também, para obtermos estimativas dentro do intervalo de variação dos parâmetros, foi necessário o uso de reparametrizações.

A aplicabilidade do modelo foi demonstrada em um conjunto de dados reais de um grupo de ratas com câncer vaginal, em que o modelo se mostrou adequado aos dados, tanto do ponto de vista clássico, quanto bayesiano. Neste exemplo, ambas abordagens resultaram em estimativas muito próximas, embora a abordagem bayesiana tenha tido erros padrões maiores.

Capítulo

3

Família Kumaraswamy Generalizada com Fração de Cura

Em análise de sobrevivência é esperado que todas unidades envolvidas no experimento falhem se acompanharmos o experimento por um longo período de tempo. No entanto, há situações em que uma parcela das unidades não apresentam o evento de interesse mesmo se acompanhadas por um longo tempo. Se acompanharmos uma lâmpada, certamente ela falhará, porém um ex-detento pode nunca apresentar recorrência no crime e dizemos que esses indivíduos são imunes, curados ou não suscetíveis ao evento de interesse e, consequentemente, sua população possui uma fração de curados. Os modelos tradicionais de sobrevivência não são capazes de captar a fração de cura, assim são necessários modelos estatísticos que incorporem a proporção de curados na população. Modelos para dados de sobrevivência com proporção de curados (também conhecidos como modelos de taxa de cura ou modelos de sobrevivência com longa duração) são de grande importância em análise de sobrevivência e em confiabilidade industrial.

Neste capítulo abordaremos o modelo unificado de longa duração para modelarmos dados de tempo de vida supondo que estes seguem distribuição Kumaraswamy exponencial.

3.1 Modelos Unificados de Fração de Cura

De uma forma geral a ideia básica do modelo unificado de fração de cura está baseada na noção de ocorrência do evento de interesse em um processo em dois estágios:

Estágio de iniciação. Seja N uma variável aleatória representando o número de causas ou riscos competitivos da ocorrência do evento de interesse. A causa de ocorrência do evento

é desconhecida, e a variável N não é observada, com distribuição de probabilidade p_n e sua cauda dadas, respectivamente, por

$$p_n = P[N = n] \quad \text{e} \quad q_n = P[N > n], \tag{3.1}$$

com $n = 0, 1, 2, ...$

Estágio de maturação. Dado $N = n$, sejam Z_k, $k = 1, ..., n$, variáveis aleatórias contínuas (não-negativas) independentes com função distribuição acumulada $F(z) = 1 - S(z)$ e independentes de N, que representam o tempo de ocorrência do evento de interesse devido à k-ésima causa. Afim de incluir os indivíduos que não são suscetíveis ao evento de interesse, seu tempo de ocorrência é definido como

$$T = \min(Z_0, Z_1, Z_2, ..., Z_N) \tag{3.2}$$

em que $P[Z_0 = \infty] = 1$, admitindo a possibilidade que uma proporção p_0 da população não apresenta a ocorrência do evento de interesse, T é uma variável aleatória observável ou censurada e Z_j e N são variáveis latentes.

Seja $\{a_n\}$ uma sequência de números reais. Se s pertence ao intervalo $[0, 1]$

$$A(s) = a_0 + a_1 s + a_2 s^2 + \cdots \tag{3.3}$$

converge, então $A(s)$ é definida como a função geradora da sequência $\{a_n\}$.

A função de sobrevivência da variável aleatória T (função de sobrevivência da população) será indicada por

$$\begin{aligned} S_{pop}(t) &= P[N = 0] + P[Z_1 > t, Z_2 > t, ..., Z_N > t, N \geq 1] \\ &= P[N = 0] + \sum_{n=1}^{\infty} P[N = n] P[Z_1 > t, Z_2 > t, ..., Z_N > t] \\ &= p_0 + \sum_{n=1}^{\infty} p_n S(t)^n \\ &= A(S(t)), \end{aligned} \tag{3.4}$$

sendo que $A(\cdot)$ é a função geradora da sequência p_n. Ou seja, a função de sobrevivência da variável aleatória T, correspondente a um modelo de longa duração em dois estágios, é uma composição da função geradora de probabilidades e a função de sobrevivência. A função de sobrevivência de longa duração em dois estágios $S_{pop}(t)$ não é uma função de sobrevivência própria.

Observe que, para a função de sobrevivência própria, $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S(t) = 0$. Já para a função de sobrevivência imprópria, $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S_{pop}(t) = P[N = 0] = p_0$. Dessa forma, p_0 é a proporção de não ocorrências do evento de interesse na população, ou seja, a fração de cura.

A função de sobrevivência populacional possui as seguintes propriedades:

- Se $p_0 = 1$, então $S_{pop}(t) = S(t)$;
- $S_{pop}(0) = 1$;
- $S_{pop}(t)$ é não crescente;
- $\lim_{t\to\infty} S_{pop}(t) = p_0$

As funções de densidade e de risco associadas à função de sobrevivência de longa duração são dadas, respectivamente, por

$$f_{pop}(t) = f(t)\frac{dA(s)}{ds}|_{s=S(t)} \tag{3.5}$$

e

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = f(t)\frac{\frac{dA(s)}{ds}|_{s=S(t)}}{S_{pop}(t)}. \tag{3.6}$$

Algumas distribuições são muito utilizadas para as funções geradoras de probabilidade, tais como Bernoulli, Binomial, Poisson, Binomial Negativa e Geométrica e contemplam diferentes modalidades de dispersão. Falaremos de cada uma delas a seguir.

- **Bernoulli**

Seja N uma variável aleatória que representa o número de causas competitivas latentes necessárias para a ocorrência de um determinado evento de interesse, que segue uma distribuição Bernoulli, com parâmetro θ. Sua função de probabilidade de massa dada por

$$P[N = n] = \theta^n(1-\theta)^{1-n},\ n = 0, 1, ...,\ \text{e}\ 0 < \theta < 1. \tag{3.7}$$

A função geradora de probabilidade de N é dada por

$$A(s) = 1 - \theta + \theta s,\ 0 \leq s \leq 1. \tag{3.8}$$

E a função de função de sobrevivência de longa duração é dada por

$$S_{pop}(t) = A(S(t)) = 1 - \theta + \theta S(t), \tag{3.9}$$

obtendo, assim, o modelo conhecido como modelo de mistura padrão (Berkson e Gage, 1952), que é um dos mais utilizados na análise de sobrevivência para ajustar dados de longa duração. Ele recebe esse nome, pois consiste de uma mistura de distribuições paramétricas, sendo uma função de sobrevivência própria para a parte da população formada pelos não curados e uma função para os curados, em que a proporção de curados é

$$p_0 = \lim_{t\to\infty} S_{pop}(t) = 1 - \theta. \tag{3.10}$$

A função de densidade e de risco são dadas respectivamente por

$$f_{pop}(t) = \theta f(t) \tag{3.11}$$

e

$$h_{pop}(t) = \frac{\theta f(t)}{1 - \theta + \theta S(t)}. \tag{3.12}$$

• Binomial

Seja o número N de riscos latentes com distribuição binomial. Uma motivação biológica é, supondo-se que existe um número K de potenciais lugares para a mutação de tumores focalizados numa região do corpo de um indivíduo afetada por uma doença, tem-se que N lugares chegam a sofrer mutações.

