Uma proposta de orientação para o uso de modelos contínuos em dados de sobrevivência discretos

A proposal for guidance on the use of Continuous Models in discrete-data survival

Elisângela Candeias Biazatti
Universidade Federal de Rondônia (UNIR), Departamento de Matemática e Estatística (DME), Ji-Paraná, RO, Brasil
https://orcid.org/0000-0002-2264-9850, elisangela.biazatti@unir.br

Eduardo Yoshio Nakano
Universidade de Brasília (UnB), Departamento de Estatística (EST), Brasília, DF, Brasil
https://orcid.org/0000-0002-9071-8512, nakano@unb.br

Informações do Artigo

Como citar este artigo
BIAZATTI, Elisângela Candeias; NAKANO, Eduardo Yoshio. Uma proposta de orientação para o uso de modelos contínuos em dados de sobrevivência discretos. **REMAT: Revista Eletrônica da Matemática**, Bento Gonçalves, RS, v. 6, n. 2, p. e4002, 25 jul. 2020. DOI: https://doi.org/10.35819/remat2020v6i2id3906

Histórico do Artigo
Submissão: 16 de fevereiro de 2020.
Aceite: 7 de abril de 2020.

Palavras-chave
Análise de Sobrevivência
Censura Intervalar
Tempo de Falha Discreto
Observações Empatadas

Resumo

Modelos discretos não são populares em análise de sobrevivência. Isso se deve principalmente à escassez de trabalhos que abordam a análise de dados discretos na presença de censura. Desta forma, a possibilidade de analisar um conjunto de dados discretos por meio de um modelo contínuo certamente traz facilidade à análise desses dados. Neste contexto, este trabalho propõe guias de decisão que podem auxiliar um pesquisador decidir sobre o uso de um modelo contínuo na análise de dados de sobrevivência originalmente discretos. Esses guias de decisão foram obtidos por meio de simulações de Monte Carlo e levam em consideração o tamanho da amostra, o percentual de censura e a proporção de observações empatadas. Os guias de decisão foram aplicados em três conjuntos de dados obtidos na literatura e se mostrou uma forma simples de decidir quando um modelo contínuo pode ser considerado para o ajuste de dados discretos.

Keywords
Survival Analysis
Interval Censoring
Discrete Time Failure
Tied Observations

Abstract

Discrete models are not popular in survival analysis. This mainly occurs due to the lack of works modeling censored discrete data. Thus, the possibility of analyzing discrete data sets through a continuous model certainly makes this analysis a little easier. In this context, this paper proposes decision guides to help a researcher to decide about the use of a continuous model in the analysis of originally discrete-data survival. These decision guides were obtained through Monte Carlo simulations, considering the sample size, the censored percentages and the proportion of ties observations. The decision guides were applied in three data sets obtained in the literature and which showed a simple way to decide when a continuous model can be considered to modelling discrete data.

1. Introdução

Em Análise de Sobrevivência, existem problemas em que os dados são coletados em sua forma discreta devido a imprecisões nas mensurações (quando o tempo é medido em dias, ou

meses completos, por exemplo). Nestes casos, o dado não é originalmente discreto. De fato, o tempo é contínuo e o que ocorre é que a mensuração do tempo é limitada. A principal característica neste caso é a observação de um grande número de tempos empatados. Esse problema pode ser tratado como um caso particular na metodologia de dados grupados, ao considerar intervalos de tempo unitários.

Por um outro lado, existem casos em que os dados de sobrevivência são originalmente discretos, quando o tempo é medido em ciclos, impactos sofridos ou sessões de um tratamento, por exemplo. Em geral, isso ocorre em problemas em que o tempo é representado por meio de dados de contagem. Nestas situações, o que se faz na prática é considerar que esses dados poderiam ser contínuos e realizar a análise por meio de um modelo contínuo (NAKANO; CARRASCO, 2006). Uma outra possibilidade é analisar esses dados discretos por meio da metodologia de dados grupados, com base em um modelo contínuo e considerando intervalos de tempos unitários.

Na última década, é possível encontrar na literatura uma série de trabalhos que apresentam propostas de distribuições discretas. Entre outros, pode-se citar Almalki e Nadarajah (2014), Bakouch, Jazi e Nadarajah (2014) e Nekoukhou e Bidram (2015). No entanto, o uso de modelos discretos não é uma prática comum na área de análise de sobrevivência. Isso se deve principalmente à escassez de trabalhos e procedimentos computacionais que abordam modelos discretos na presença de censura.

Neste contexto, o objetivo deste trabalho foi verificar em quais situações é aceitável utilizar um modelo contínuo ou a metodologia de dados grupados para analisar dados discretos (dados originalmente discretos) de sobrevivência. Mais especificamente, este trabalho propôs guias para o uso de modelos contínuos (adotando-se ou não a metodologia de dados grupados) na análise de dados discretos de sobrevivência. Por se tratar de dados de sobrevivência, optamos neste trabalho por adotar a distribuição Weibull e lognormal. A distribuição Weibull foi escolhida por ser uma distribuição muito utilizada na modelagem de dados de sobrevivência devido a sua versatilidade e relativa simplicidade (RINNE, 2008), além de possuir a sua “versão discreta”, dada pela distribuição Weibull discreta de Nakagawa e Osaki (NAKAGAWA; OSAKI, 1975). Adicionalmente, a escolha da distribuição lognormal foi devido ao fato da mesma permitir funções de risco não monótonas.

2. Metodologia

2.1. Distribuições Weibull e Weibull discreta

A distribuição Weibull foi proposta originalmente por Wallodi Weibull em 1951 (WEIBULL, 1951) e, desde então, devido em grande parte à sua simplicidade e flexibilidade, tem sido uma das distribuições de probabilidades mais utilizadas na modelagem de dados biomédicos e também

industriais. Sua função de densidade, função de sobrevivência e função de risco são dadas, respectivamente, por:

$$\begin{aligned} f(t) &= \alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}},\ \ t>0, \\ S(t) &= e^{-\lambda t^{\alpha}},\ \ t>0 \text{ e} \\ h(t) &= \alpha\lambda t^{\alpha-1},\ \ t>0. \end{aligned} \tag{1}$$

Em (1), $\alpha > 0$ e $\lambda > 0$ são os parâmetros de forma e escala, respectivamente.

A distribuição Weibull discreta de Nakagawa e Osaki (1975) pode ser obtida agrupando os tempos em intervalos unitários (NAKANO; CARRASCO, 2006). A partir de (1) e considerando $q = e^{-\lambda}$, as funções de probabilidade, sobrevivência e risco da distribuição) Weibull discreta são dadas por:

$$\begin{aligned} f(t) &= q^{t^{\alpha}} - q^{(t+1)^{\alpha}},\ \ t=0,1,2,\ldots \\ S(t) &= q^{(t+1)^{\alpha}},\ \ \ t=0,1,2,\ldots \text{ e} \\ h(t) &= 1 - q^{(t+1)^{\alpha}-t^{\alpha}},\ \ \ t=0,1,2,\ldots. \end{aligned} \tag{2}$$

em que $\alpha > 0$ e $\alpha < q < 1$ são os parâmetros da distribuição. Assim, como na distribuição Weibull contínua, a função de risco da distribuição Weibull discreta é monótona decrescente se $\alpha < 1$, constante se $\alpha = 1$ e monótona crescente se $\alpha > 1$ (VILA; NAKANO; SAULO, 2019).

2.2. Distribuições lognormal e lognormal discreta

A distribuição lognormal é uma distribuição bastante utilizada na modelagem de dados de sobrevivência. Isto está relacionado ao fato da mesma permitir acomodar funções de risco unimodais, o que pode ser adequado em diversas situações práticas. Uma variável aleatória T tem distribuição lognormal se $Y = log\,(T)$ tem uma distribuição normal. Assim, a função densidade, a função de sobrevivência e a função risco da distribuição lognormal são dadas, respectivamente, por:

$$\begin{aligned} f(t) &= \frac{1}{\sigma t\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{log\,(t)-\mu}{\sigma}\right)^2\right\},\ \ t>0, \\ S(t) &= 1 - \Phi\left(\frac{log\,(t)-\mu}{\sigma}\right),\ \ t>0\ e \\ h(t) &= 1 - q^{(t+1)^{\alpha}-t^{\alpha}},\ \ \ t=0,1,2,\ldots. \end{aligned} \tag{3}$$

Em (3), $-\infty < \mu < \infty$ e $\sigma > 0$ são os parâmetros de locação e escala, respectivamente, e Φ é a função de distribuição acumulada da normal padrão.

