Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Intervalos de confiança bootstrap para algumas estatísticas não paramétricas em dados censurados

por

Tulio Paixão Santos

Brasília, 6 de junho de 2022

Intervalos de confiança bootstrap para algumas estatísticas não paramétricas em dados censurados

por

Tulio Paixão Santos

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Brasília, 6 de junho de 2022

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Prof. Dr. Eduardo Yoshio Nakano

Orientador, EST/UnB

Prof. Dr. André Luiz Fernandes Cançado

EST/UnB

Profa. Dra. Giovana Oliveira Silva

DEst/UFBA

Eu não vim até aqui pra desistir agora.

(Engenheiros do Hawaii)

Dedico este trabalho a mim mesmo.

Agradecimentos

Agradeço a todos que contribuíram para a conclusão deste trabalho, de forma direta e indireta.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

O objetivo deste trabalho é obter intervalos de confiança para algumas quantidades não paramétricas em análise de sobrevivência utilizando o método de reamostragem bootstrap e considerando o estimador de Kaplan-Meier como referência. Os intervalos de confiança foram obtidos para as estimativas de Função de Sobrevivência, Função de Risco Acumulado, quantis, Vida Média e Função de Vida Residual Média. O desempenho do intervalo proposto foi avaliado por meio da probabilidade de cobertura estimada. Por fim, foi ilustrada uma aplicação da metodologia proposta através de uma aplicação de dados reais, comparando os intervalos de bootstrap com os obtidos por métodos assintóticos.

Palavras-Chave: Análise de Sobrevivência; Bootstrap; Intervalo de Confiança; Reamostragem; Dados Censurados.

Abstract

The aim of this work is to obtain confidence intervals for some nonparametric quantities in survival analysis using the method of bootstrap resampling and considering Kaplan-Meier estimator as a reference. The confidence intervals were obtained for the estimates of Survival Function, Accumulated Risk Function, quantiles, Mean Life and Mean Residual Life Function. The performance of proposed interval was evaluated through the estimated probability of coverage. Finally, an application of the proposed methodology was illustrated through a real data application, comparing bootstrap intervals with those obtained by assymptotics methods.

Keywords: Survival analysis; Bootstrap; Confidence Interval; Resampling; Censored Data.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas duas décadas do século passado. A razão deste crescimento é o desenvolvimento e aprimoramento de técnicas estatísticas combinados com computadores cada vez mais velozes. Uma evidência quantitativa deste sucesso é o número de aplicaçoes de análise de sobrevivência em medicina (Colosimo e Giolo, 2006).

O objeto de estudo desta área é o tempo até a ocorrência de um certo evento, como por exemplo, o tempo de duração de uma lâmpada, o tempo de vida que pessoas têm depois do surgimento de alguma doença, o tempo gasto para que um remédio faça o efeito desejado. Estes são alguns exemplos da aplicabilidade da análise de sobrevivência.

Uma das medidas mais importantes nesse ramo de estudo é a censura. Ela pode ser definida como a ausência da ocorrência do evento no tempo de análise, por motivos diversos. Um exemplo em que isso ocorre é quando um paciente está sendo estudado após o surgimento de uma doença, mas morre por outra causa. Então a morte do paciente ocorreu por motivos externos e o dado do indivíduo é censurado. Neste trabalho será considerado apenas o caso de censura mais comum, conhecido como censura à direita, para a qual é possível aproveitar-se a informação do tempo do objeto em estudo até a ocorrência do evento aleatório.

A análise de sobrevivência pode ser paramétrica (com o uso de distribuiçoes de probabili-

dade) e não-paramétrica (sem o uso de distribuiçoes de probabilidade). Os estimadores não-paramétricos mais comuns na análise de sobrevivência são os de Nelson-Aalen e Kaplan-Meier e serão utilizados neste trabalho.

Neste contexto, a proposta deste trabalho é estimar o intervalo de confiança dos principais estimadores da análise de sobrevivência por meio de técnicas de reamostragem bootstrap, proposto inicialmente por Efron (1979). Estudos de simulação serão apresentados comparando os resultados obtidos pelos estimadores assintóticos presentes na literatura com aqueles obtidos pelo bootstrap. O desempenho dos intervalos propostos é avaliado por meio de sua probabilidade de cobertura. Todos os cálculos e simulaçoes realizados neste trabalho foram feitos por meio do software livre R (R Core Team, 2021).

Capítulo 2

Conceitos em Análise de Sobrevivência

2.1 Censura

Em análise de sobrevivência, um dos problemas mais comuns está relacionado à medição da variável resposta, que é temporal e não pode ser medida instantaneamente e independente do tamanho da resposta. Valores altos de tempo podem não ser medidos tão facilmente por uma série de motivos: às vezes o paciente precisa abandonar o estudo antes do fim; o indivíduo falece no meio de um tratamento; o indivíduo muda de localidade. Em indústrias de eletrônicos onde o interesse é verificar o tempo de sobrevivência de um certo aparelho, o tempo gasto para essa verificação pode causar prejuízos à indústria.

Nesses casos, apesar de não ser possível observar o tempo até a ocorrência do evento de interesse, tem-se o tempo até o abandono do estudo. Essa informação é muito útil para a análise de sobrevivência.

Desse modo, existe a necessidade de se criar uma variável dicotômica δ que indica se o indivíduo foi ou não observado. Essa variável é definida como censura, e assume o valor 1 se o tempo de sobrevida for observado, e 0 se for censurado, isto é,

$$\delta = \begin{cases} 0, \text{ se o tempo foi censurado} \\ 1, \text{ se o tempo foi observado.} \end{cases}$$

A censura pode ser à direita, à esquerda ou intervalar. Neste trabalho iremos tratar apenas de censura à direita. Os três tipos de censura à direita são:

- Censura do Tipo I: neste tipo de censura, o tempo de duração do experimento é pré-determinado. O tempo de falha é, então, incompleto se os elementos em estudo não falharem. Portanto, o objeto é censurado se não falhar até o tempo pré-estabelecido.

- Censura do Tipo II: ocorre quando é determinado um número k ($k \leq n$) de elementos a falharem no experimento. O pesquisador observa os elementos de estudo até que as k falhas aconteçam. o valor n se refere ao tamanho da amostra em estudo.

- Censura aleatória: é a mais geral e engloba as demais, pode acontecer quando um ou mais componentes não puderem ser acompanhados até o final do experimento ou ainda quando estes falharem por motivos distintos do interesse do estudo. Esse tipo de censura é a mais comum e acontece de forma natural, sem a manipulação do pesquisador.