Adotamos uma reparametrização do modelo binomial: $\theta^* = \theta/(1+\theta)$. Assim, seja N uma variável aleatória que representa o número de causas competitivas latentes necessárias para a ocorrência de um determinado evento de interesse com distribuição binomial, sua função de probabilidade de massa é dada por

$$P[N = n^*] = \binom{K}{n^*} \theta^{*n^*} (1-\theta^*)^{K-n^*}, \;\; n^* = 0, 1, \ldots, K, \quad 0 < \theta^* < 1, \tag{3.13}$$

com $\mathrm{E}[N] = K\theta^*$ e $\mathrm{Var}[N] = K\theta(1-\theta^*)$, em que K é um número inteiro positivo. A função geradora de probabilidade de N é dada por

$$A(s) = (1 - \theta^* + \theta^* s)^K, \quad 0 \leq s \leq 1. \tag{3.14}$$

A função de função de sobrevivência de longa duração é dada por

$$S_{pop}(t) = A(S(t)) = S_{pop}(t) = \left[1 - \theta^* + \theta^* S(t)\right]^K, \tag{3.15}$$

sendo que a fração de cura p_0 é dada por

$$p_0 = \lim_{t \to \infty} S_{pop}(t) = (1 - \theta^*)^K > 0. \tag{3.16}$$

As funções de densidade e de risco associadas à função de sobrevivência de longa duração binomial são dadas respectivamente por

$$f_{pop}(t) = K\theta^* f(t)(1 - \theta^* + \theta^* S(t))^{(K-1)} \;\text{ e }\; h_{pop}(t) = \frac{K\theta^* f(t)}{\{1 - \theta^* + \theta^* S(t)\}}, \tag{3.17}$$

em que $f(t) = -S'(t)$ é uma função de densidade própria.

Observa-se que, no caso em que K cresce a fração de cura p_0 decresce.

• **Poisson**

Seja o número N de causas do evento de interesse com distribuição de probabilidade de Poisson. Assim, considerando a função geradora de probabilidade da distribuição de Poisson $A(s) = \exp[\theta(1-s)]$, obtemos as funções de sobrevivência, densidade e de risco da população, dadas respectivamente por

$$S_{pop}(t) = A(S(t)) = \exp\left[-\theta F(t)\right], \tag{3.18}$$

$$f_{pop}(t) = -\frac{dS_{pop}(t)}{dt} = \theta f(t) \exp\left[-\theta F(t)\right] \tag{3.19}$$

e

$$h_{pop}(t) = \theta f(t). \tag{3.20}$$

Dessa forma, de (3.18) tem-se a fração de cura dada por $p_0 = \lim_{t\to\infty} S_{pop}(t) = \exp(-\theta)$.

• **Geométrica**

Seja o número de causas competitivas N seguindo uma distribuição geométrica, então temos que N tem função de probabilidade definida por

$$P(N = n) = \left(\frac{\theta}{1+\theta}\right)^n (1+\theta)^{-1},$$

$n = 0, 1, \ldots, 0 < \theta < 1$.

A função geradora de probabilidades é dada por

$$A(s) = \sum_{n=0}^{\infty} p_n s^n = \{1 + \theta(1-s)\}^{-1}, \quad 0 \leq s \leq 1, \tag{3.21}$$

de tal forma que a função de sobrevivência de longa duração para o modelo geométrico será dada por

$$S_{pop}(t) = \{1 + \theta F(t)\}^{-1}, \tag{3.22}$$

tendo assim uma fração de cura de $p_0 = 1/(1+\theta)$.

Dessa forma, a função de densidade e de risco da população são dadas respectivamente por

$$f_{pop}(t) = -\frac{dS_{pop}(t)}{dt} = \theta f(t)\{1 + \theta F(t)\}^{-2} \tag{3.23}$$

e

$$h_{pop}(t) = \frac{\theta f(t)\{1 + \theta F(t)\}^{-2}}{\{1 + \theta F(t)\}^{-1}}. \tag{3.24}$$

• **Binomial Negativa**

neste caso, temos o número de causas competitivas N seguindo uma distribuição binomial negativa, então N tem função de probabilidade definida por

$$P(N=n)=\frac{\Gamma(n+\eta^{-1})}{n!\Gamma(\eta^{-1})}\left(\frac{\eta\theta}{1+\eta\theta}\right)^{n}(1+\eta\theta)^{-1/\eta},$$

$n=0,1,\ldots,\theta>0$, $\eta\geq-1$ e $1+\eta\theta>0$, e então $\mathrm{E}(N)=\theta$ e $\mathrm{Var}(N)=\theta+\eta\theta^2$.

A função geradora de probabilidades é dada por

$$A(s)=\sum_{n=0}^{\infty}p_n s^n=\{1+\eta\theta(1-s)\}^{-1/\eta},\quad 0\leq s\leq 1. \tag{3.25}$$

Dessa forma, a função de sobrevivência de longa duração para o modelo binomial negativo é dada por

$$S_{pop}(t)=\{1+\eta\theta F(t)\}^{-1/\eta}, \tag{3.26}$$

e temos que a fração de curados na população é

$$p_0=\lim_{t\to\infty}S_{pop}(t)=(1+\eta\theta)^{-1/\eta}. \tag{3.27}$$

A função densidade do modelo (3.26) é

$$f_{pop}(t)=-\frac{dS_{pop}(t)}{dt}=\theta f(t)\{1+\eta\theta F(t)\}^{-1-1/\eta}, \tag{3.28}$$

em que $f(t)=-S'(t)$. Além disso, a função de risco correspondente é dada por

$$h_{pop}(t)=\theta f(t)\{1+\eta\theta F(t)\}^{-1}. \tag{3.29}$$

Observamos alguns casos particulares deste modelo: de 3.25, quando $\eta\to 0$, obtemos a função de densidade da distribuição de Poisson, se $\eta=-1$, caímos na distribuição de Bernoulli, se $\eta=1$ temos a distribuição geométrica. Observamos ainda, das expressões da esperança e variância do modelo, que a variância do número de causas competindo é bem flexível. Se $-1/\theta<\eta<0$, há uma subdispersão em relação à distribuição de Poisson, enquanto que $\eta>0$ há uma sobredispersão.

Na Tabela 3.1 são apresentados a função de sobrevivência de longa duração, a densidade imprópria e fração de cura correspondentes aos modelos descritos anteriormente.

Como a distribuição binomial negativa é mais geral, a seguir iremos apresentar o modelo de longa duração com a Kum-G para este caso.

Tabela 3.1: Função de sobrevivência $S_{pop}(t)$, função de densidade $f_{pop}(t)$ e fração de cura para diferentes distribuições do número de causas latentes, N.