De forma similar à distribuição Weibull discreta, a distribuição lognormal discreta pode ser obtida agrupando os tempos em intervalos unitários. Assim, as funções de probabilidade, sobrevivência e risco da lognormal discreta podem ser escritas como:

$$
\begin{aligned}
f(t) &= \Phi\left(\frac{log\,(t+1)-\mu}{\sigma}\right) - \Phi\left(\frac{log\,(t)-\mu}{\sigma}\right), \quad t = 0, 1, 2, \ldots \\
S(t) &= 1 - \Phi\left(\frac{log\,(t+1)-\mu}{\sigma}\right), \quad t = 0, 1, 2, \ldots e \\
h(t) &= \frac{\Phi\left(\frac{\log\,(t+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\log\,(t)-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log\,(t)-\mu}{\sigma}\right)}, \quad t = 0, 1, 2, \ldots
\end{aligned}
\tag{4}
$$

Em (4), $-\infty < \mu < \infty$ e $\sigma > 0$ são os parâmetros da distribuição e Φ é a função de distribuição acumulada da normal padrão.

2.3. Função de verossimilhança

Considerando um mecanismo de censura à direta, a função de verossimilhança pode ser obtida de modo que a contribuição do i-ésimo indivíduo da amostra é dada pela:

- função densidade de probabilidades (ou função de probabilidade, no caso discreto), $f(t)$, se o i-ésimo indivíduo falhar no tempo t;
- função de sobrevivência, $S(t)$, se o i-ésimo indivíduo for censurado no tempo t.

Assim, a função de verossimilhança é dada por

$$L(\theta) = \prod_{i \in F} f(t_i) \prod_{i \notin F} S(t_i)$$

em que θ é o vetor de parâmetros a serem estimados e F denota o conjunto de indivíduos que falharam (não censurados).

Uma outra possibilidade de analisar dados discretos de sobrevivência é por meio da metodologia de dados grupados, com base em um modelo contínuo e considerando intervalos de tempos unitários. Neste caso específico, o j-ésimo intervalo de tempo é definido por $l_j = [j, j+1), j = 0, 1, 2, \ldots, k$, em que k é o número de tempos distintos observados.

Neste caso, a função de verossimilhança pode ser obtida de modo que a contribuição do i-ésimo indivíduo da amostra é dado por (HASHIMOTO, 2008; BIAZATTI, 2017):

- Se o i-ésimo indivíduo falhar no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $1 - \frac{S(j+1)}{S(j)}$;
- Se o i-ésimo indivíduo sobreviver (ou seja, estiver sob risco) no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $\frac{S(j+1)}{S(j)}$;
- Se o i-ésimo indivíduo for censurado no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $\left(\frac{S(j+1)}{S(j)}\right)^{\frac{1}{2}}$.

Assim, segundo a metodologia de dados grupados, a função de verossimilhança é dada por:

$$L(\theta) = \prod_{j=1}^{k} \left\{ \prod_{i \in F_j} \left[1 - \frac{S(j+1)}{S(j)} \right] \times \prod_{i \in R_j} \frac{S(j+1)}{S(j)} \times \prod_{i \in C_j} \left[\frac{S(j+1)}{S(j)} \right]^{\frac{1}{2}} \right\},$$

em que θ é o vetor de parâmetros a serem estimados, F_i denota o conjunto de indivíduos que falharam no j-ésimo intervalo, R_i denota o conjunto de indivíduos sob risco no j-ésimo intervalo e C_i denota o conjunto de indivíduos censurados no j-ésimo intervalo.

2.4. Avaliação do desempenho dos modelos e simulação dos dados

As simulações apresentadas neste trabalho têm como objetivo comparar as estimativas da função de sobrevivência pelos modelos discreto e contínuo e pela metodologia de dados grupados. Estimativas da função de sobrevivência foram obtidas pelo estimador de Kaplan-Meier (KAPLAN; MEIER, 1958) e por meio das três metodologias, verificando em quais situações pode ser razoável decidir pelo uso de distribuições contínuas (considerando ou não a metodologia de dados grupados) na análise de dados discretos. O critério para a tomada desta decisão será baseado no tamanho da amostra, percentual de censura e em uma medida de Proporção de Empates (pe), proposto por Chalita, Colosimo e Demétrio (2002):

$$pe = \frac{d-k}{n},$$

em que n é o tamanho da amostra, d é o número total de falhas e k é o número total de falhas distintas observadas na amostra.

Os modelos serão avaliados por meio do Erro Absoluto Médio (EAM) das estimativas da função de sobrevivência, definido por:

$$EAM = \frac{1}{k} \sum_{j=1}^{k} \left| \hat{S}(t_j) - \hat{S}_{KM}(t_j) \right|, \quad (5)$$

em que $\hat{S}(.)$ é a estimativa da função de sobrevivência por meio das três metodologias e $\hat{S}_{KM}(.)$ é a estimativa de Kaplan-Meier da função de sobrevivência. Em (5), t_j $(j = 1, 2, ..., k)$ são os tempos de falhas distintos.

A discordância entre as estimativas das funções de sobrevivência obtidas pelos modelos discreto e contínuo (com ou sem dados grupados) será avaliado por meio da Divergência Absoluta Média (DAM), definida por:

$$DAM = \frac{1}{J} \sum_{j=0}^{J} \left| \hat{S}_c(j) - \hat{S}_d(j) \right|, \quad (6)$$

em que $J = max\{t_1,\ t_2, ...,\ t_n\}$, $\hat{S}_c(.)$ é a estimativa da função de sobrevivência obtida pelo modelo contínuo (com ou sem dados grupados) e $\hat{S}_d(.)$ é a estimativa da função de sobrevivência obtida pelo modelo discreto.

O estudo de simulação considerou amostras de tamanho $n =$ 20, 30, 40, 50, 60, 70, 80, 90, 100, 200 e 500 com 0%, 10%, 20%, 50% e 75% de censura. As amostras dos tempos de sobrevivência discretos foram geradas segundo as distribuições Weibull e lognormal discretas e considerando um mecanismo de censura aleatório. Os Apêndices A e B apresentam os procedimentos adotados para a geração dos tempos censurados. Os dados foram gerados considerando diversos valores para os parâmetros, segundo uma distribuição de probabilidade. Os valores do parâmetro α da distribução Weibull discreta foram escolhidos aleatoriamente segundo uma distribuição uniforme no intervalo (0,25 ; 2) e os valores do parâmetro q segundo uma distribuição uniforme no intervalo (0,6 ; 0,99). Já os parâmetros μ e σ da distribuição lognormal discreta foram escolhidos de acordo com distribuições uniformes nos intervalos (–1 ; 3) e (0,25 ; 5), respectivamente. Esse procedimento permitiu simular os dados para uma vasta gama de valores dos parâmetros e também garantiu variação nos valores da proporção de empates (pe) nas amostras geradas.

Neste estudo de simulação foram realizados $M =$ 10.000 réplicas de Monte Carlo. Para cada uma dessas M amostras, os respectivos modelos foram ajustados segundo as três metodologias (modelo discreto, modelo contínuo e dados grupados) e foram obtidos os valores do EAM e da DAM (Expressões 5 e 6). Note que as estimativas de máxima verossimilhança dos modelos Weibull e lognormal contínuo não podem ser obtidas no caso em que pelo menos uma observação não censurada é igual a zero. Nestes casos, para a estimação dos parâmetros do modelo contínuo, os valores de t iguais a zero foram substituídos por $t = 0{,}5$.