2.2 Representação do Tempo de Sobrevivência

Na análise de Sobrevivência, o comportamento da variável aleatória que descreve o tempo de sobrevivência de determinado objeto é especificado pela função de sobrevivência ou pela função de risco. Se a função de sobrevivência é conhecida, a função de risco pode ser facilmente especificada. Este capítulo apresenta a definição do tempo de sobrevivência, a função de risco, função de risco acumulado, percentis, tempo de vida médio e vida média residual.

2.2.1 Variáveis aleatórias contínuas

A seguir serão listadas as funções utilizadas na análise de sobrevivência, para as variáveis aleatórias contínuas não-negativas ($T \geq 0$).

Função densidade de probabilidades

Seja T uma variável aleatória não-negativa contínua. A função Densidade de Probabilidades de T, $f(t)$, é uma função que satisfaz as seguintes condições (Meyer, 1983):

i) $f(t) \geq 0, \forall\, t \geq 0$;

ii) $\int_0^\infty f(t)\, dt$ = 1;

iii) $P(a \leq T \leq b)$ = $\int_a^b f(t)$ dt ,$\forall\ 0 \leq a < b$.

Função de Sobrevivência

Seja T uma variável aleatória não-negativa contínua. A função de Sobrevivência ($S(t)$) é definida como a probabilidade da observação em estudo não falhar até o tempo t, ou seja, a probabilidade de sobrevivência além de t. Esta função é definida por

$$S(t) = P[T > t] = \int_t^\infty f(t)\, dt, t \geq 0. \tag{2.1}$$

A função de sobrevivência também pode ser utilizada para se determinar os quantis do tempo de sobrevivência. Seja t_p o p-ésimo quantil da variável aleatória contínua T, i.e., $P[T \leq t_p]$ = p. Assim,

$$t_p = S^{-1}(1-p),\ \forall\ 0 < p < 1. \tag{2.2}$$

Função de Risco (ou Taxa de Falha)

Conhecida como Função de risco (ou taxa de falha), e denotada por $h(t)$, esta função representa o risco instantâneo que o indivíduo tem de experimentar o evento de interesse em um determinado tempo t. No caso de uma variável aleatória contínua, esta função é definida como a razão do limite da probabilidade condicional de um indivíduo experimentar o evento de interesse no intervalo de tempo $[t, t + \Delta t)$ dado que o mesmo não tenha experimentado o evento de

interesse antes de t, sobre o intervalo de tempo Δt . A função $h(t)$ é expressa por

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.3}$$

Em se tratando de variáveis aleatórias contínuas, $h(t)$ assume valores reais positivos e essa função não tem limite superior. A função $h(t)$ descreve como o risco (taxa de falhar) se modifica com o passar do tempo. Por esse motivo, essa função é muito utilizada para descrever o comportamento do tempo de sobrevivência. Alguns autores consideram que a taxa de falha é mais informativa que a função de sobrevivência, pois diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de risco podem diferir drasticamente (Colosimo e Giolo, 2006).

Função de Risco Acumulado (ou Taxa de Falha Acumulada)

A função de Risco Acumulado (ou taxa de Falha Acumulada), denotada por $H(t)$, fornece o risco acumulado do indivíduo no tempo t. Para variáveis aleatórias contínuas, esta função é definida por

$$H(t) = \int_0^t h(u)\, du, t \geq 0. \tag{2.4}$$

Relações entre as funções anteriores

Como existe uma relações entre as funções citadas acima, dada uma dessas funções, as demais podem ser obtidas. Abaixo são apresentadas as relações entre as funções $f(t), S(t), h(t)$ e $H(t)$, para o caso contínuo.

A função de risco pode ser definida como

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.5}$$

A função densidade de probabilidade de T, $f(t)$, é obtida a partir da derivação da função de distribuição acumulada $F(t)$, ou seja,

$$f(t) = \frac{d}{dt}F(t). \tag{2.6}$$

Como $F(t) = 1 - S(t)$, então

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t). \tag{2.7}$$

A partir das equações anteriores, tem-se também que

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t). \tag{2.8}$$

Integrando ambos os termos na equação acima, obtém-se que

$$\log S(t) = -H(t), \tag{2.9}$$

ou seja,

$$S(t) = \exp[-H(t)]. \tag{2.10}$$

Além disso, tem-se também que

$$f(t) = h(t)\exp[-H(t)]. \tag{2.11}$$

Momentos e Vida Média Residual

Os momentos da variável aleatória T e a Vida Média Residual são quantidades de interesse na área de sobrevivência, que podem ser obtidas através da função de sobrevivência.

Se T é uma variável aleatória contínua não-negativa, então seu r-ésimo momento, se existir,

pode ser obtido por (James, 2015)

$$E(T^r) = r \int_0^\infty t^{r-1} S(t)\, dt,\ \forall\, r \geq 1. \tag{2.12}$$

Com base na equação acima, a média e variância do tempo de vida são dadas, respectivamente, por

$$E(T) = \int_0^\infty S(t)\, dt \tag{2.13}$$

e

$$Var(T) = 2 \int_0^\infty tS(t)\, dt - \left[\int_0^\infty S(t)\, dt \right]^2. \tag{2.14}$$

A Vida Média Residual é uma função que representa a expectativa média de vida de um indivíduo que não falhou até o tempo t, e pode ser escrita como

$$v(t) = E[T - t | T \geq t] = \frac{1}{S(t)} \int_t^\infty S(u)\, du,\ \forall\, t \geq 0. \tag{2.15}$$

2.2.2 Variáveis aleatórias discretas

A seguir serão listadas as funções utilizadas na análise de sobrevivência, para as variáveis aleatórias discretas não-negativas.

Função de probabilidades

Seja T uma variável aleatória discreta que assume valores inteiros não-negativos, i.e, t=0,1,2,.... A função de Probabilidades de T, $p(t)$, é uma função que satisfaz

i) $p(t) \geq 0$, $\forall\, t$ =0,1,2,...;

ii) $\sum_{t=0}^{\infty} p(t)$ = 1.

Função de Sobrevivência

Seja T uma variável aleatória discreta cujos possíveis valores são os inteiros não negativos. A probabilidade de uma observação não falhar até o tempo t é expressa por

$$S(t) = P[T > t] = \sum_{k=t+1}^{\infty} p(k) = \sum_{k=t+1}^{\infty} P(T = k),\ t = 0, 1, 2, ... \tag{2.16}$$

A função $S(t)$ pode ser entendida como uma função escada que decresce onde t tem probabilidade positiva, e é constante nos demais pontos.

Função de Risco (ou Taxa de Falha)

A Função de Risco (ou Função Taxa de Falha) representa a probabilidade condicional da observação experimentar o evento de interesse no tempo t, dado que não experimentou o evento antes de t, e é definida, no caso discreto, por

$$h(t) = P(T = t|T \geq t),\ t = 0, 1, 2, ... \tag{2.17}$$

É importante observar que no caso de variáveis aleatórias discretas, $h(t)$ é limitada no intervalo [0,1], pois se trata de uma probabilidade condicional.