Distribuição	$S_{pop}(t)$	$f_{pop}(t)$	p_0
Bernoulli(θ)	$1-\theta+\theta S(t)$	$\theta f(t)$	$1-\theta$
Binomial(K,θ^*)	$(1-\theta^*+\theta^* S(t))^K$	$K\theta^* f(t)(1-\theta^*+\theta^* S(t))^{K-1}$	$(1-\theta^*)^K$
Poisson(θ)	$\exp(-\theta F(t))$	$\theta f(t)\exp(-\theta F(t))$	$e^{-\theta}$
Geométrica(θ)	$\{1+\theta F(t)\}^{-1}$	$\theta f(t)\{1+\theta F(t)\}^{-2}$	$1/(1+\theta)$
Binomial Negativa(τ,θ)	$\{1+\eta\theta F(t)\}^{-1/\eta}$	$\theta f(t)\{1+\eta\theta F(t)\}^{-1-1/\eta}$	$(1+\eta\theta)^{-1/\eta}$

3.2 Família Kum-G Binomial Negativa de Longa Duração

Considerando a distribuição Binomial Negativa para o número de causas competitivas e o tempo seguindo a distribuição Kum-G, iremos obter uma família de distribuições de longa duração, em que a função de sobrevivência populacional do modelo é dada por

$$S_{pop}(t) = \{1+\eta\theta F(t)\}^{-1/\eta} = \{1+\eta\theta[1-\left[1-G(t)^\lambda\right]^\varphi]\}^{-1/\eta}, \tag{3.30}$$

com fração de curados na população dada por

$$p_0 = (1+\eta\theta)^{-1/\eta}. \tag{3.31}$$

Assim, basta substituir a função $G(t)$ pela função distribuição acumulada de alguma distribuição e teremos um modelo de Kum-G binomial negativa de longa duração.

A função de densidade populacional é

$$f_{pop}(t) = \theta\lambda\varphi g(t)G(t)^{\lambda-1}\left[1-G(t)^\lambda\right]^{\varphi-1}\{1+\eta\theta[1-\left[1-G(t)^\lambda\right]^\varphi]\}^{-1-1/\eta} \tag{3.32}$$

e a função de risco populacional é dada por

$$h_{pop}(t) = \theta\lambda\varphi g(t)G(t)^{\lambda-1}\left[1-G(t)^\lambda\right]^{\varphi-1}\{1+\eta\theta[1-\left[1-G(t)^\lambda\right]^\varphi]\}^{-1}. \tag{3.33}$$

Na Tabela 3.2 mostramos os casos particulares deste modelo. Vale ressaltar que para cada $G(t)$, teremos diferentes distribuições.

A seguir, trataremos do caso em que $G(t)$ tem distribuição exponencial.

Tabela 3.2: Função de sobrevivência $S_{pop}(t)$, função de densidade $f_{pop}(t)$ e fração de cura para diferentes distribuições do número de causas latentes, N.

Parametrização	Modelo	$S_{pop}(t)$
$\eta \to 0$	Poisson	$\exp\{-\theta[1-[1-G(t)^{\lambda}]^{\varphi}]\}$
$\eta = -1$	Bernoulli	$1-\theta+\theta[1-G(t)^{\lambda}]^{\varphi}$
$\eta = 1$	geométrica	$\{1+\theta[1-[1-G(t)^{\lambda}]^{\varphi}]\}^{-1}$

3.2.1 Família Kum-Exp Binomial Negativa de Longa Duração

Considerando $G(t)$ seguindo uma distribuição exponencial e substituindo em 3.30, temos a família Kum-Exp binomial negativa de longa duração, em que sua função de sobrevivência populacional é dada por

$$S_{pop}(t) = \{1+\eta\theta[1-\left[1-(1-e^{-\alpha t})^{\lambda}\right]^{\varphi}]\}^{-1/\eta}. \tag{3.34}$$

A função de densidade e a função de risco populacional desse modelo são, respectivamente,

$$f_{pop}(t) = \theta\varphi\lambda\alpha \mathrm{e}^{-\alpha t}(1-\mathrm{e}^{-\alpha t})^{\lambda-1}[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi-1}\{1+\eta\theta[1-[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}]\}^{-1-1/\eta}, \tag{3.35}$$

e

$$h_{pop}(t) = \theta\varphi\lambda\alpha \mathrm{e}^{-\alpha t}(1-\mathrm{e}^{-\alpha t})^{\lambda-1}[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi-1}\{1+\eta\theta[1-[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}]\}^{-1}. \tag{3.36}$$

Para realizar a estimação dos parâmetros, usaremos novamente o método de máxima verossimilhança. Considerando dados de sobrevivência, a partir dos tempos observados $\boldsymbol{t} = (t_1, t_2, ..., t_n)$ e do vetor de parâmetros $\boldsymbol{\mu} = (\varphi, \lambda, \alpha, \theta)$, a função de verossimilhança do modelo é dada por

$$\begin{aligned} L(\boldsymbol{\mu}) &= \prod_{i=1}^{n}\left[\theta\varphi\lambda\alpha \mathrm{e}^{-\alpha t}(1-\mathrm{e}^{-\alpha t})^{\lambda-1}[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi-1}\{1+\eta\theta[1-[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}]\}^{-1}\right]^{\delta_i} \\ &\quad \{1+\eta\theta[1-\left[1-(1-e^{-\alpha t})^{\lambda}\right]^{\varphi}]\}^{-1/\eta} \end{aligned} \tag{3.37}$$

na qual δ_i é o indicador de falha (1 indica falha e 0 indica censura).

O logaritmo da função de verossimilhança é

$$\begin{aligned} \ell(\boldsymbol{\mu}) &= \sum_{i=1}^{n}\delta_i \log\left[\theta\varphi\lambda\alpha \mathrm{e}^{-\alpha t}(1-\mathrm{e}^{-\alpha t})^{\lambda-1}[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi-1}\{1+\eta\theta[1-[1-(1-\mathrm{e}^{-\alpha t})^{\lambda}]^{\varphi}]\}^{-1}\right] \\ &\quad + \sum_{i=1}^{n}\left(\frac{-1}{\eta}\right)\log\{1+\eta\theta[1-\left[1-(1-e^{-\alpha t})^{\lambda}\right]^{\varphi}]\}. \end{aligned} \tag{3.38}$$

Não é possível encontrar as estimativas analiticamente, então é necessário o uso de um método numérico.

3.2.1.1 Aplicação

Analisamos um conjunto de dados de casais dos EUA em que o evento de interesse é o divórcio, adaptado de um exemplo encontrado em (Lillard e Panis, 2000) e bastante conhecido na literatura. Neste caso, um casal pode vir a nunca se divorciar, então uma parcela desta população tem uma fração de curados (os que não se divorciam). Nele há 3371 casais sendo que há a presença de 2339 censuras. O tempo máximo observado foi de 73,07 anos. Na Figura 3.1 temos o estimador de Kaplan-Meier baseado nos dados.

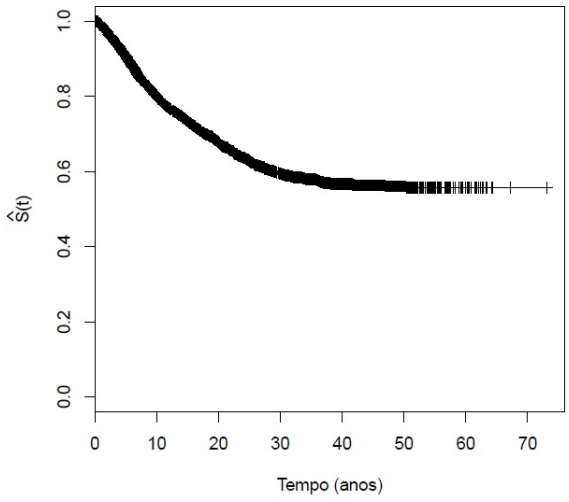

Figura 3.1: Curva de Kaplan-Meier do conjunto de dados.