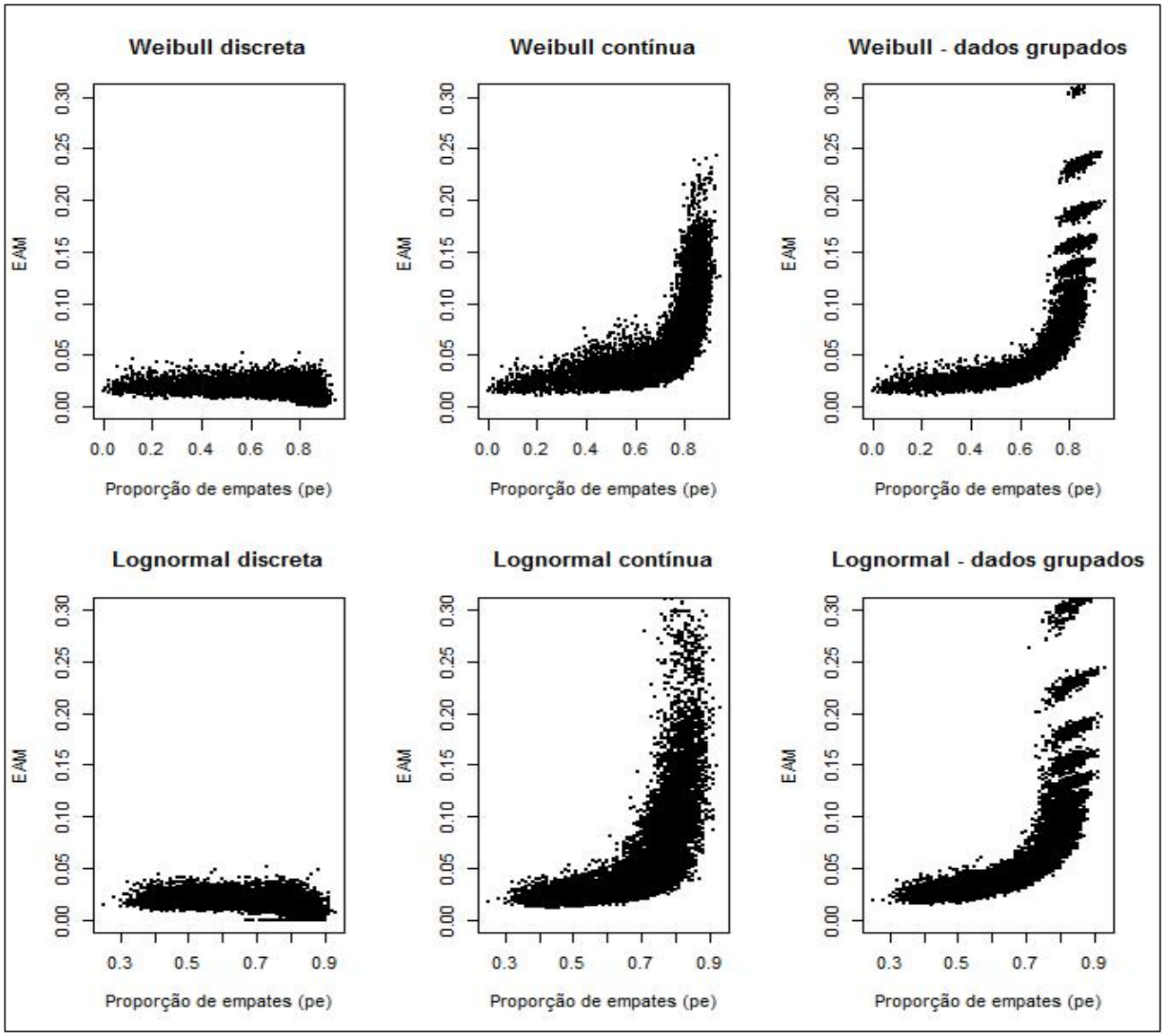
3. Resultados das simulações e definição dos pontos de corte

A Figura 1 apresenta os valores do EAM de acordo com os percentuais de empates (pe). Os valores do EAM foram obtidos segundo as distribuições Weibull e lognormal (discretos, contínuos e segundo a metodologia de dados grupados) e considerando uma amostra de $n = 100$ observações e 10% de censura.

Como pode ser visto na Figura 1, a proporção de empates (pe) não influencia o valor do EAM quando a inferência é realizada pelas distribuições discretas. Este é um resultado esperado, visto que os dados são provenientes de distribuições discretas. Já os $EAMs$ das análises realizadas por meio de suas respectivas distribuições contínuas ou por meio da metodologia de dados grupados aumentam à medida que crescem os valores da pe . Esses mesmos comportamentos foram observados nos modelos Weibull e lognormal e também para diferentes tamanhos de amostras e percentuais de censura. Esse resultado mostra que a inferência de

dados de sobrevivência discretos por meio de distribuições contínuas (mesmo que via metodologia de dados grupados) não é adequado, quando a amostra apresenta alta proporção de valores empatados.

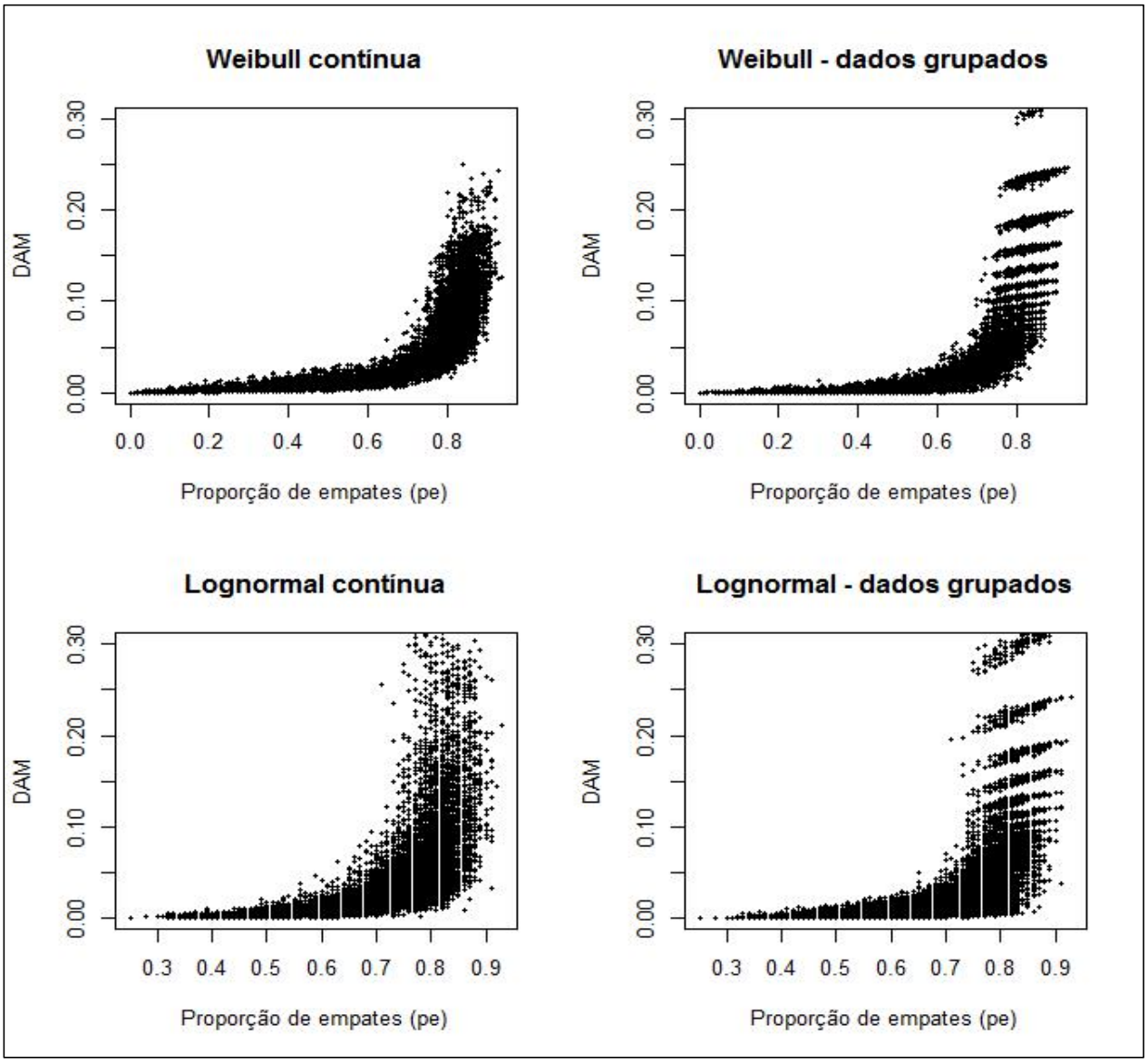
Figura 1 – Valores do *EAM* de acordo com os percentuais de empates (pe) para amostras de tamanho $n = 100$ e 10% de censura.



Fonte: Dados da pesquisa (2020).

Para verificar o quanto as estimativas da função de sobrevivência obtidas pelo modelo contínuo (com e sem dados grupados) divergem do modelo discreto, os valores da DAM foram obtidos. A Figura 2 apresenta os valores da DAM de acordo com os percentuais de empates (pe) segundo as distribuições Weibull e lognormal (discretos, contínuos e segundo a metodologia de dados grupados) e considerando uma amostra de $n = 100$ observações e 10% de censura. Como esperado (visto que o EAM independe da *pe* para o modelo discreto), a DAM também aumenta à medida que crescem os valores da pe .