Função de Risco Acumulado (ou Taxa de Falha Acumulada)

A Função de Risco Acumulado ($H(t)$), fornece o risco acumulado do indivíduo no tempo t. Além disso, esta função assume valores reais positivos e não tem limite superior, e é definida por

$$H(t) = \sum_{k=0}^{t} h(k),\ t = 0, 1, 2, ... \tag{2.18}$$

Relações entre as funções anteriores

Assim como no caso em que as variáveis aleatórias são contínuas, no caso discreto também pode-se usar algumas funções para encontrar as demais utilizadas em análise de sobrevivência. A seguir serão apresentadas algumas relações importantes.

A Função de Risco pode ser expressa por

$$h(t) = \frac{p(t)}{p(t) + S(t)},\ t = 0, 1, 2, ... \tag{2.19}$$

Da equação acima, $p(t)$ pode ser escrita como

$$p(t) = \frac{h(t)}{1 - h(t)} S(t),\ t = 0, 1, 2, ... \tag{2.20}$$

A função de probabilidade ainda pode ser escrita da seguinte forma

$$p(t) = \begin{cases} 1 - S(0), \text{se t=0} \\ S(t-1) - S(t), \text{se t=1,2,...} \end{cases} \tag{2.21}$$

A Função de Sobrevivência pode ser escrita da seguinte forma

$$S(t) = S(0) \prod_{k=1}^{t} \frac{S(k)}{S(k-1)} = \prod_{k=0}^{t} [1 - h(k)],\ t = 0, 1, 2, ... \tag{2.22}$$

É possível também escrever a Função de Probabilidade como

$$p(t) = \frac{h(t)}{1 - h(t)} = \prod_{k=0}^{t} [1 - h(k)],\ t = 0, 1, 2, ... \tag{2.23}$$

Momentos e Vida Média Residual

Seja T uma variável aleatória discreta não-negativa. Se existir, o r-ésimo momento dessa variável aleatória é definido por:

$$E(T^r) = \sum_{t=0}^{\infty} t^r p(t), \ \forall \ r \geq 1. \tag{2.24}$$

O primeiro e segundo momentos da variável aleatória T podem ser escritos, respectivamente, como

$$E(T) = \sum_{t=0}^{\infty} S(t) \tag{2.25}$$

e

$$E(T^2) = E(T) + 2\sum_{t=1}^{\infty} tS(t). \tag{2.26}$$

A Função Vida Média Residual para uma variável aleatória discreta T é definida por

$$v(t) = E[T - t|T \geq t] = \frac{1}{p(t) + S(t)} \sum_{k=t}^{\infty} S(k), \ t = 0, 1, 2, ... \tag{2.27}$$

Capítulo 3

Estimativa não paramétrica

Neste capítulo serão apresentados métodos não paramétricos para estimar as principais estimativas em análise de sobrevivência.

3.1 Estimação para dados não censurados

A seguir serão apresentados precedimentos para estimativas da funções citadas acima, para dados sem a presença de censura.

3.1.1 Função de Probabilidades

Seja T uma variável aleatória contínua não-negativa. A Função Densidade de Probabilidades pode ser estimada por (Nakano, 2017)

$$\hat{f(t)} = \frac{\text{nº de indivíduos que experimentaram o evento de interesse em } [t, t+\Delta t)}{\Delta t \times (\text{nº total de indivíduos})}, \qquad (3.1)$$

em que Δt é a amplitude do intervalo.

Se T for uma variável aleatória discreta, a Função de Probabilidades é estimada por

$$\hat{p(t)} = \frac{\text{nº de indivíduos que experimentaram o evento de interesse em } t}{\text{nº total de indivíduos}}, t = 0, 1, 2, ... \quad (3.2)$$

3.1.2 Função de Sobrevivência

A Função de Sobrevivência pode ser estimada como a proporção de indivíduos que não falharam até o tempo t, e é estimado por

$$\hat{S(t)} = \frac{\text{nº de indivíduos com tempo de sobrevivência maior que } t}{\text{nº total de indivíduos}}, \forall\, t \geq 0. \quad (3.3)$$

Pode haver o interesse de ordenar os valores em análise de sobrevivência obsevados na amostra. Seja $t_{(j)}$ o j-ésimo valor ordenado da amostra observada, j=1,2,...,k. Considerando $t_{(0)}$=0 e $t_{(k+1)}$=∞, tem-se que

$$0 = t_{(0)} < \min\{t_1, t_2, ..., t_n\} = t_{(1)} < t_{(2)} < ... < t_{(k)} = \max\{t_1, t_2, ..., t_n\} < t_{(k+1)} = \infty.$$

É importante observar que $k \leq n$ (no caso em que a amostra observada não apresenta empates, tem-se que $k = n$), sendo n o tamanho da amostra.

Assim, o estimador de $S(t)$ pode ser escrito como

$$\hat{S(t)} = \frac{n_j - d_j}{n}, \forall\, t_{(j)} \leq t < t_{(j+1)}, j = 0, 1, ..., k \quad (3.4)$$

em que n_j é o número de indivíduos que estão sob risco no tempo $t_{(j)}$ e d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=0,1,...,k. Assim, a função de sobrevivência é como uma função escada com degraus nos tempos de falha observados de tamanho d_j/n.

3.2 Estimação para dados censurados

A seção anterior trouxe estimativas para dados não censurados. Nesta seção, serão apresentadas estimativas para a função de sobrevivência e risco acumulado, para dados com censura à direita. Serão apresentados os estimadores não paramétricos de Kaplan-Meier (Kaplan e Meier, 1958) e o estimador de Nelson-Aalen (Nelson, 1972) (Aalen, 1978).

3.2.1 Estimador de Kaplan-Meier

O estimador não paramétrico de Kaplan é um estimador da função de sobrevivência bastante popular e o mesmo se encontra presente nos principais softwares estatísticos. Este estimador é também conhecido na literatura como estimador limite-produto e é definido como (Kaplan e Meier, 1958)

$$\hat{S}_{KM}(t) = \prod_{j:t_{(j)} \leq t} \left[1 - \frac{d_j}{n_j}\right], \tag{3.5}$$

em que n_j é o número de indivíduos que estão sob risco no tempo $t_{(j)}$ e d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=0,1,...,k.

Se uma variável aleatória tem distribuição contínua, foi visto no Capítulo 1 que a relação entre a Função de Sobrevivência e a Função de Risco Acumulado é $S(t) = \exp[-H(t)]$. Isolando $H(t)$, o estimador de Kaplan-Meier da Função de Risco Acumulado é dado por

$$\hat{H}_{KM}(t) = -\log \hat{S}_{KM}(t), \tag{3.6}$$

sendo $\hat{S}_{KM}(t)$ o estimador de Kaplan-Meier da Função de Sobrevivência.