Há vários indicativos de que estamos em um caso de longa duração: a curva de Kaplan-Meier que se estabiliza em um valor acima de zero, o tempo grande de estudo com muitas censuras além de um conhecimento prévio que temos que um casal pode vir a nunca se divorciar.

Na Figura 3.2 temos o TTT plot com censuras do conjunto de dados. Neste caso temos um indicativo de que a função de risco é crescente, nos indicando que a distribuição Kum-Exp é uma das distribuições adequadas aos dados.

Assim vamos aplicar o modelo Kum-Exp binomial negativa de longa duração e seus casos particulares para podermos modelar a função de sobrevivência dos tempos.

Os resultados das estimativas de máxima verossimilhança dos parâmetros de cada um dos modelos da família Kum-Exp binomial negativa de longa duração são apresentados na Tabela 3.3, bem como os desvios padrões e intervalo de confiança correspondentes.

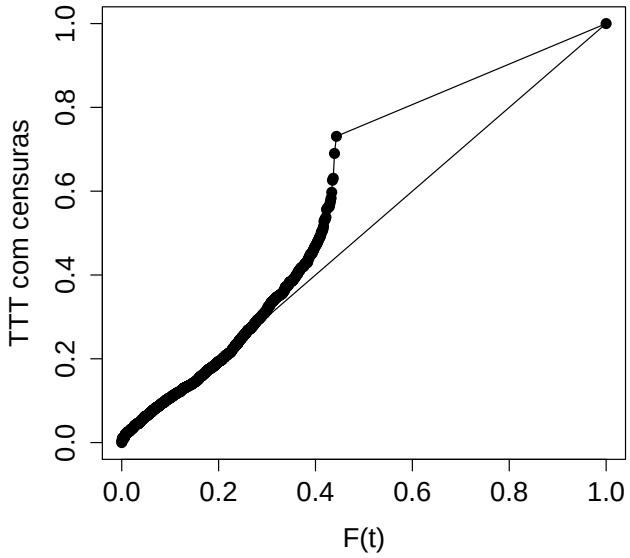

Figura 3.2: TTT plot com censuras dos dados.

Tabela 3.3: Estimativas dos parâmetros dos modelos de longa duração

Modelo	Parâmetros	EMV	Desvio padrão	IC 95%	p_0
Binomial Negativa	α	$0,07$	$1,04$	$(0,01;0,52)$	$0,54$
	λ	$1,55$	$0,07$	$(1,34;1,80)$	
	φ	$0,84$	$1,14$	$(0,09;7,94)$	
	θ	$0,85$	$0,45$	$(0,35;2,05)$	
	η	$0,98$	$1,27$	$(-1,51;3,47)$	
Bernoulli	α	$0,31$	$0,54$	$(0,11;0,91)$	$0,54$
	λ	$1,79$	$0,13$	$(1,38;2,34)$	
	φ	$0,22$	$0,61$	$(0,07;0,74)$	
	θ	$0,46$	$0,03$	$(0,43;0,49)$	
Geométrica	α	$0,04$	$1,06$	$(0,0;0,3)$	$0,53$
	λ	$1,48$	$0,06$	$(1,31;1,68)$	
	φ	$1,74$	$1,38$	$(0,12;25,83)$	
	θ	$0,87$	$0,07$	$(0,75;1,00)$	
Poisson	α	$0,11$	$0,85$	$(0,02;0,58)$	$0,54$
	λ	$1,58$	$0,11$	$(1,27;1,96)$	
	φ	$0,59$	$1,02$	$(0,08;4,36$	
	θ	$0,61$	$0,05$	$(0,55;0,67)$	

Na Tabela 3.4 temos os valores dos critérios AIC, BIC e a log-verossimilhança de cada modelo.

Tabela 3.4: Critérios AIC, BIC e Log-verossimilhança dos modelos

Modelo	AIC	BIC	Log-verossimilhança
Binomial Negativa	$10298,66$	$10329,27$	$-5144,33$
Bernoulli	$10293,34$	$10317,83$	$-5142,67$
Geométrica	$10296,26$	$10320,75$	$-5144,13$
Poisson	$10295,26$	$10319,76$	$-5143,63$

Dessa forma, o modelo selecionado é o modelo Kum-Exp Bernoulli de longa duração, embora não tenha muita diferença em relação aos outros modelos.

A Figura 3.3 apresenta o gráfico do ajuste da curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelo modelo. Observamos que a curva de sobrevivência estimada pelo modelo está próxima da curva de Kaplan-Meier, confirmando que o modelo é adequado para os dados. Também vemos que a curva se estabiliza em aproximadamente $0,55$, muito próximo à proporção de cura estimada da população pelo modelo.

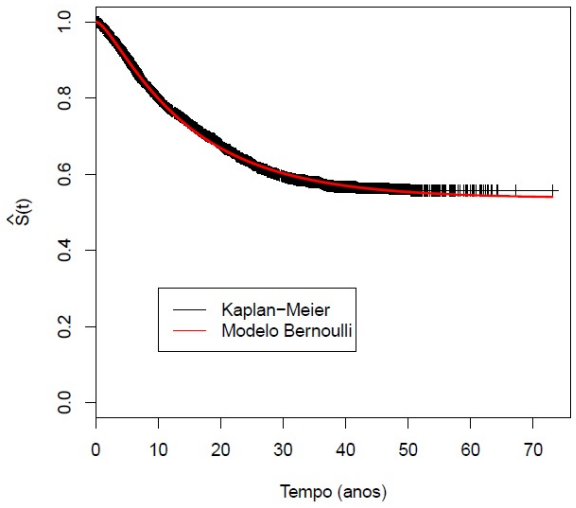

Figura 3.3: Curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelo modelo.

3.3 Considerações Finais

Neste capítulo apresentamos a teoria de modelos unificados de longa duração, estudamos o caso que o número de causas competitivas segue uma distribuição binomial negativa, que é bem geral e contém três casos particulares, supondo que o tempo de vida dos indivíduos em risco seguem distribuição Kumaraswamy generalizada, e vimos um caso na distribuição Kum-Exp.

A vantagem desse modelo reside na capacidade de modelar dados que contenham uma fração de curados na população e tempos de vida dos indivíduos em risco com diversas formas, já que a distribuição Kum-Exp é bem versátil. O modelo proposto abrange alguns modelos especiais, sendo importante o uso de critérios de seleção de modelo para verificar o que se encaixa melhor à aplicação.

A aplicabilidade do modelo foi demonstrada em um conjunto de dados reais de casais que se divorciam, em que o modelo Kum-Exp Bernoulli de longa duração se ajustou melhor aos dados, embora os outros também tenham se ajustado muito bem. Podemos notar que neste exemplo vários modelos podem ser aplicados, visto que o risco é monótono.

Capítulo

4

Família Kum-G de Longa Duração na Presença de Covariáveis

Os estudos em análise de sobrevivência muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Por exemplo, se estamos interessados no tempo até a morte de um indivíduo devido ao câncer, vários fatores podem influenciar nesse tempo, como a idade do paciente, estado psicológico, sexo, peso, entre outras. Dessa forma, modelos de sobrevivência que incluem covariáveis se tornam imprescindíveis em muitos estudos.