Figura 2 – Valores da DAM de acordo com os percentuais de empates (pe) para amostras de tamanho $n = 100$ e 10% de censura.



Fonte: Dados da pesquisa (2020).

3.1. Obtenção dos pontos de corte para pe

Como visto pelos resultados apresentados na Figura 2, à medida que a proporção de empates (pe) aumenta, maior é a divergência entre as estimativas da função de sobrevivência obtidas pelos modelos contínuos quando comparado com o modelo discreto (que é o padrão-ouro). No entanto, é importante destacar que essa divergência (DAM) é pequena para baixos valores da pe . Desta forma, é interessante definir pontos de corte para que o pesquisador possa decidir (a partir da *pe* observadas em sua amostra), se é razoável realizar a análise por meio de um modelo contínuo (optando ou não pela metodologia de dados grupados). Assume-se, neste trabalho, que o uso de um modelo contínuo em dados discretos é razoável desde que $DAM \leq 0{,}05$. Isto significa que a estimativa da função de sobrevivência não irá divergir, em média, por mais de 0,05, quando comparado com seu respectivo modelo discreto.

Neste contexto, este trabalho propõe a adoção de dois pontos de corte (Corte 1 < Corte 2) para a pe, de forma que se pe < Corte 1, o uso de um modelo contínuo em dados discretos será considerado razoável e se pe > Corte 2, deve-se utilizar um modelo discreto para a análise dos dados. Para definir estes pontos de corte, primeiramente é importante listar os quatro possíveis resultados na tomada de decisão: 1) Erro I – Dizer que o modelo contínuo é adequado quando não é (pe < Corte 1 e DAM > 0,05); 2) Acerto I – Dizer que o modelo contínuo é adequado quando é (pe < Corte 1 e $DAM \leq 0{,}05$); 3) Acerto II – Dizer que o modelo contínuo não é adequado quando não é (pe > Corte 2 e DAM > 0,05); e 4) Erro II – Dizer que o modelo contínuo não é adequado quando é (pe > Corte 2 e $DAM \leq 0{,}05$).

Neste trabalho, os Cortes 1 e 2 foram especificados de forma controlar os Erros I e II. O Ponto de Corte 1 é definido como o maior valor cuja probabilidade do Erro I (estimado por meio dos valores obtidos nas simulações de Monte Carlo) não ultrapasse 5%. Assim, o uso de um modelo contínuo em dados discreto será razoável se pe < Corte 1. De forma similar, o Ponto de Corte 2 é definido como o menor valor cuja a probabilidade do Erro II não ultrapasse 5%. Desta forma, pe > Corte 2 indica a necessidade do uso de um modelo discreto.

Para uma amostra de tamanho $n = 100$ e 10% de censura, os pontos de corte para a proporção de empates para o modelo Weibull contínuo foram: Corte 1 = 0,77 e Corte 2 = 0,80. Isto é, uma amostra de 100 observações discretas com 10% de censuras e proporção de empates menor que 0,77 pode ser modelada por meio de uma distribuição contínua. No entanto, se essa amostra apresentar uma proporção de empates maior que 0,80, isto significa que não é adequado realizar a análise desses dados por meio de uma distribuição contínua. Note que se Corte 1 $\leq$ pe $\leq$ Corte 2, a decisão de aceitar (ou rejeitar) um modelo contínuo gera Erros I e II maiores que 5%. Nestes casos, o diagnóstico será considerado inconclusivo e, portanto, a decisão sobre o uso de um modelo contínuo deve ser feita a partir da DAM obtida. Esse mesmo procedimento pode ser aplicado para a distribuição lognormal (ou qualquer outra distribuição de probabilidades) e também para a inferência baseada na metodologia de dados grupados.

A Tabela 1 e a Figura 3 apresentam os pontos de cortes da proporção de empates (pe) de acordo com o tamanho da amostra e percentual de censura para os modelos contínuos Weibull e lognormal (optando ou não pela metodologia de dados grupados). É possível notar que os resultados obtidos pelo modelo Weibull e lognormal foram similares. Ademais, os pontos de corte da proporção de empates (pe) aumentam à medida que o tamanho da amostra cresce, isto é, quanto maior o tamanho da amostra, maior é a proporção de empates que pode ser observada na amostra. Em contrapartida, o percentual de censura apresentou um comportamento inverso. Quanto maior o percentual de censura menor são os pontos de corte da pe. Este comportamento está de acordo com os resultados apresentados em Nakano e Carrasco (2006), que mostraram

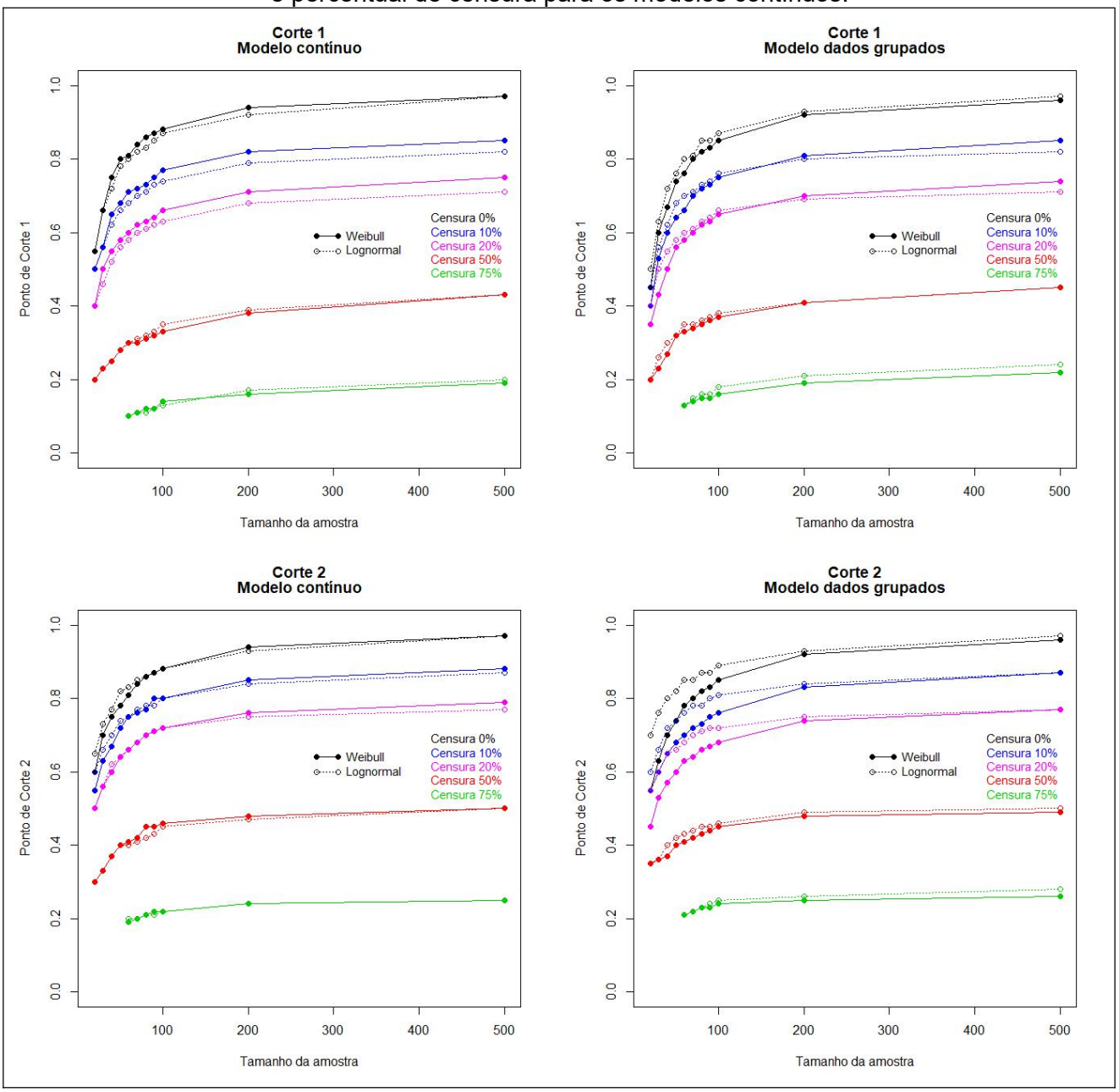
que o aumento da censura causa um efeito contrário ao aumento do tamanho da amostra na análise de dados discretos por modelos contínuos. Ainda, o aumento dos pontos de corte da pe à medida que a amostra cresce também é um resultado esperado, visto que amostras maiores tendem a apresentar maior amplitude dos dados, fazendo com que ocorra um menor erro na aproximação de modelos contínuos em dados discretos (NAKANO; CARRASCO, 2006). Resultados similares podem ser observados com os pontos de corte da *pe* para o uso da metodologia de dados grupados, isto é, os pontos de corte crescem à medida que o tamanho da amostra aumenta e decrescem à medida que aumenta o percentual de censura.