Intervalo de Confiança Simétrico para S(t)

Todo estimador está sujeito a variações amostrais, e essas variações podem ser expressas pela variância. A variância assintótica do estimador de Kaplan-Meier da Função de Sobrevi-

vência pode ser estimada pela Fórmula de Greenwood (Greenwood, 1926), dada por

$$\hat{Var}(\hat{S}_{KM}(t)) = \hat{S}^2_{KM}(t) \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}, \tag{3.7}$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da Função de Sobrevivência, n_j é o número de indivíduos que estão sob risco no tempo $t_{(j)}$ e d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=0,1,...,k.

Agora, é possível obter um intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$, que é expresso por

$$I.C.[S(t)] : \left[\hat{S}_{KM}(t) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{S}_{KM}(t))}; \hat{S}_{KM}(t) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{S}_{KM}(t))}\right], \tag{3.8}$$

em que $z_{(1-\frac{\alpha}{2})}$ é o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Os intervalos simétricos para a função de sobrevivência de Kaplan-Meier, descritos em (3.8), podem ser obtidos no software livre R, com o comando $survfit$ da biblioteca $survival$.

Intervalo log baseado na Função de Risco Acumulado

Outra maneira de se obter um intervalo de confiança para $S(t)$ baseia-se na Função de Risco Acumulado, através da equivalência $H(t) = -\log S(t)$. A variância estimada para o estimador de $H(t)$ de Kaplan-Meier é dada por

$$\hat{Var}(\hat{H}_{KM}(t)) = \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{3.9}$$

Portanto, um intervalo com $100(1-\alpha)\%$ de confiança para $H(t)$, $\forall\, t$ fixo, é expresso por

$$I.C.[H(t)] : \left[\hat{H}_{KM}(t) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}; \hat{H}_{KM}(t) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}\right], \tag{3.10}$$

resultando no seguinte intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$

$$I.C.[S(t)] : \left[e^{-z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}}\hat{S}_{KM}(t); e^{z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{KM}(t))}}\hat{S}_{KM}(t)\right], \tag{3.11}$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, $\hat{H}_{KM}(t)$ é o estimador de Kaplan-Meier da Função de Risco Acumulado e $z_{(1-\frac{\alpha}{2})}$ é o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Os intervalos do tipo $\log$ para a função de sobrevivência de Kaplan-Meier, descritos em (3.11), podem ser obtidos no software livre R, com o comando $survfit$ da biblioteca $survival$.

Intervalo log-log

Os intervalos apresentados acima possuem algumas limitações. Se a estimativa de $S(t)$ estiver próxima de 0 ou de 1, o intervalo simétrico pode apresentar limite inferior negativo ou limite superior maior que 1, respectivamente. Para o intervalo log, o limite superior também pode ser maior que 1. Apesar de ser truncado em 0 ou em 1, esse problema pode ser resolvido aplicando uma transformação alternativa em $S(t)$.

Kalbfleisch e Prentice (2002) sugeriram aplicar o logaritmo na função de risco acumulado, ou a transformação "log-log" na função de sobrevivência de Kaplan-Meier, isto é,

$$\hat{U}_{KM}(t) = \log\{-\log \hat{S}_{KM}(t)\}. \tag{3.12}$$

A variância de $\hat{U}_{KM}(t)$ é expressa por

$$\hat{Var}(\hat{U}_{KM}(t)) = \frac{1}{\left[\log \hat{S}_{KM}(t)\right]^2} \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{3.13}$$

Portanto, um intervalo com $100(1-\alpha)\%$ de confiança para $U(t)$, $\forall\ t$ fixo, é expresso por

$$I.C.[U(t)] : \left[\hat{U}_{KM}(t) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{KM}(t))}; \hat{U}_{KM}(t) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{KM}(t))}\right], \tag{3.14}$$

resultando no intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$, dado por

$$I.C.[S(t)] : \left[(\hat{S}_{KM}(t))^{exp\left\{z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{KM}(t))}\right\}}; (\hat{S}_{KM}(t))^{exp\left\{-z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{KM}(t))}\right\}}\right], \tag{3.15}$$

em que $\hat{S}_{KM}(t)$ é o estimador de Kaplan-Meier da função de sobrevivência, $\hat{U}_{KM}(t)$ é o estimador de Kaplan-Meier da função log-log e $z_{(1-\frac{\alpha}{2})}$ é o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Os intervalos do tipo $\log - \log$ para a função de sobrevivência de Kaplan-Meier, descritos em (3.15), podem ser obtidos no software livre R, com o comando $survfit$ da biblioteca $survival$.

3.2.2 Estimador de Nelson-Aalen

Como para o estimador de Kaplan-Meier, o estimador de Nelson- Aalen também se baseia na relação entre a função de sobrevivência e a função de risco acumulado. Nelson propôs o estimador em 1972 (Nelson, 1972), enquanto Aalen provou as propriedades assintóticas em

1978 (Aalen, 1978). O estimador de Nelson-Aalen é dado por

$$\hat{H}_{NA}(t) = \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j}, \tag{3.16}$$

em que n_j é o número de indivíduos que estão sob risco no tempo $t_{(j)}$ e d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=0,1,...,k.

Como a função de sobrevivência está diretamente relacionada com a função de Risco Acumulado, e estimador de Nelson-Aalen da função de sobrevivência é dado por

$$\hat{S}_{NA}(t) = exp\{-\hat{H}_{NA}(t)\} = exp\left\{-\sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j}\right\}. \tag{3.17}$$

Intervalo de Confiança Simétrico para S(t)

A variância assintótica do estimador de Nelson-Aalen para a Função de Sobrevivência é dada por (Aalen e Johansen, 1978)

$$\hat{Var}(\hat{S}_{NA}(t)) = \hat{S}^2_{NA}(t) \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j^2}, \tag{3.18}$$

em que $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da Função de Sobrevivência, n_j é o número de indivíduos que estão sob risco no tempo $t_{(j)}$ e d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, j=0,1,...,k.

Portanto um intervalo com $100(1-\alpha)\%$ de confiança para $H(t)$, $\forall\, t$ fixo, é expresso por

$$I.C.[S(t)] : \left[\hat{S}_{NA}(t) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{S}_{NA}(t))}; \hat{S}_{NA}(t) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{S}_{NA}(t))}\right], \tag{3.19}$$

sendo $z_{(1-\frac{\alpha}{2})}$ o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Uma observação importante é que para o intervalo simétrico, o limite superior pode ser

maior que 1. Nestes casos, o intervalo deve ser truncado em 1.