Consideramos o modelo de fração de cura Bernoulli, também chamado modelo de mistura padrão, para que possamos analisar dados de tempo de vida com censura (como é feito em modelos usuais de sobrevivência), proporção de curados (como inserido através dos modelos de fração de cura) e também levando em conta o efeito de covariáveis. Taconeli (2013) estudou a modelagem de covariáveis que veremos a seguir com os tempos seguindo uma distribuição Weibull e neste trabalho supomos que os tempos seguem a distribuição Kum-Exp, por esta ser mais flexível.

Uma vantagem do modelo de mistura padrão é sua identificabilidade no caso em que a função de sobrevivência S(t) é especificada através de uma distribuição de probabilidade. Li *et al.* (2001) mostraram que o modelo de mistura padrão é identificável quando a função de sobrevivência própria é especificada parametricamente e a fração de curados é um parâmetro (e que também pode ser função de covariáveis).

Seja $\boldsymbol{Z}$ representando o conjunto de covariáveis que afetam a probabilidade de cura e $\boldsymbol{X}$ o grupo de covariáveis que afetam a função de sobrevivência condicional. O modelo pode ser

reescrito como

$$S_{pop}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{Z}) = 1 - \theta(\boldsymbol{Z}) + \theta(\boldsymbol{Z})S(\boldsymbol{t}|\boldsymbol{X}) \tag{4.1}$$

em que $\theta(\boldsymbol{Z})$ é a probabilidade de um paciente ser não curado dependendo de $\boldsymbol{Z}$ e $S(\boldsymbol{t}|\boldsymbol{X})$ é a função de sobrevivência da distribuição do tempo de falha de pacientes não curados dependendo $\boldsymbol{X}$. Vale ressaltar que em alguns casos práticos as covariáveis $\boldsymbol{Z}$ e $\boldsymbol{X}$ podem ser as mesmas e que também podem ter casos em que não há covariáveis influenciando na proporção de cura ou nos indivíduos em risco.

Considerando o modelo de riscos proporcionais e que o tempo dos indivíduos em risco segue a distribuição Kum-G, podemos reescrever 4.1 da seguinte forma

$$S_{pop}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{Z}) = 1 - \theta(\boldsymbol{Z}) + \theta(\boldsymbol{Z}) \left\{ \left[1 - G(\boldsymbol{t})^{\lambda}\right]^{\varphi} \right\}^{\mathrm{e}^{\boldsymbol{\beta X}}}, \tag{4.2}$$

em que $\boldsymbol{\beta}$ representa o vetor de parâmetros que serão estimados para as covariáveis associadas aos não imunes. A função densidade de probabilidade deste modelo será

$$f_{pop}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{Z}) = \theta(\boldsymbol{Z})\lambda\varphi g(\boldsymbol{t})G(\boldsymbol{t})^{\lambda-1}\left[1 - G(\boldsymbol{t})^{\lambda}\right]^{\varphi-1}\mathrm{e}^{\boldsymbol{\beta X}}\left\{\left[1 - G(\boldsymbol{t})^{\lambda}\right]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}-1}. \tag{4.3}$$

Para modelar os efeitos das covariáveis na taxa de cura, podemos utilizar diferentes funções de ligações. Definindo $\boldsymbol{b}$ como sendo o vetor de parâmetros que serão estimados para as covariáveis associadas à fração de cura e a função de ligação logito temos o modelo de regressão logístico dado por

$$\theta(\boldsymbol{Z}) = \frac{\mathrm{e}^{\boldsymbol{bZ}}}{1 + \mathrm{e}^{\boldsymbol{bZ}}}, \tag{4.4}$$

a função de ligação probit tem o seguinte modelo de regressão

$$\theta(\boldsymbol{Z}) = \Phi(\boldsymbol{bZ}),$$

em que Φ corresponde à função de distribuição acumulada de uma distribuição normal padrão e a função de ligação complementar log-log

$$\theta(\boldsymbol{Z}) = \mathrm{e}^{-\mathrm{e}^{\boldsymbol{bZ}}}$$

em que o processo de estimação empregada é similar para as três funções de ligação.

Em Taconeli (2013) podemos ver que utilizando as três funções de ligação diferentes, chegou-se a resultados muito parecidos. Desta forma, neste trabalho consideramos apenas a função de ligação logística.

4.1 Kumaraswamy Exponencial Bernoulli na Presença de Covariáveis

Se considerarmos que os tempos seguem uma distribuição Kum-Exp, o modelo de regressão Kum-Exp Bernoulli será

$$S_{pop}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{Z}) = 1-\theta(\boldsymbol{Z})+\theta(\boldsymbol{Z})\left\{\left[1-(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda}\right]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}}, \tag{4.5}$$

em que $\theta(\boldsymbol{Z}) = \frac{\mathrm{e}^{\boldsymbol{bZ}}}{1+\mathrm{e}^{\boldsymbol{bZ}}}$.

Dessa forma, a proporção de curados na população será

$$p_0 = 1-\theta(\boldsymbol{Z}) = 1-\frac{\mathrm{e}^{\boldsymbol{bZ}}}{1+\mathrm{e}^{\boldsymbol{bZ}}}. \tag{4.6}$$

A função densidade é dada por

$$\begin{aligned} f_{pop}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{Z}) &= \theta(\boldsymbol{Z})\lambda\varphi\alpha\mathrm{e}^{-\alpha\boldsymbol{t}}(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda-1}\left[1-(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda}\right]^{\varphi-1}\mathrm{e}^{\boldsymbol{\beta X}} \\ &\quad \left\{\left[1-(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda}\right]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}-1} \end{aligned} \tag{4.7}$$

e a função de risco

$$h_{pop}(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{Z}) = \frac{\theta(\boldsymbol{Z})\lambda\varphi\alpha\mathrm{e}^{-\alpha\boldsymbol{t}}(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda-1}\left[1-(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda}\right]^{\varphi-1}\mathrm{e}^{\boldsymbol{\beta X}}\left\{\left[1-(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda}\right]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}-1}}{1-\theta(\boldsymbol{Z})+\theta(\boldsymbol{Z})\left\{[1-(1-\mathrm{e}^{-\alpha\boldsymbol{t}})^{\lambda}]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}}}. \tag{4.8}$$

Para a estimação dos parâmetros do modelo 4.5 consideramos a função de verossimilhança, dada por

$$\begin{aligned} L(\boldsymbol{\eta}) &= \prod_{i=1}^{n}\left[h_{pop}(t_i|\boldsymbol{X},\boldsymbol{Z})\right]^{\delta_i} S_{pop}(t_i|\boldsymbol{X},\boldsymbol{Z}) \\ &= \prod_{i=1}^{n}\left[\frac{\theta(\boldsymbol{Z})\lambda\varphi\alpha\mathrm{e}^{-\alpha t_i}(1-\mathrm{e}^{-\alpha t_i})^{\lambda-1}\left[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}\right]^{\varphi-1}\mathrm{e}^{\boldsymbol{\beta X}}\left\{\left[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}\right]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta x}}-1}}{1-\theta(\boldsymbol{Z})+\theta(\boldsymbol{Z})\left\{[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}}}\right]^{\delta_i} \\ &\quad \left[1-\theta(\boldsymbol{Z})+\theta(\boldsymbol{Z})\left\{\left[1-(1-\mathrm{e}^{-\alpha t_i})^{\lambda}\right]^{\varphi}\right\}^{\mathrm{e}^{\boldsymbol{\beta X}}}\right], \end{aligned} \tag{4.9}$$

em que $\boldsymbol{t} = (t_1, t_2, ..., t_n)$ são os tempos observados, δ_i é a variável indicadora de censura e $\boldsymbol{\eta} = (\varphi, \lambda, \alpha, \boldsymbol{\beta}, \boldsymbol{b})$ é o vetor com todos os parâmetros do modelo.