Tabela 1 – Pontos de cortes (Corte 1; Corte 2) da proporção de empates (pe) de acordo com o tamanho da amostra e percentual de censura para os modelos contínuos.

n	Censura 0%		Censura 10%		Censura 20%		Censura 50%		Censura 75%	
	W*	LN*	W*	LN*	W*	LN*	W*	LN*	W*	LN*
Modelo contínuo										
20	0,55;0,60	0,55;0,65	0,50;0,55	0,50;0,60	0,40;0,50	0,40;0,50	0,20;0,30	0,20;0,30	**	
30	0,66;0,70	0,66;0,73	0,56;0,63	0,56;0,66	0,50;0,56	0,46;0,56	0,23;0,33	0,23;0,33	**	
40	0,75;0,75	0,72;0,77	0,65;0,67	0,62;0,70	0,55;0,60	0,52;0,62	0,25;0,37	0,25;0,37	**	
50	0,80;0,78	0,78;0,82	0,68;0,72	0,66;0,74	0,58;0,64	0,56;0,64	0,28;0,40	0,28;0,40	**	
60	0,81;0,81	0,80;0,83	0,71;0,75	0,68;0,75	0,60;0,66	0,58;0,66	0,30;0,41	0,30;0,40	0,10;0,19	0,10;0,20
70	0,84;0,84	0,82;0,85	0,72;0,76	0,70;0,77	0,62;0,68	0,60;0,68	0,30;0,42	0,31;0,41	0,11;0,20	0,11;0,20
80	0,86;0,86	0,83;0,86	0,73;0,77	0,71;0,78	0,63;0,70	0,61;0,70	0,31;0,45	0,32;0,42	0,12;0,21	0,11;0,21
90	0,87;0,87	0,85;0,87	0,75;0,80	0,73;0,78	0,64;0,71	0,62;0,71	0,32;0,45	0,33;0,43	0,12;0,22	0,12;0,21
100	0,88;0,88	0,87;0,88	0,77;0,80	0,74;0,80	0,66;0,72	0,63;0,72	0,33;0,46	0,35;0,45	0,14;0,22	0,13;0,22
200	0,94;0,94	0,92;0,93	0,82;0,85	0,79;0,84	0,71;0,76	0,68;0,75	0,38;0,48	0,39;0,47	0,16;0,24	0,17;0,24
500	0,97;0,97	0,97;0,97	0,85;0,88	0,82;0,87	0,75;0,79	0,71;0,77	0,43;0,50	0,43;0,50	0,19;0,25	0,20;0,25
Dados grupados										
20	0,45;0,55	0,50;0,70	0,40;0,55	0,45;0,60	0,35;0,45	0,40;0,55	0,20;0,35	0,20;0,35	**	
30	0,60;0,63	0,63;0,76	0,53;0,60	0,56;0,66	0,43;0,53	0,50;0,60	0,23;0,36	0,26;0,36	**	
40	0,67;0,70	0,72;0,80	0,60;0,65	0,62;0,72	0,50;0,57	0,55;0,65	0,27;0,37	0,30;0,40	**	
50	0,74;0,74	0,76;0,82	0,64;0,68	0,68;0,74	0,56;0,60	0,58;0,66	0,32;0,40	0,32;0,42	**	
60	0,76;0,78	0,80;0,85	0,66;0,70	0,70;0,76	0,58;0,63	0,60;0,68	0,33;0,41	0,35;0,43	0,13;0,21	0,13;0,21
70	0,80;0,80	0,81;0,85	0,70;0,72	0,71;0,78	0,60;0,64	0,61;0,70	0,34;0,42	0,35;0,44	0,14;0,22	0,15;0,22
80	0,82;0,82	0,85;0,87	0,72;0,73	0,73;0,78	0,62;0,66	0,63;0,71	0,35;0,43	0,36;0,45	0,15;0,23	0,16;0,23
90	0,83;0,83	0,85;0,87	0,73;0,75	0,74;0,80	0,63;0,67	0,64;0,72	0,36;0,44	0,37;0,45	0,15;0,23	0,16;0,24
100	0,85;0,85	0,87;0,89	0,75;0,76	0,76;0,81	0,65;0,68	0,66;0,72	0,37;0,45	0,38;0,46	0,16;0,24	0,18;0,25
200	0,92;0,92	0,93;0,93	0,81;0,83	0,80;0,84	0,70;0,74	0,69;0,75	0,41;0,48	0,41;0,49	0,19;0,25	0,21;0,26
500	0,96;0,96	0,97;0,97	0,85;0,87	0,82;0,87	0,74;0,77	0,71;0,77	0,45;0,49	0,45;0,50	0,22;0,26	0,24;0,28

*W: Weibull; LN: lognormal. **É esperado menos de 13 observações não censuradas. Mesmo que não sejam observados empates, o uso de modelos contínuos é inadequado.
Fonte: Dados da pesquisa (2020).

Figura 3 – Pontos de cortes da proporção de empates (pe) de acordo com o tamanho da amostra e percentual de censura para os modelos contínuos.



Fonte: Dados da pesquisa (2020).

Com base nos resultados apresentados na Figura 3, este trabalho propõe guias para o uso de modelos contínuos ou da metodologia de dados grupados para a análise de dados de sobrevivência discretos. Estes guias consideraram as decisões mais conservadoras baseadas no estudo de simulação realizado. Isto é, esses guias adotaram o menor valor de Corte 1 e o maior valor de Corte 2 nos cenários envolvidos. Esse procedimento garante que as probabilidades dos Erros I e II não ultrapasse os limites definidos (5%). As Tabela 2 e 3 apresentam esses guias com as decisões em diversos cenários para o uso de um modelo contínuo e para o uso da metodologia de dados grupados, respectivamente.

Tabela 2– Guia para o uso de um modelo contínuo em dados discretos, segundo tamanho da amostra, percentual de censura e proporção de empates.

Tamanho da amostra	**Percentual de censura**				
	Até 10%	**10% a 20%**	**20% a 50%**	**50% a 75%**	**Mais de 75%**
$n < 20$	Recomendado modelo discreto	Recomendado modelo discreto	Recomendado modelo discreto	Recomendado modelo discreto	Recomendado modelo discreto
$20 \leq n \leq 30$	Corte 1 = 0,50 Corte 2 = 0,73	Corte 1 = 0,40 Corte 2 = 0,66	Corte 1 = 0,20 Corte 2 = 0,56	Recomendado modelo discreto	Recomendado modelo discreto
$30 < n \leq 50$	Corte 1 = 0,56 Corte 2 = 0,82	Corte 1 = 0,46 Corte 2 = 0,74	Corte 1 = 0,23 Corte 2 = 0,64	Recomendado modelo discreto	Recomendado modelo discreto
$50 < n \leq 100$	Corte 1 = 0,66 Corte 2 = 0,88	Corte 1 = 0,56 Corte 2 = 0,80	Corte 1 = 0,28 Corte 2 = 0,72	Corte 1 = 0,10 Corte 2 = 0,46	Inconclusivo
$100 < n \leq 200$	Corte 1 = 0,74 Corte 2 = 0,94	Corte 1 = 0,63 Corte 2 = 0,85	Corte 1 = 0,33 Corte 2 = 0,76	Corte 1 = 0,13 Corte 2 = 0,48	Inconclusivo
$n > 200$	Corte 1 = 0,79 Corte 2 = 0,97	Corte 1 = 0,68 Corte 2 = 0,88	Corte 1 = 0,38 Corte 2 = 0,79	Corte 1 = 0,16 Corte 2 = 0,50	Inconclusivo

Notas: Um modelo contínuo pode ser considerado quando $pe <$ Corte 1; um modelo discreto deve ser utilizado quando $pe >$ Corte 2; se Corte 1 $\leq pe \leq$ Corte 2, o diagnóstico é inconclusivo. Fonte: Dados da pesquisa (2020).