Os intervalos simétricos para a função de sobrevivência de Nelson-Aalen, descritos em (3.19), podem ser obtidos no software livre R, com o comando $survfit$ da biblioteca $survival$.

Intervalo log baseado na Função de Risco Acumulado

Outra maneira de se obter um intervalo de confiança para $S(t)$ é baseado na Função de Risco Acumulado, através da equivalência $H(t) = -\log S(t)$. A variância estimada para o estimador de $H(t)$ de Nelson-Aalen é dada por

$$\hat{Var}(\hat{H}_{NA}(t)) = \sum_{j:t_{(j)} \leq t} \frac{d_j}{n_j^2}. \tag{3.20}$$

Portanto, um intervalo com $100(1-\alpha)\%$ de confiança para $H(t)$, $\forall\ t$ fixo, é expresso por

$$I.C.[H(t)] : \left[\hat{H}_{NA}(t) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{NA}(t))}; \hat{H}_{NA}(t) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{NA}(t))}\right], \tag{3.21}$$

resultando no intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$, dado por

$$I.C.[S(t)] : \left[e^{-z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{NA}(t))}}\hat{S}_{NA}(t); e^{z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{H}_{NA}(t))}}\hat{S}_{NA}(t)\right], \tag{3.22}$$

em que $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função de sobrevivência, $\hat{H}_{NA}(t)$ é o estimador de Nelson-Aalen da Função de Risco Acumulado e $z_{(1-\frac{\alpha}{2})}$ é o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Uma observação importante é que para o intervalo $\log$, o limite superior pode ser maior que 1. Nestes casos, o intervalo deve ser truncado em 1.

Os intervalos do tipo $\log$ para a função de sobrevivência de Nelson-Aalen, descritos em (3.22), podem ser obtidos no software livre R, com o comando $survfit$ da biblioteca $survival$.

Intervalo log-log

Os intervalos apresentados acima podem apresentar limite superior maior que 1, e portanto devem ser truncados em 1, já que a função de Sobrevivência é uma probabilidade limitada em [0,1]. Esse problema pode ser resolvido com a transformação "log-log" na função de sobrevivência de Nelson-Aalen, dada por

$$\hat{U}_{NA}(t) = log\{-log\hat{S}_{NA}(t)\}. \tag{3.23}$$

A variância de $\hat{U}_{NA}(t)$ é expressa por

$$\hat{Var}(\hat{U}_{NA}(t)) = \frac{1}{\left[log\hat{S}_{NA}(t)\right]^2} \sum_{j:t_{(j)}\leq t} \frac{d_j}{n_j^2}. \tag{3.24}$$

Portanto, um intervalo com $100(1-\alpha)\%$ de confiança para $U(t)$, $\forall\ t$ fixo, é expresso por

$$I.C.[U(t)] : \left[\hat{U}_{NA}(t) - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{NA}(t))}; \hat{U}_{NA}(t) + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{NA}(t))}\right], \tag{3.25}$$

resultando no intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$, dado por

$$I.C.[S(t)] : \left[(\hat{S}_{NA}(t))^{exp\left\{z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{NA}(t))}\right\}}; (\hat{S}_{NA}(t))^{exp\left\{-z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\hat{U}_{NA}(t))}\right\}}\right], \tag{3.26}$$

em que $\hat{S}_{NA}(t)$ é o estimador de Nelson-Aalen da função de sobrevivência, $\hat{U}_{NA}(t)$ é o estimador de Nelson-Aalen da função $\log - \log$ e $z_{(1-\frac{\alpha}{2})}$ é o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Os intervalos do tipo $\log - \log$ para a função de sobrevivência de Nelson-Aalen, descritos em (3.26), podem ser obtidos no software livre R, com o comando $survfit$ da biblioteca $survival$.

3.2.3 Estimador dos Percentis, Média e Vida Média Residual

Dada a função de sobrevivência, é possível obter estimativas dos percentis do tempo de sobrevivência, tempo de vida médio e vida média residual. A seguir serão apresentandos os procedimentos para o cálculo destas quantidades, utilizando as funções de sobrevivência de Kaplan-Meier e Nelson-Aalen.

Percentis do tempo de sobrevivência

O p-ésimo quantil do tempo de sobrevivência pode ser ser expresso por $t_p = S^{-1}(1-p)$. Então, t_p pode ser obtido por:

$$t_p = inf\{t : S(t) \leq 1-p\}, \text{ para } 0 < p < 1. \tag{3.27}$$

Como a função de sobrevivência tem forma de escada, pode-se obter a estimativa de t_p por interpolação linear. Mas quando o maior tempo da amostra for censurado, a função de sobrevivência não atinge o ponto zero, não sendo possível o cálculo dos quantis próximos a 1. Esta situação ocorre na função de sobrevivência de Nelson-Aalen, que não atinge o zero nem mesmo quando não existem observações censuradas. Para correção desse problema, pode-se limitar a estimativa empírica da função de sobrevivência ao maior tempo da amostra, ou seja, pode-se considerar que $S(t)$=0 para $t \geq \max\{t_1, t_2, ..., t_n\}$.

Portanto, um estimador do p-ésimo quantil, t_p é expresso por

$$\hat{t_p} = t_{(u)} + \frac{[t_{(u+1)} - t_{(u)}][\hat{S}(t_{(u)}) - (1-p)]}{[\hat{S}(t_{(u)}) - \hat{S}(t_{(u+1)})]}, \text{ para } 0 < p < 1, \tag{3.28}$$

em que $[t_{(u)}; t_{(u+1)}]$ é o intervalo que contém o p-ésimo quantil, ou seja, o intervalo que satisfaz $\hat{S}(t_{(u+1)}) < 1-p \leq \hat{S}(t_{(u)})$ e $\hat{S}(t_{(u)})$ é a estimativa empírica da função de sobrevivência em $t_{(u)}$, u=0,1,...,k+1. Tem-se que $t_{(k+1)}$ = max$\{t_1, t_2, ..., t_n\}$ e $\hat{S}(t_{(k+1)})$=0.

Tempo médio de sobrevivência

O tempo médio de sobrevivência pode ser representado pela área sob a função de sobrevivência. Como a estimativa não paramétrica da função de sobrevivência tem formato de escada, a área que traz a média é a soma de áreas de retângulos com base $(t_{(j+1)} - t_{(j)})$ e altura $\hat{S}(t_{(j)})$, $j = 1, 2, ..., k$. Mas quando o maior tempo da amostra for censurado, a função de sobrevivência não atinge o ponto zero e o último retângulo terá área infinita. Para correção desse problema, pode-se limitar a estimativa empírica da função de sobrevivência ao maior tempo da amostra, ou seja, pode-se considerar que $S(t)$=0 para $t \geq \max\{t_1, t_2, ..., t_n\}$.