4.2 Aplicação

Nesta aplicação reutilizaremos o conjunto de dados sobre divórcio utilizado no capítulo anterior, pois se trata de um caso de longa duração, mas agora utilizaremos covariáveis afim de verificarmos quais fatores influencia na função de sobrevivência e na fração de curados.

O conjunto de dados tem três variáveis:

Educação: educação do marido,

$$\begin{cases} 0 = \text{ menos de 12 anos,} \\ 1 = \text{ 12 a 15 anos,} \\ 2 = \text{ 16 ou mais anos,} \end{cases}$$

Etnia: etnia do marido,

$$\begin{cases} 1 = \text{ se o marido é negro,} \\ 0 = \text{ caso contrário.} \end{cases}$$

e **Misturado:** etnia do casal,

$$\begin{cases} 1 = \text{ se o marido e a mulher têm diferentes etnias (definidas como negro ou outro),} \\ 0 = \text{ caso contrário.} \end{cases}$$

Analisamos o conjunto de dados considerando o modelo com covariáveis tanto na fração de cura quanto na função de sobrevivência dos não curados. Afim de testar quais covariáveis são significativas no modelo, consideramos todas as covariáveis na função de sobrevivência e na fração de cura. Através da análise dos intervalos de confiança assintóticos, retiramos uma a uma as covariáveis não significativas.

Por fim, o modelo ficou com apenas a covariável *misturado* na fração de cura e nenhuma covariável foi significativa na função de sobrevivência dos não curados.

Desta forma, cinco parâmetros foram estimados: os três referentes à distribuição de probabilidade da função de sobrevivência dos não curados, Kum-exp (λ, φ e α) e $b = (b_0, b_1)$, que é o conjunto de parâmetros utilizados na regressão relativa à probabilidade de cura (não divorciados), em que b_0 representa o intercepto e b_1 está associado à covariável *misturado*. Exibimos na Tabela 4.1 as estimativas de máxima verossimilhança dos parâmetros do modelo ajustado e seus respectivos desvios-padrão e intervalos de confiança assintótico.

Tabela 4.1: Estimativas de máxima verossimilhança e seus respectivos desvios-padrão e intervalos de confiança assintótico do modelo

Parâmetro	Estimativa	Desvio padrão	IC95%
α	$0,33$	$0,57$	$(0,11; 1,02)$
λ	$1,85$	$0,15$	$(1,37; 2,50)$
φ	$0,21$	$0,64$	$(0,06; 0,73)$
b_0	$-0,22$	$0,06$	$(-0,35; -0,10)$
b_1	$0,41$	$0,12$	$(0,18; 0,65)$

Portanto, a função de sobrevivência estimada é:

$$S_{pop}(\text{anos}|\text{misturado}) = 1 - \theta(\boldsymbol{Z}) + \theta(\boldsymbol{Z}) \left\{ \left[1 - (1 - \mathrm{e}^{-0,33\text{anos}})^{1,85} \right]^{0,21} \right\}, \qquad (4.10)$$

em que

$$\theta(\boldsymbol{Z}) = \frac{\mathrm{e}^{-0,22+0,41\text{misturado}}}{1+\mathrm{e}^{-0,22+0,41\text{misturado}}}. \tag{4.11}$$

Assim, a proporção de cura estimada pelo modelo é de 0,55 se a covariável *misturado* valer 0 e 0,45 se o valor da covariável for 1, ou seja, 55% dos casais não se divorciam quando o casal tem a mesma etnia e 45% dos casais de etnia diferente não se divorciam.

Na Figura 4.1, são comparadas a função de sobrevivência gerada pelo modelo final com a curva de Kaplan-Meier, pelos níveis da covariável *misturado*.

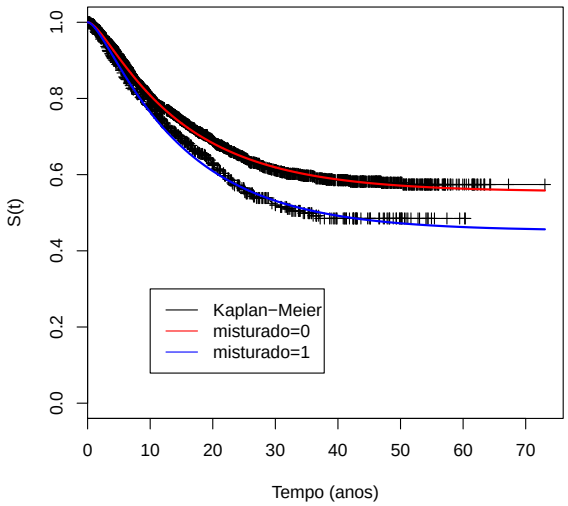

Figura 4.1: Curva de Kaplan-Meier juntamente com a curva de sobrevivência estimada pelo modelo por nível de covariável.

Observamos que em ambas as combinações dos níveis da covariável, o modelo se ajustou bem aos dados, pois a curva de sobrevivência estimada pelo modelo se aproximou razoavelmente bem da curva de Kaplan-Meier. Também, confirmamos que os casais com maior probabilidade de não divorciar são os com a mesma etnia.

4.3 Considerações Finais

Neste capítulo apresentamos um modelo de longa duração na presença de covariáveis utilizando a distribuição Kum-Exp. Este modelo é bem completo, já que considera que uma parcela da população está imune ao evento de interesse e ainda identifica fatores que influenciam na proporção de curados e na função de sobrevivência.

Demonstramos a aplicabilidade deste modelo utilizando um conjunto de dados de divórcio e verificamos que a covariável que indica se o casal é de etnia diferente influencia na proporção de cura.

Capítulo

5

Conclusões e Propostas de Trabalhos Futuros

Neste trabalho fizemos uma breve revisão em análise de sobrevivência e mostramos a distribuição Kumaraswamy, juntamente com suas propriedades. Em seguida, apresentamos a distribuição Kumaraswamy generalizada e estudamos dois de seus casos particulares, a Kumaraswamy Weibull modificada e, com mais detalhes, a Kumaraswamy exponencial. Também, utilizamos a metodologia de Análise de Sobrevivência para modelar dados seguindo estas distribuições e mostramos duas formas de estimação de parâmetros, por máxima verossimilhança e por uma abordagem bayesiana.

Utilizamos o modelo unificado de fração de cura para dados de longa duração com a distribuição Kum-Exp com a distribuição do número de causas competitivas seguindo a distribuição binomial negativa. Este modelo apresenta três casos particulares, sendo um modelo a ser usado em diversos casos. Quando não temos indivíduos curados no estudo, a função de sobrevivência acaba sendo a mesma do modelo Kum-Exp.

Também consideramos covariáveis no modelo de longa duração Bernoulli, afim de se identificar fatores que influenciam na função de sobrevivência e na proporção de curados. Temos assim um modelo que se encaixa em muitos casos práticos.