Tabela 3– Guia para o uso de um modelo contínuo com a metodologia de dados grupados em dados discretos, segundo tamanho da amostra, percentual de censura e proporção de empates.

Tamanho da amostra	**Percentual de censuras**				
	Até 10%	**10% a 20%**	**20% a 50%**	**50% a 75%**	**Mais de 75%**
$n < 20$	Recomendado modelo discreto	Recomendado modelo discreto	Recomendado modelo discreto	Recomendado modelo discreto	Recomendado modelo discreto
$20 \leq n \leq 30$	Corte 1 = 0,40 Corte 2 = 0,76	Corte 1 = 0,35 Corte 2 = 0,66	Corte 1 = 0,20 Corte 2 = 0,60	Recomendado modelo discreto	Recomendado modelo discreto
$30 < n \leq 50$	Corte 1 = 0,53 Corte 2 = 0,82	Corte 1 = 0,43 Corte 2 = 0,74	Corte 1 = 0,23 Corte 2 = 0,66	Recomendado modelo discreto	Recomendado modelo discreto
$50 < n \leq 100$	Corte 1 = 0,64 Corte 2 = 0,89	Corte 1 = 0,56 Corte 2 = 0,81	Corte 1 = 0,32 Corte 2 = 0,72	Corte 1 = 0,13 Corte 2 = 0,46	Inconclusivo
$100 < n \leq 200$	Corte 1 = 0,75 Corte 2 = 0,93	Corte 1 = 0,65 Corte 2 = 0,84	Corte 1 = 0,37 Corte 2 = 0,75	Corte 1 = 0,16 Corte 2 = 0,49	Inconclusivo
$n > 200$	Corte 1 = 0,80 Corte 2 = 0,97	Corte 1 = 0,69 Corte 2 = 0,87	Corte 1 = 0,41 Corte 2 = 0,77	Corte 1 = 0,19 Corte 2 = 0,50	Inconclusivo

Notas: Um modelo contínuo com a metodologia de dados grupados pode ser considerado quando $pe <$ Corte 1; um modelo discreto deve ser utilizado quando $pe >$ Corte 2; se Corte 1 $\leq pe \leq$ Corte 2, o diagnóstico é inconclusivo. Fonte: Dados da pesquisa (2020).

Note que os guias apresentados pelas Tabela 2 e 3 recomendam o uso de um modelo discreto quando a amostra apresentar menos do que 20 observações. Isso é devido ao fato de amostras pequenas apresentarem poucos valores distintos, fazendo que ocorra um maior erro na

aproximação dos modelos contínuos em dados discretos. Ainda, essa mesma recomendação é feita quando o percentual de censura é alto, pois o aumento da censura causa um efeito similar à diminuição do tamanho da amostra (quanto maior o percentual de censura, menor é o número de falhas distintas). Pode-se notar, a partir das Tabela 2 e 3, que a metodologia de dados grupados apresenta limites da pe menores para amostras pequenas com baixo percentual de censura e limites maiores para amostras grandes com alto percentual de censura.

4. Aplicações

4.1. Aplicação 1

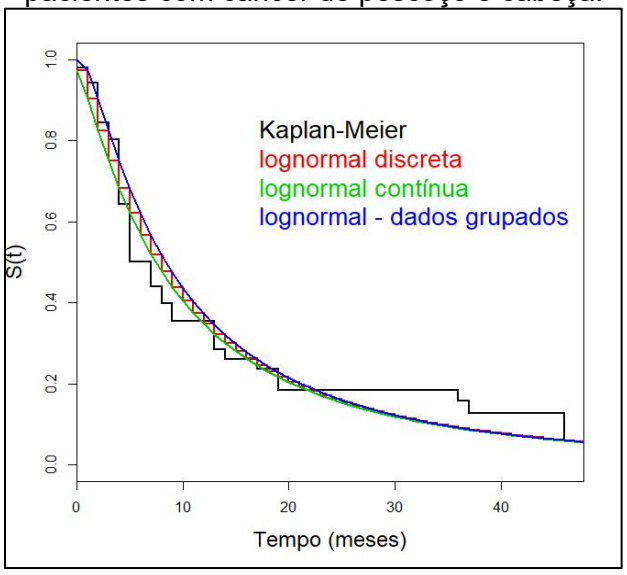
Efron (1988) apresentou um estudo com 51 pacientes com câncer de pescoço e cabeça realizado pelo Grupo de Oncologia do Norte da Califórnia. Os tempos de sobrevivência em meses destes pacientes são apresentados na Tabela 4.

Tabela 4 – Tempos de sobrevivência em meses de pacientes com câncer de pescoço e cabeça.

0; 1; 1; 2; 2; 2$^+$; 2; 2; 2; 3; 3; 4; 4; 4; 4; 4; 4; 4; 4; 5; 5; 5; 5; 5; 5; 5; 6$^+$; 7; 7; 7; 8; 8; 9; 9$^+$; 9; 10$^+$; 13; 13; 13; 14; 17; 17$^+$; 19; 19; 36; 36$^+$; 37; 40$^+$; 44$^+$; 46$^+$; 46

Nota: "+" indica observações censuradas à direita.
Fonte: Efron (1988, p. 415).

Figura 4 – Estimativas da função de sobrevivência do tempo de sobrevivência de pacientes com câncer de pescoço e cabeça.



Fonte: Dados da Tabela 4.

O evento de interesse é o óbito do paciente e a variável resposta T é o tempo, em meses completos, do início do estudo até o óbito do paciente ou censura. Aqui, $T = 0$ indica que o paciente morreu antes de completar um mês de diagnóstico positivo. Para este exemplo, a distribuição lognormal discreta apresentou um melhor ajuste quando comparado com a distribuição Weibull discreta ($EAM_{\text{Weibull discreta}} = 0{,}0788$; $EAM_{\text{lognormal discreta}} = 0{,}0486$). Desta forma, foi adotada a distribuição lognormal para os ajustes dos modelos discreto e contínuos.

Esta amostra é composta de $n = 51$ observações, das quais 9 (17,6%) são censuradas. A proporção de empates é $pe = 0{,}5098$, que é inferior aos limites inferiores (Cortes 1) sugeridos

pelas Tabelas 2 e 3 para uma amostra de 50 a 100 observações com 10% a 20% de censura. Desta forma, tem-se que não há perdas significativas ao considerar um modelo contínuo (com ou sem a metodologia de dados grupados) para o ajuste desse conjunto de dados. De fato, como pode ser visto na Figura 4, a divergência entre as estimativas da função de sobrevivência do modelo discreto e dos modelos contínuos foram pequenas ($DAM_{\text{contínuo}} = 0{,}0091$ e $DAM_{\text{dados grupados}} = 0{,}0182$).

4.2. Aplicação 2

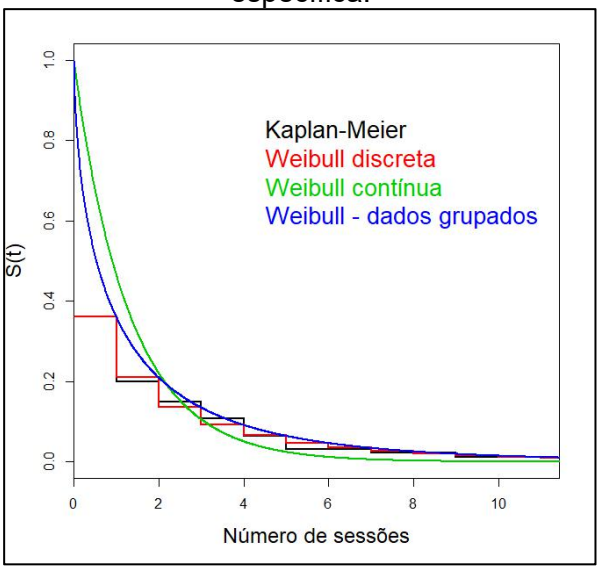
Os dados dessa aplicação são referentes ao estudo apresentado por Corrêa *et al.* (2016) e Silva *et al.* (2017), que mediram o tempo até o alívio da dor em pacientes com dor lombar crônica não específica submetidos a sessões de 30 minutos de um tratamento baseado em estimulações com uma corrente interferencial. Cada um dos participantes realizou três sessões do tratamento por semana em dias alternados, totalizando 12 sessões. O evento de interesses é o alívio ou diminuição da dor lombar, que foi definido como a redução em no mínimo 50% da escala numérica de dor em relação ao valor observado no início do tratamento. A variável resposta T foi definida como o número de sessões malsucedidas anteriores à sessão que aliviou ou diminuiu a dor lombar. Neste estudo, $T = 0$ indica que o paciente teve o alívio da dor logo na primeira sessão do tratamento. Os números de seções malsucedidas até o alívio da dor são apresentados na Tabela 5.