Portanto, um estimador do tempo médio de sobrevivência é dado por

$$\bar{T} = \sum_{j=0}^{k} (t_{(j+1)} - t_{(j)}) \hat{S}(t_{(j)}), \tag{3.29}$$

em que $\hat{S}(t_{(j)})$ é a estimativa empírica da função de sobrevivência em $t_{(j)}$, $j = 0, 1, ..., k$ e $t_{(k+1)}$ = $\max\{t_1, t_2, ..., t_n\}$.

Considerando o estimador de Kaplan-Meier, a variância de $\bar{T}$ é dada por

$$\hat{Var}(\bar{T}_{KM}) = \frac{d}{d-1} \sum_{j=1}^{k} \left(\left[\sum_{l=j}^{k} (t_{(l+1)} - t_{(l)}) \hat{S}_{KM}(t_{(l)}) \right]^2 \frac{d_j}{n_j(n_j - d_j)} \right). \tag{3.30}$$

Considerando o estimador de Nelson-Aalen, a variância de $\bar{T}$ é dada por

$$\hat{Var}(\bar{T}_{NA}) = \frac{d}{d-1} \sum_{j=1}^{k} \left(\left[\sum_{l=j}^{k} (t_{(l+1)} - t_{(l)}) \hat{S}_{NA}(t_{(l)}) \right]^2 \frac{d_j}{n_j^2} \right). \tag{3.31}$$

Em (3.32) e (3.33), n_j é o número de indivíduos que estão sob risco no tempo $t_{(j)}$, d_j representa o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$, $j = 1, 2, ..., k + 1$, com $t_{(k+1)}$ = $\max\{t_1, t_2, ..., t_n\}$, e d é o número total de observações que não foram censuradas na amostra.

Portanto, um intervalo assintótico com $100(1-\alpha)\%$ de confiança para o tempo médio de sobrevivência, com base na função de sobrevivência de Kaplan-Meier ou de Nelson-Aalen, é dado por

$$I.C.[\mu] : \left[\bar{T} - z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\bar{T})}; \bar{T} + z_{(1-\frac{\alpha}{2})}\sqrt{\hat{Var}(\bar{T})}\right], \tag{3.32}$$

onde $z_{(1-\frac{\alpha}{2})}$ é o quantil $(1-\frac{\alpha}{2})$ da distribuição normal padrão.

Para uma amostra sem dados censurados, considerando a função de sobrevivência de Kaplan-Meier, a estimativa do tempo médio de sobrevivência e sua variância, se reduzem, respectivamente, a

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n} t_i, \tag{3.33}$$

$$\hat{Var}(\bar{T}_{KM}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(T_i - \bar{T})^2 \tag{3.34}$$

Para a função de sobrevivência de Nelson-Aalen, o mesmo não acontece para uma amostra sem dados censurados.

Vida média residual

A vida média residual no tempo t é a expectativa média de vida de uma observação que não falhou até t, e é representada pela área sob a curva de sobrevivência à direita do ponto t, dividido pelo valor da função de sobrevivência no ponto t. A estimativa da função vida média residual é dada por

$$\hat{v}(t) = \frac{1}{\hat{S}(t)}\left[(t_{(u)} - t)\hat{S}(t) + \sum_{j:t_{(j)}\geq t}^{k}(t_{(j+1)} - t_{(j)})\hat{S}(t_{(j)})\right], \ \forall\, t \geq 0, \tag{3.35}$$

sendo $\hat{S}(.)$ a estimativa da função de sobrevivência, $t_{(k+1)}$ = max$\{t_1, t_2, ..., t_n\}$ e u=inf$\{j : t_{(j)} \geq t\}$. Quando $t = 0$, $\hat{v}(t)$ = $\bar{T}$.

A função vida média residual é definida por $v(t) = E(T - t|T \geq t)$. Então, uma forma alternativa de se obter a estimativa $\hat{v}(t)$ é calcular a esperança condicional $E(T|T \geq t)$ e subtrair t, ou seja

$$\hat{v}(t) = E(T|T \geq t) - t, \ \forall \, t \geq 0. \tag{3.36}$$

A variância estimada da vida média residual, $\hat{Var}(\hat{v}(t)) = Var(\hat{E}(T|T \geq t))$, pode ser obtida por meio das expressões (3.32) e (3.33), excluindo da amostra as observações menores que t.

Capítulo 4

Técnica de Reamostragem Bootstrap

4.1 Introdução

O bootstrap é um método de reamostragem proposto por Bradley Efron (Efron, 1979), e seu uso vem ganhando força graças ao avanço computacional, devido ao seu trabalho computacionalmente intensivo.

O termo $bootstrap$ vem de uma expressão da língua inglesa *"to pull oneself up by one's bootstrap"*, cuja ideia é de que uma pessoa pode se livrar de um afogamento puxando pelo cadarço do próprio sapato, ou seja, algo impossível.

O método consiste em coletar reamostras, ou seja, amostras da amostra original. É importante destacar que esta reamostragem é feita com reposição, e as reamostras possuem o mesmo tamanho da amostra original. Para facilitar o entendimento, imagine uma amostra original de tamanho n. A técnica consiste em realizar n sorteios, com reposição, desta amostra original. O procedimento deve ser repetido B vezes, até existirem B amostras bootstrap. Essas B amostras servem para fazer inferências e testar hipóteses acerca do parâmetro de estudo.

4.2 Algoritmo

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória independente e identicamente distribuída com distribuição desconhecida F, ou seja, $X_i \sim F$. Cada X_i tem igual probabilidade, $1/n$, de ser sorteado, sendo n o tamanho da amostra. Considere que θ é o parâmetro de interesse, cujo estimador depende de F, isto é, $\hat{\theta} = f(X_1, X_2, ..., X_n)$. Os passos para se obter B amostras bootstrap consistem em:

1. Gerar uma amostra $x^* = (x_1^*, x_2^*, ..., x_n^*)$ com reposição da amostra original $x_1, x_2, ..., x_n$.

2. Calcular a b-ésima estimativa $\hat{\theta}^{(b)}$ da b-ésima amostra de bootstrap, $b = 1, 2, ..., B$.

4.3 Estimativa bootstrap

Após a realização dos passos descritos na Seção 4.2, a estimativa pontual bootstrap é dada por

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}, \tag{4.1}$$

e a estimativa do erro padrão de $\hat{\theta}$ via bootstrap é o desvio padrão amostral das estimativas de bootstrap $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(B)}$, dado por

$$se(\hat{\theta}^*) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \hat{\theta}^*)^2}. \tag{4.2}$$

4.4 Intervalos de Confiança via bootstrap

Nesta seção, serão mostrados os principais intervalos de confiança bootstrap para o parâmetro de interesse θ. São eles:

- Intervalo bootstrap normal padrão;

- Intervalo bootstrap básico;
- Intervalo bootstrap percentil;
- Intervalo bootstrap t;
- Intervalo bootstrap BCa.