Através de conjuntos de dados reais, foi possível verificar que todos os modelos estudados neste trabalho são de fato aplicáveis, já que verificamos que eles foram adequados aos dados, fazendo uma comparação entre a curva de sobrevivência estimada pelos modelos e a curva de Kaplan-Meier.

Para trabalhos futuros, propomos considerar outros casos da distribuição Kum-G no modelo de longa duração e no modelo de longa duração com covariáveis, como a distribuição Kumaraswamy Weibull, Kumaraswamy Weibull modificada, Kumaraswamy log normal, entre outras. Propomos também utilizar o teste da razão de verossimilhança para fazer a seleção de

covariáveis e para testar parâmetros. Além disso, sugerimos fazer estimação bayesiana no modelo de longa duração e no modelo de longa duração com covariáveis, e uma opção é considerar prioris com distribuição uniforme.

Referências Bibliográficas

AALEN, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

AARSET, M. V. The null distribution for a test of constant vs "bathtub"failure rate. *Scand.Journal of Stat.*, pages 55–61, 1985.

ANSCOMBE, F. J. Estimating a mixed-exponential response law. *J. Amer. Statist. Assoc.*, v. 56, p. 493–502, 1961. ISSN 0162-1459.

BARLOW, R. E.; CAMPO, R. A. Total time on test processes and applications to failure data analysis. 1975.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v. 47(259), p. pp. 501–515, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2281318.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 11(1), p. pp. 15–53, 1949. ISSN 00359246. URL http://www.jstor.org/stable/2983694.

BOURGUIGNON, M.; SILVA, R. B.; ZEA, L. M.; CORDEIRO, G. M. The Kumaraswamy Pareto distribution. *J. Stat. Theory Appl.*, v. 12(2), p. 129–144, 2013. ISSN 1538-7887. doi: 10.2991/jsta.2013.12.2.1. URL http://dx.doi.org/10.2991/jsta.2013.12.2.1.

BROADHURST, R.; MALLER, R. Estimating the numbers of prison terms in criminal careers from one-step probabilities of recidivism. *Journal of Quantitative Criminology*, v. 7(3), p. 275–290, 1991. ISSN 0748-4518. doi: 10.1007/BF01063234. URL http://dx.doi.org/10.1007/BF01063234.

CASELLA, G.; BERGER, R. L. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. A new Bayesian model for survival data with a surviving fraction. *J. Amer. Statist. Assoc.*, v. 94(447), p. 909–919, 1999. ISSN 0162-1459. doi: 10.2307/2670006. URL http://dx.doi.org/10.2307/2670006.

COLOSIMO, E.; GIOLO, S. *Análise de sobrevivência aplicada*. ABE - Projeto Fisher. Edgard Blücher, 2006. ISBN 9788521203841. URL http://books.google.com.br/books?id=g0-uOgAACAAJ.

CORDEIRO, G. M.; DE CASTRO, M. A new family of generalized distributions. *J. Stat. Comput. Simul.*, v. 81(7), p. 883–898, 2011. ISSN 0094-9655. doi: 10.1080/00949650903530745. URL http://dx.doi.org/10.1080/00949650903530745.

CORDEIRO, G. M.; PESCIM, R. R.; ORTEGA, E. M. M. The Kumaraswamy generalized half-normal distribution for skewed positive data. *J. Data Sci.*, v. 10(2), p. 195–224, 2012. ISSN 1680-743X.

CORDEIRO, G. M.; ORTEGA, E. M.; SILVA, G. O. The kumaraswamy modified weibull distribution: theory and applications. *Journal of Statistical Computation and Simulation*, v. 84(7), p. 1387–1411, 2014. doi: 10.1080/00949655.2012.745125. URL http://dx.doi.org/10.1080/00949655.2012.745125.

CORREA, M. A.; NOGUEIRA, D. A.; FERREIRA, E. B. Kumaraswamy normal and azzalini's skew normal modeling asymmetry. *Sigmae*, v. 1(1), p. 65–83, 2012.

DE PASCOA, M. A. R.; ORTEGA, E. M. M.; CORDEIRO, G. M. The Kumaraswamy generalized gamma distribution with application in survival analysis. *Stat. Methodol.*, v. 8(5), p. 411–433, 2011. ISSN 1572-3127. doi: 10.1016/j.stamet.2011.04.001. URL http://dx.doi.org/10.1016/j.stamet.2011.04.001.

DE SANTANA, T. V. F.; ORTEGA, E. M.; CORDEIRO, G. M.; SILVA, G. O. The kumaraswamy-log-logistic distribution. *Journal of Statistical Theory and Applications*, v. 11(3), p. 265–291, 2012.

FAREWELL, V. T. A model for a binary variable with time-censored observations. *Biometrika*, v. 64(1), p. 43–46, 1977.

GARG, M. On generalized order statistics from Kumaraswamy distribution. *Tamsui Oxf. J. Math. Sci.*, v. 25(2), p. 153–166, 2009. ISSN 1561-8307.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

GOLDMAN, A. I. Survivorship analysis when cure is a possibility: a monte carlo study. *Statistics in Medicine*, v. 3(2), p. 153–163, 1984.

HUSSIAN, M.; A AMIN, E. Estimation and prediction for the kumaraswamy-inverse rayleigh distribution based on records. *International Journal of Advanced Statistics and Probability*, v. 2(1), p. 21–27, 2014.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, v. 53, p. 457–481, 1958. ISSN 0162-1459.

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

LAI, C.; XIE, M.; MURTHY, D. A modified weibull distribution. *Reliability, IEEE Transactions on*, v. 52(1), p. 33–37, 2003.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, 1982.

LI, C.-S.; TAYLOR, J. M.; SY, J. P. Identifiability of cure models. *Statistics & Probability Letters*, v. 54(4), p. 389–395, 2001.

LILLARD, L. A.; PANIS, C. W. A. Aml multilevel multiprocess statistical software. 2000.

MEEKER, W. Q.; ESCOBAR, L. A. *Statistical methods for reliability data*, volume 314. John Wiley & Sons, 1998.

NADARAJAH, S.; ELJABRI, S. The Kumaraswamy GP distribution. *J. Data Sci.*, v. 11(4), p. 739–766, 2013. ISSN 1680-743X.

NELSON, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, v. 14(4), p. 945–966, 1972.

PIKE, M. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, pages 142–161, 1966.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

RODRIGUES, J.; CANCHO, V. G.; DE CASTRO, M.; LOUZADA-NETO, F. On the unification of long-term survival models. *Statistics & Probability Letters*, v. 79(6), p. 753–759, 2009.

SHAHBAZ, M. Q.; SHAHBAZ, S.; BUTT, N. S. The Kumaraswamy-inverse Weibull distribution. *Pak. J. Stat. Oper. Res.*, v. 8(3), p. 479–489, 2012. ISSN 1816-2711.

SUN, F. B.; KECECIOGLU, D. B. A new method for obtaining the ttt plot for a censored sample. *Reliability and Maintainability Symp.*, 1999.

TACONELI, J. P. Modelo de mistura paramétrico com fragilidade na presença de covariáveis. 2013.

YAKOVLEV, A. Y.; TSODIKOV, A. D. *Stochastic models of tumor latency and their biostatistical applications*, volume 1. OECD Publishing, 1996.