Tabela 5 – Estimativas da função de sobrevivência do tempo até o alívio da dor em pacientes com dor lombar crônica não específica.

Tempo*	Frequência	
	Observações não censuradas	Observações censuradas
0	64	0
1	16	0
2	5	1
3	4	0
4	4	0
5	3	0
7	1	0
9	1	0
11	0	1

*Número de sessões malsucedidas.
Fonte: Corrêa *et al*. (2016, dados obtidos com os autores).

Figura 5 – Estimativas da função de sobrevivência do tempo até o alívio da dor em pacientes com dor lombar crônica não específica.



Fonte: Dados da Tabela 5.

Para este exemplo, ambas distribuições Weibull discreta e lognormal discreta apresentam um bom ajuste dos dados ($EAM_{\text{Weibull discreta}} = 0{,}0074$; $EAM_{\text{lognormal discreta}} = 0{,}0104$). Desta forma,

devido ao menor valor do EAM , foi adotada aqui a distribuição Weibull para os ajustes dos modelos discreto e contínuos.

Neste exemplo, a amostra é composta de $n = 100$ observações, das quais duas (2%) são censuradas. A proporção de empates observada na amostra é $pe = 0{,}9$. Como o valor da pe é maior que os limites superiores (Cortes 2) sugeridos pelas Tabelas 2 e 3, para uma amostra de 50 a 100 observações e até 10% de censura, tem-se que não é adequado considerar um modelo contínuo (optando ou não pela metodologia de dados grupados) para o ajuste deste conjunto de dados. De fato, como pode ser visto na Figura 5, a divergência entre as estimativas da função de sobrevivência do modelo discreto e dos modelos contínuos foram grandes ($DAM_{\text{contínuo}} = 0{,}0940$ e $DAM_{\text{dados grupados}} = 0{,}0818$).

4.3. Aplicação 3

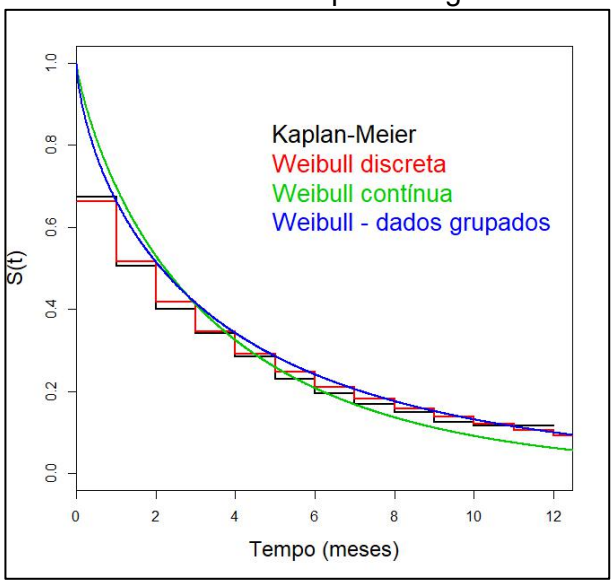
A Tabela 6 apresenta dados do tempo para a gravidez de casais que desejam ter uma criança (AALEN, 1987; TIETZE, 1968). Nesta aplicação, a variável resposta T representa o número de meses (completos) entre a interrupção do uso de qualquer tipo de contraceptivo e a gravidez. Os casais foram acompanhados por 12 meses. Note que, neste caso, $T = 0$ indica que a mulher engravidou antes de completar um mês após a interrupção do uso de contraceptivos.

Tabela 6 – Tempo (em meses) até a gravidez ($n = 611$).

Tempo*	Frequência	
	Observações não censuradas	Observações censuradas
0	199	0
1	103	0
2	64	0
3	36	12
4	33	7
5	30	9
6	18	5
7	13	5
8	9	5
9	10	5
10	3	2
11	0	5
12	0	38

Fonte: Aalen (1987, p. 20).

Figura 6 – Estimativas da função de sobrevivência do tempo até a gravidez.



Fonte: Dados da Tabela 6.

As distribuições Weibull discreta e lognormal discreta apresentam um bom ajuste dos dados ($EAM_{\text{Weibull discreta}} = 0{,}0119$; $EAM_{\text{lognormal discreta}} = 0{,}0099$) nesta aplicação. Devido à sua popularidade e maior facilidade computacional, foi adotada aqui a distribuição Weibull para os ajustes dos modelos discreto e contínuos.

A amostra é composta de $n=$ 611 observações, das quais 93 (15,2%) são censuradas. A proporção de empates observada na amostra é $pe=0{,}83$. Visto que o valor da pe está entre os limites inferiores (Cortes 1) e superiores (Cortes 2) sugeridos pelas Tabelas 2 e 3, para uma amostra com mais de 200 observações e com 10% a 20% de censura, chega-se a um resultado inconclusivo. Desta forma, a princípio não é possível decidir entre permitir ou rejeitar o uso de modelos contínuos para a análise desses dados apenas com informações do tamanho da amostra, do percentual de censuras e da proporção de empates. No entanto, a decisão sobre o uso de modelos contínuos para a análise dos dados da Tabela 6 poderá ser tomada após o ajuste dos modelos contínuos (com ou sem a metodologia de dados grupados) e o cálculo de suas divergências em relação ao modelo discreto. A Figura 6 apresenta as estimativas da função de sobrevivência para os dados da Tabela 6. Observa-se que as divergências entre a estimativa obtida pelos modelos contínuos com aquela obtida pelo modelo discreto foram grandes ($DAM_{\text{contínuo}} =$ 0,0694 e $DAM_{\text{dados grupados}} =$ 0,0657), indicando que não é adequado considerar um modelo contínuo (optando ou não pela metodologia de dados grupados) para o ajuste desse conjunto de dados.

5. Considerações Finais

O uso de modelos discretos não é uma prática comum na área de análise de sobrevivência, devido à escassez de trabalhos e procedimentos computacionais que abordam modelos discretos na presença de censuras. Desta forma, a possibilidade de analisar um conjunto de dados discretos por meio de um modelo contínuo sem dúvida traz facilidade à análise desses dados. No entanto, os resultados obtidos neste trabalho corroboram com aqueles apresentados por Nakano e Carrasco (2006), que mostraram que nem sempre é adequado o uso de modelos contínuos em dados discretos. A principal consequência é a divergência entre as estimativas da função de sobrevivência obtidas pelos modelos discreto e contínuo. As simulações realizadas mostraram que o tamanho da amostra, o percentual de censura e a quantidade de observações empatadas na amostra influenciam nessa divergência.

Este trabalho propôs guias que podem ajudar um pesquisador a decidir sobre o uso de um modelo contínuo na análise de dados discretos de sobrevivência. Os guias propostos têm como base dois pontos de cortes e sugerem que o uso de um modelo contínuo possa ser considerado quando $pe <$ Corte 1 e seu uso deva ser descartado quando $pe >$ Corte 2. Ademais, se Corte 1 $\leq pe \leq$ Corte 2, o diagnóstico é inconclusivo. Neste caso, o pesquisador pode decidir sobre o uso de um modelo contínuo com base na precisão das estimativas obtidas. Os resultados mostraram que os pontos de corte são maiores à medida que o tamanho da amostra aumenta e o percentual de censura diminui. Isto é, quanto maior for o número de observações não censuradas, maior será a proporção de empates que um modelo contínuo suporta. Ademais, a metodologia de

dados grupados apresenta pontos de corte menores para amostras pequenas com baixo percentual de censura e pontos de corte maiores para amostras grandes com alto percentual de censura. Além disso, é possível notar que os pontos de corte não são influenciados pela distribuição adotada e/ou forma da função de risco, visto que os resultados obtidos pelas distribuições Weibull e lognormal foram similares.

Agradecimentos

Os autores agradecem à Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) pelo apoio financeiro para a realização deste trabalho. Processo n. 00193-00002161/2018-19.

Referências

AALEN, O. O. Two examples of modelling heterogeneity in survival analysis. **Scandinavian Journal of Statistics**, v. 14, n. 1, p. 19-25, 1987.

ALMALKI, S. J.; NADARAJAH, S. A new discrete modified Weibull distribution. **IEEE Transactions on Reliability**, v. 63, n. 1, p. 68-80, 2014. DOI: https://doi.org/10.1109/TR.2014.2299691.