4.4.1 Intervalo bootstrap normal padrão

Seja $\hat{\theta}$ um estimador de θ. Seja também $se(\hat{\theta})$ a estimativa do erro padrão de $\hat{\theta}$. Se $\hat{\theta}$ é uma média amostral e o tamanho a amostra é suficientemente grande, pelo Teorema Central do Limite tem-se que

$$Z = \frac{\hat{\theta} - E[\hat{\theta}]}{se(\hat{\theta})} \sim N(0,1). \tag{4.3}$$

Então, se $\hat{\theta}$ é um estimador não viesado de θ, um intervalo de confiança $100(1-\alpha)\%$ para θ é dado por

$$I.C.[\theta] : [\hat{\theta} - z_{1-\alpha/2} se(\hat{\theta}); \hat{\theta} + z_{1-\alpha/2} se(\hat{\theta})], \tag{4.4}$$

em que $z_{1-\alpha/2}$ é o quantil $(1-\alpha/2)$ de uma distribuição normal padrão. Em bootstrap, a estimativa do erro padrão corresponde à estimativa do erro padrão a partir das réplicas bootstrap, $se(\hat{\theta^*})$, apresentado na Seção 4.2. Assim, o intervalo de confiança é dado por

$$I.C.[\theta] : [\hat{\theta^*} - z_{1-\alpha/2} se(\hat{\theta^*}); \hat{\theta^*} + z_{1-\alpha/2} se(\hat{\theta^*})]. \tag{4.5}$$

4.4.2 Intervalo bootstrap básico

Este intervalo tem como característica transformar a distribuição das réplicas do estimador, pois subtrai o valor observado da estatística.

Seja T é um estimador de θ e a_α tal que

$$P(T - \theta > a_\alpha) = 1 - \alpha \rightarrow P(T - a_\alpha > \theta) = 1 - \alpha.$$

Assim, o intervalo com $100(1 - 2\alpha)\%$ é dado por

$$[t - a_{1-\alpha}; t - a_\alpha], \tag{4.6}$$

em que t é um estimador de T.

Definindo como $\hat{b}_\alpha = \hat{\theta}_\alpha - \hat{\theta}$ o percentil de ordem α de $\hat{\theta}^* - \hat{\theta}$, o limite superior é definido como $\hat{\theta} - \hat{b}_{\alpha/2}$ e o limite inferior é definido como $\hat{\theta} - \hat{b}_{1-\alpha/2}$.

Portanto, o intervalo de confiança bootstrap básico de $100(1 - \alpha)\%$ é dado por

$$I.C.[\theta] : [2\hat{\theta} - \hat{\theta}_{1-\alpha/2}; 2\hat{\theta} - \hat{\theta}_{\alpha/2}], \tag{4.7}$$

em que $\hat{\theta}_\alpha$ é o quantil α das réplicas bootstrap.

4.4.3 Intervalo bootstrap percentil

Para o cálculo do Intervalo de confiança bootstrap percentil, é utilizada a distribuição empírica das réplicas bootstrap como a distribuição de referência. Suponha que $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(B)}$ são as réplicas bootstrap da estatística $\hat{\theta}$. O Intervalo percentil é dado por

$$I.C.[\theta] : [\hat{\theta}_{\alpha/2}; \hat{\theta}_{1-\alpha/2}], \tag{4.8}$$

em que $\theta_{\alpha/2}$ e $\theta_{1-\alpha/2}$ são, respectivamente, os quantis empíricos $\alpha/2$ e $(1-\alpha/2)$ obtidos a partir da amostra bootstrap.

4.4.4 Intervalo bootstrap t

O bootstrap t não usa a distribuição t-Student, mas sim a distribuição amostral de uma estatística gerada por reamostragem. Portanto, um intervalo de confiança bootstrap t de $100(1-\alpha)\%$ é dado por

$$I.C.[\theta] : [\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}); \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta})], \tag{4.9}$$

em que $\hat{se}(\hat{\theta})$, $t^*_{1-\alpha/2}$ e $t^*_{\alpha/2}$ são obtidos através dos passos descritos a seguir:

- Calcular a estatística observada $\hat{\theta}$.
- Realizar a amostragem com reposição de x para obter a b-ésima amostra bootstrap $x^{(b)} = (x^{(b)}_1, ..., x^{(b)}_n)$.
- Calcular $\hat{\theta}$ da b-ésima amostra boostrap $x^{(b)}$.
- Calcular ou estimar e erro-padrão $\hat{se}(\hat{\theta}^{(b)})$ de cada amostra bootstrap.
- Calcular $t^{(b)} = \frac{\hat{\theta}^{(b)} - \hat{\theta}}{\hat{se}(\hat{\theta}^{(b)})}$.
- Encontrar os quantis $t^*_{1-\alpha/2}$ e $t^*_{\alpha/2}$ da amostra ordenada de $t^{(b)}$.
- Calcular $\hat{se}(\hat{\theta})$, ou seja, o desvio padrão amostral das réplicas $\hat{\theta}^{(b)}$.
- Computar os limites de confiança: $[\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}); \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta})]$.

4.4.5 Intervalo bootstrap BCa

O intervalo bootstrap BCa é o melhor intervalo do bootstrap, pois corrige o viés e a assimetria. Estes intervalos são uma versão modificada dos intervalos percentis. Um intervalo de confiança bootstrap BCa de $100(1-\alpha)\%$ é dado por (Efron e Tibshirani, 1994)

$$I.C.[\theta] : [\hat{\theta}_{\alpha_1}; \hat{\theta}_{\alpha_2}] \tag{4.10}$$

em que

$$\alpha_1 = \phi^{-1}\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right),$$

$$\alpha_2 = \phi^{-1}\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right),$$

$$z_\alpha = \phi^{-1}(\alpha),$$

$$\hat{z}_0 = \phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\{\hat{\theta}^{(b)} < \theta\}\right),$$

e

$$\hat{a} = \frac{\sum_{i=1}^{n}(\bar{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6(\sum_{i=1}^{n}(\bar{\theta}_{(.)} - \hat{\theta}_{(i)})^2)^{3/2}},$$

em que $\bar{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$ e $\hat{\theta}_{(i)}$ é a estimativa de θ sem a observação x_i.

Capítulo 5

Reamostragem com dados censurados

Este capítulo apresenta um estudo de simulações para verificar as probabilidades de cobertura dos intervalos de confiança apresentados para diferentes tamanhos de amostras e percentuais de censura.