Apêndice

A

TTT Plot

A.1 Método TTT Plot Sem Censura

Para verificar o comportamento da função de risco dos tempos observados, utilizamos um método gráfico baseado no teste do tempo total do teste (TTT), na qual é descrito por Aarset (1985). Seja $0 \leq T_{1:n} \leq T_{2:n} \leq \ldots \leq T_{n:n}$ uma amostra ordenada de um estudo de tempo de vida de n componentes. Então, a estatística TTT, para cada tempo de falha T_i, baseada nesta amostra, é dado por

$$\begin{cases} T_i = \sum_{j=1}^{i} T_{j:n} + (n-i)T_{i:n}, \text{ para } i = 1, 2, \ldots, n, \\ T_0 = 0 \text{ para } i = 0 \end{cases}$$

e a estatística TTT escalada para cada tempo de falha é definida como

$$\varphi_i = T_i / T_n,$$

para $i = 1, 2, \ldots, n$, e $T_n = \sum_{j=1}^{n} T_{j:n}$.

A fda observada de tempo de falha é

$$u_i = \hat{F}(T_{i:n}) = i/n, \ i = 1, 2, \ldots, n,$$

Assim, devemos plotar $\varphi_i \; versus \; u_i$ e conectar os pontos com segmentos de linha.

A.2 Método TTT Plot Com Censura

Assuma que um total de n itens são tempos de vida em que r falhas são observadas enquanto $(n-r)$ itens são censurados. Os tempos de falha ou censura são arranjados em ordem crescente como segue

$$0 < TC_{11} \leq TC_{12} \leq \ldots \leq TC_{1m_1} < TF_1$$
$$< TC_{21} \leq TC_{22} \leq \ldots \leq TC_{2m_2} < TF_2$$
$$< \ldots < TF_r < TC_{r+1,1} \leq TC_{r+1,2} \leq \ldots \leq TC_{r+1,m_{r+1}}$$

em que

- TC_j = tempo de falha da j-ésima falha, $j = 1, 2, \ldots, r$,
- TC_{ji} = tempo da i-ésima censura entre $(j-1)$-ésima e a j-ésima falha, $i = 1, 2, \ldots, m_i$ e $j = 1, 2, \ldots, r+1$,
- $TF_0 = 0$ indica o começo do estudo e TF_{r+1} indica o final do estudo.

Quando censuras estão presentes entre falhas, o número de ordem de falha torna-se incerto. Neste caso, um número de ordem médio (M) é tipicamente estimado como segue

$$M_j = M_{j-1} + I_j, \tag{A.1}$$

em que

- j=1,2,...,r,
- M_{j-1} = número de ordem médio da $(j-1)$-ésima falha,
- M_j = número de ordem médio da j-ésima falha,
- I_j = incremento do número de ordem médio da j-ésima falha devido à censura (s) entre a $(j-1)$-ésima e a j-ésima falha,
- $I_j = \frac{(n+1)-M_{j-1}}{1+K}$,
- K = número de itens além do atual conjunto de censuras, isto é, além $(TC_{j1}, TC_{j2}, \ldots, TC_{jm_j})$.

Observações:

1. Se o primeiro tempo corresponder a uma falha, então $M_1 = 1$. Se o segundo tempo também é uma falha, então $M_2 = M_1 + 1 = 2$. $\cdots$. Isto é válido até uma censura for encontrada.
2. Se o primeiro tempo corresponder a uma censura, então $M_0 = 0$.
3. Se não houver censuras entre a $(j-1)$-ésima e a j-ésima falhas, I_j é igual a I_{j-1}.

Determinado o número médio de ordem para cada tempo de falha, a correspondente função de distribuição acumulada do tempo de falha e a função de sobrevivência podem ser estimadas como segue:

$$F^M(T) = \begin{cases} \frac{M_{j-1}}{n} \text{ para } TF_{j-1} \leq T < TF_j,\ j = 1, 2, \ldots, r+1, \\ 1 \text{ para } T \geq T_{end}, \end{cases}$$

e

$$S^M(T) = \begin{cases} 1 - \frac{M_{j-1}}{n} \text{ para } TF_{j-1} \leq T < TF_j,\ j = 1, 2, \ldots, r+1, \\ 1 \text{ para } T \geq T_{end}. \end{cases}$$

em que

- $TF_0 = 0$,
- $M_0 = 0$,

- $TF_{r+1} = T_{end}$, uma pseudo falha.

Especificamente, no tempo $T = 0, TF_1, TF_2, \cdots, TF_r$ e T_{end} temos

$$\begin{cases} \text{Para } T = 0,\ F_0 = 0 \text{ e } C_0 = 1 \\ \text{Para } T = TF_1,\ F_1 = \frac{M_1}{n} \text{ e } C_1 = 1 - \frac{M_1}{n}, \\ \text{Para } T = TF_2,\ F_2 = \frac{M_2}{n} \text{ e } C_2 = 1 - \frac{M_2}{n}, \\ \cdots \\ \text{Para } T = TF_r,\ F_r = \frac{M_r}{n} \text{ e } C_r = 1 - \frac{M_r}{n}, \\ \text{Para } T = TF_{end},\ F_{end} = 1 \text{ e } C_{end} = 0. \end{cases}$$

Então, o tempo total em teste (TTT) até cada falha observada TF_j, $j = 1, 2, \ldots, r+1$, é dado por

$$TTT_j = n \int_0^{TF_j} C^M(T) dT. \tag{A.2}$$

Mais explicitamente,

$$\begin{cases} \text{Para } T = 0,\ TTT_0 = 0, \\ \text{Para } T = TF_1,\ TTT_1 = TTT_0 + n.S_0.(TF_1 - 0), \\ \text{Para } T = TF_2,\ TTT_2 = TTT_1 + n.S_1.(TF_2 - TF_1), \\ \cdots \\ \text{Para } T = TF_r,\ TTT_r = TTT_{r-1} + n.S_{r-1}.(TF_r - TF_{r-1}), \\ \text{Para } T = TF_{end},\ TTT_{end} = TTT_r + n.S_r.(TF_{end} - TF_r). \end{cases}$$

Assim, a estatística escalada TTT é dada por

$$\begin{cases} \text{Para } T = 0,\ \varphi_0 = 0, \\ \text{Para } T = TF_1,\ \varphi_1 = TTT_1/TTT_{end}, \\ \text{Para } T = TF_2,\ \varphi_2 = TTT_2/TTT_{end}, \\ \cdots \\ \text{Para } T = TF_r,\ \varphi_r = TTT_r/TTT_{end}, \\ \text{Para } T = TF_n,\ \varphi_{end} = 1. \end{cases}$$

Finalmente, plotando $(F_0,\ \varphi_0)$, $(F_1,\ \varphi_1)$, ..., $(F_r,\ \varphi_r)$ e $(F_{end},\ \varphi_{end})$ e conectá-los com segmentos de linha fornecidos pelo TTT escalado para esta amostra censurada.

A função de risco cresce (decresce) se o TTT plot é côncavo (convexo). Se o gráfico aproxima de uma linha diagonal temos função de risco constante e, se a curvatura é côncava e depois convexa a função de risco tem forma unimodal. Se o gráfico apresentar curvatura convexa e depois côncava a função de risco é em forma de banheira.

O gráfico TTT plot é apenas uma condição suficiente e não necessária para indicar a forma da função de risco e será utilizado como um indicador de seu comportamento.