BAKOUCH, H. S.; JAZI, M. A.; NADARAJAH, S. A new discrete distribution. **Statistics: A Journal of Theoretical and Applied Statistics**, v. 48, n. 1, p. 200-240, 2014. DOI: https://doi.org/10.1080/02331888.2012.716677.

BIAZATTI, E. C. **Modelo de regressão log Weibull com fração de cura para dados**. 2017. 60 f. Dissertação (Mestrado em Estatística) – Programa de Pós-Graduação em Estatística, Universidade de Brasília, Brasília/DF, 2017.

CHALITA, L. V. A. S.; COLOSIMO, E. A.; DEMÉTRIO, C. B. G. Likelihood Approximations and discrete models for tied survival data. **Communications in Statistics: Theory and Methods**, v. 31, n. 7, p. 1215-1229, 2002.

CORRÊA, J. B.; COSTA, L. O.; OLIVEIRA, N. T.; LIMA, W. P.; SLUKA, K. A.; LIEBANO, R. E. Effects of the carrier frequency of interferential current on pain modulation and central hypersensitivity in people with chronic nonspecific low back pain: A randomized placebo-controlled trial. **European Journal of Pain**, v. 20, n. 10, p. 1653-1666, 2016. DOI: https://doi.org/10.1002/ejp.889.

EFRON, B. Logistic regression, survival analysis and the Kaplan-Meier Curve. **Journal of the American Statistical Association**, v. 83, n. 402, p. 414-425, 1988. DOI: https://doi.org/10.2307/2288857.

HASHIMOTO, E. M. **Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados**. 2008. 121 f. Tese (Doutorado em Estatística e Experimentação Agronômica) – Programa de Pós-Graduação em Estatística e Experimentação Agronômica, Universidade de São Paulo, Piracicaba/SP, 2008.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457-481, 1958. DOI: https://doi.org/10.2307/2281868.

NAKAGAWA, T.; OSAKI, S. The discrete Weibull distribution. **IEEE Transactions on Reliability**, v. 24, n. 5, p. 300-301, 1975. DOI: https://doi.org/10.1109/TR.1975.5214915.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. **TEMA: Tendências em Matemática Aplicada e Computacional**, v. 7, n. 1, p. 91-100, 2006.

NEKOUKHOU, V; BIDRAM, H. The exponentiated discrete Weibull distribution. **Sort**, v. 39, n. 1, p. 127-146, 2015.

RINNE, H. **The Weibull distribution: a handbook**. Boca Raton: Taylor&Francis, 2008.

SILVA, J. F.; LIEBANO, R. E.; CORRÊA, J. B.; MATSUSHITA, R. Y.; NAKANO, E. Y. Análise do tempo para o alívio da intensidade da dor em pacientes com dor lombar crônica não específica via modelo de riscos proporcionais de Cox. **Ciência e Natura**, v. 39, n. 2, p. 233-243, 2017. DOI: https://doi.org/10.5902/2179460X24102.

TIETZE, C. Fertility after discontinuation of intrauterine and oral contraception. **International Journal of Fertility**, v. 13, n. 4, p. 385-389, 1968.

VILA, R.; NAKANO, E. Y.; SAULO, H. Theoretical results on the discrete Weibull distribution of Nakagawa and Osaki. **Statistics: A Journal of Theoretical and Applied Statistics**, v. 53, n. 2, p. 339-363, 2019. DOI: https://doi.org/10.1080/02331888.2018.1550645.

WEIBULL, W. A statistical distribution function of wide applicability. **Journal of Applied Mechanics**, v. 18, n. 3, p. 293-297, 1951.

APÊNDICE A – Mecanismo para geração de censuras aleatórias da distribuição Weibull discreta

Considere uma variável aleatória T representando o tempo até a falha e C uma outra variável aleatória independente de T, representando o tempo até a censura. Assumindo que $T \sim$ Weibull(α, λ_1) e $C \sim$ Weibull(α, λ_2), tem-se, segundo a parametrização apresentada em (1), que

$$\pi = \mathrm{P}(C < T) = \mathrm{P}(C \leq T) = \lambda_1/(\lambda_1 + \lambda_2).$$

Note que π é a probabilidade de uma observação ser censurada, considerando a distribuição Weibull contínua. Considerando a reparametrização $q_i = \exp\{-i\}$ em (2), tem-se para a distribuição Weibull discreta que

$$\pi = \log(\mathrm{q}_2)/[\log(\mathrm{q}_1) + \log(\mathrm{q}_2)],$$

que resulta em $q_2 = q_1^{\pi/(1-\pi)}$.

Dessa forma, se $T \sim$ Weibull-Discreta(α, q_1) e $C \sim$ Weibull-Discreta(α, q_2) então $\pi \cong \mathrm{P}(C < T)$ quando $q_2 = q_1^{\pi/(1-\pi)}$. Assim, o algoritmo utilizado neste trabalho para gerar tempos segundo uma distribuição Weibull-Discreta(α, q) com censura aleatória é apresentado a seguir.

Algoritmo 1** – **Geração de dados de uma Weibull-Discreta(α, q) com $\pi \times$ 100% de censuras.

Passo 1. Gerar o tempo de falha $T \sim$ Weibull-Discreta(α, q);

Passo 2. Gerar o tempo de censura $C \sim$ Weibull-Discreta(α, q_2), com $q_2 = q_1^{\pi/(1-\pi)}$;

Passo 3. Obter o tempo observado $t = \min\{T, C\}$;

Passo 4. Obter o indicador de censura δ:

Passo 4.1. Gerar $u \sim$ Bernoulli($p = 0{,}5$)

Passo 4.2. Definir o valor de δ

- se $T < C$, então $\delta = 1$;
- se $T > C$, então $\delta = 0$;
- se $T = C$ e $u = 1$, então $\delta = 1$;
- se $T = C$ e $u = 0$, então $\delta = 0$.

Passo 5. Repetir os Passos 1 a 5 até obter a amostra desejada.

APÊNDICE B – Mecanismo para geração de censuras aleatórias da distribuição lognormal discreta

Considere uma variável aleatória T representando o tempo até a falha e C uma outra variável aleatória idependente de T, representando o tempo até a censura. Assumindo que $T \sim$ Lognormal(μ_1, σ^2) e $C \sim$ Lognormal(μ_2, σ^2), tem-se, segundo a parametrização apresentada em (3), que

$$\pi = P(C < T) = P(C \leq T) = P(\log[C] \leq \log[T]) = \Phi[(\mu_1 - \mu_2)/ (2\sigma^2)^{1/2}],$$

que resulta em

$$\mu_2 = \mu_1 - z_{(\pi)}(2\sigma^2)^{1/2}.$$

Aqui, π é a probabilidade de uma observação ser censurada, considerando a distribuição lognormal contínua, $\Phi[.]$ é a função de distribuição acumulada da distribuição normal padrão e $z_{(\pi)}$ é o quantil π da distribuição normal padrão.

Dessa forma, se $T \sim$ Lognormal-Discreta(μ_1, σ^2) e $C \sim$ Lognormal-Discreta(μ_2, σ^2), então $\pi \cong P(C < T)$ quando $\mu_2 = \mu_1 - z_{(\pi)}(2\sigma^2)^{1/2}$. Assim, o algoritmo utilizado neste trabalho para gerar tempos segundo uma distribuição Lognormal-Discreta(μ, σ^2) com censura aleatória é apresentado a seguir.

Algoritmo 2 – Geração de dados de uma Lognormal-Discreta(μ, σ^2) com $\pi \times 100\%$ de censuras.

Passo 1. Gerar o tempo de falha $T \sim$ Lognormal-Discreta(μ, σ^2);

Passo 2. Gerar o tempo de censura $C \sim$ Lognormal-Discreta(μ_2, σ^2), com $\mu_2 = \mu - z_{(\pi)}(2\sigma^2)^{1/2}$;

Passo 3. Obter o tempo observado $t = \min\{T, C\}$;

Passo 4. Obter o indicador de censura δ:

Passo 4.1. Gerar $u \sim$ Bernoulli($p = 0{,}5$)

Passo 4.2. Definir o valor de δ

- se $T < C$, então $\delta = 1$;
- se $T > C$, então $\delta = 0$;
- se $T = C$ e $u = 1$, então $\delta = 1$;
- se $T = C$ e $u = 0$, então $\delta = 0$.

Passo 5. Repetir os Passos 1 a 5 até obter a amostra desejada.