5.1 Reamostragem bootstrap para dados censurados

Na Análise de Sobrevivência, todas as estimativas bootstrap que foram descritas no Capítulo 4 são aplicadas no caso de dados não censurados. Para dados censurados, ou seja, na forma $\{(x_1, \delta_1), ..., (x_n, \delta_n)\}$, em que x_j é a j-ésima observação censurada ou não, com seu respectivo indicador de censura δ_j (aqui, δ_j=1 se x_j não for censurado e δ_j=0 se for censurado à direita). O procedimento bootstrap é análogo em comparação ao caso de dados não censurados, desta vez levando em conta que os dados estão em pares (x_j, δ_j). Estudos que realizam reamostragem bootstrap para dados de sobrevivência podem ser vistos em Carrasco et al. (2012) e Carrasco e Nakano (2016), dentre outros. Serão retiradas B amostras bootstrap $(X^{(b)}, \delta^{(b)})$, $b = 1, 2, ..., B$, sendo essas amostras retiradas pelo método de amostragem aleatória simples e com reposição. Para cada amostra bootstrap, estima-se uma estatística de interesse, $\hat{\theta}^{(b)}$.

5.2 Exemplo para a distribuição exponencial

Nesta seção serão apresentados estudos simulados com a intenção de verificar se a probabilidade de cobertura dos intervalos de confiança construídos está próxima de seu nível de confiança, ou seja, se os intervalos construídos tem 95% de confiança, espera-se que em 95% das vezes o intervalo contenha o verdadeiro valor do parâmetro de interesse. Para geração de dados com censura aleatória, será adotada a metodologia proposta por Oliveira (2021), descrita na sequência.

Considere uma variável aleatória T, que representa o tempo de falha, e uma variável aleatória C, que representa a censura. $T \sim Exp(\lambda_1)$ e $C \sim Exp(\lambda_2)$. Considere também que T e C são independentes. A proporção de censura p desejada pode ser obtida através da equação

$$p = P(C < T) = 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_2}{\lambda_1 + \lambda_2}, \tag{5.1}$$

que resulta em

$$\lambda_2 = \frac{p}{(1-p)}\lambda_1. \tag{5.2}$$

Segundo Oliveira (2021), é importante destacar que o percentual de censura obtido não é, necessariamente, igual ao valor pré-fixado, p, pois o percentual de censura gerado é uma quantidade aleatória. O algoritmo proposto por Oliveira (2021) para geração de amostras de tempo de sobrevivência com Censura Aleatória é dado pelos seguintes passos:

- Passo 1: Obter os valores dos parâmetros da distribuição do tempo de censura C a partir dos parâmetros da distribuição do tempo de sobrevivência desejado e proporção de censuras p.
- Passo 2: Gerar $T_1, T_2, ..., T_n$ da distribuição de interesse.
- Passo 3: Gerar $C_1, C_2, ..., C_n$ da distribuição de censura cujos parâmetro foram definidos

no passo 1.

- Passo 4: Gerar o indicador de censura δ_i:
  - Se $T_i \leq C_i$, δ_i = 1.
  - Se $T_i > C_i$, δ_i =0.

Neste trabalho, será considerado que $T \sim Exp(2)$ e, consequentemente, $C \sim Exp\left(\frac{p}{(1-p)}2\right)$, onde p é o percentual de censura desejado. Foi realizado um estudo de simulação que apresenta o percentual de vezes em que os intervalos de confiança obtidos via bootstrap continham o verdadeiro valor do parâmetro da distribuição de T. O estudo foi feito testando diferentes tamanhos de amostra e diferentes percentuais de censura. Os intervalos de confiança foram obtidos por meio de B=1000 reamostras bootstrap e as probabilidades de cobertura foram calculadas a partir de M=1000 réplicas de Monte Carlo. É importante destacar que, neste estudo, os intervalos foram feitos para o parâmetro λ_1, da distribuição Exponencial T.

A Tabela 5.1 e Figura 5.1 apresentam as probabilidades de cobertura dos intervalos de Bootstrap Normal, Básico, Percentil e BCa de 95% de confiança para o parâmetro da distribuição Exponencial. Os resultados apresentados consideram diferentes tamanhos de amostras e diferentes percentuais de censura.

Como visto na Tabela 5.1 e Figura 5.1, os intervalos bootstrap Percentil e BCa foram o que apresentaram os melhores resultados, tendo suas probabilidades de cobertura próximas de 0.95 (nível de confiança nominal dos intervalos) em amostras grandes ($n \geq 50$). Já os intervalos bootstrap Normal Padrão e Básico não apresentaram resultados satisfatórios, mesmo com amostras grandes. Desta forma, este trabalho irá considerar apenas os intervalos boostrap Percentil e BCa nos Capítulos 6 e 7.

Tabela 5.1: Probabilidades de cobertura para os Intervalos Bootstrap Normal, Básico, Percentil e BCa para o parâmetro da Exponencial para diversos tamanhos de amostra e % de censura.

n	**% de censura**	**Probabilidade de Cobertura**			
		Bootstrap Normal	**Bootstrap Básico**	**Bootstrap Percentil**	**Bootstrap BCa**
20	**0%**	0.908	0.887	0.914	0.918
	5%	0.903	0.891	0.911	0.918
	10%	0.906	0.888	0.908	0.926
	20%	0.914	0.890	0.927	0.941
	50%	0.927	0.875	0.933	0.941
30	**0%**	0.907	0.897	0.904	0.917
	5%	0.908	0.893	0.916	0.912
	10%	0.914	0.910	0.913	0.918
	20%	0.909	0.892	0.916	0.915
	50%	0.911	0.878	0.917	0.929
50	**0%**	0.916	0.901	0.918	0.928
	5%	0.931	0.925	0.941	0.954
	10%	0.935	0.924	0.943	0.953
	20%	0.936	0.930	0.948	0.951
	50%	0.940	0.921	0.945	0.939
100	**0%**	0.933	0.926	0.940	0.952
	5%	0.938	0.939	0.937	0.943
	10%	0.935	0.938	0.937	0.945
	20%	0.933	0.932	0.937	0.941
	50%	0.951	0.938	0.940	0.940
200	**0%**	0.948	0.946	0.955	0.951
	5%	0.927	0.924	0.938	0.945
	10%	0.939	0.934	0.942	0.945
	20%	0.930	0.922	0.937	0.942
	50%	0.945	0.935	0.934	0.933
500	**0%**	0.959	0.959	0.958	0.956
	5%	0.936	0.933	0.930	0.938
	10%	0.937	0.934	0.939	0.939
	20%	0.939	0.936	0.935	0.936
	50%	0.944	0.938	0.944	0.946

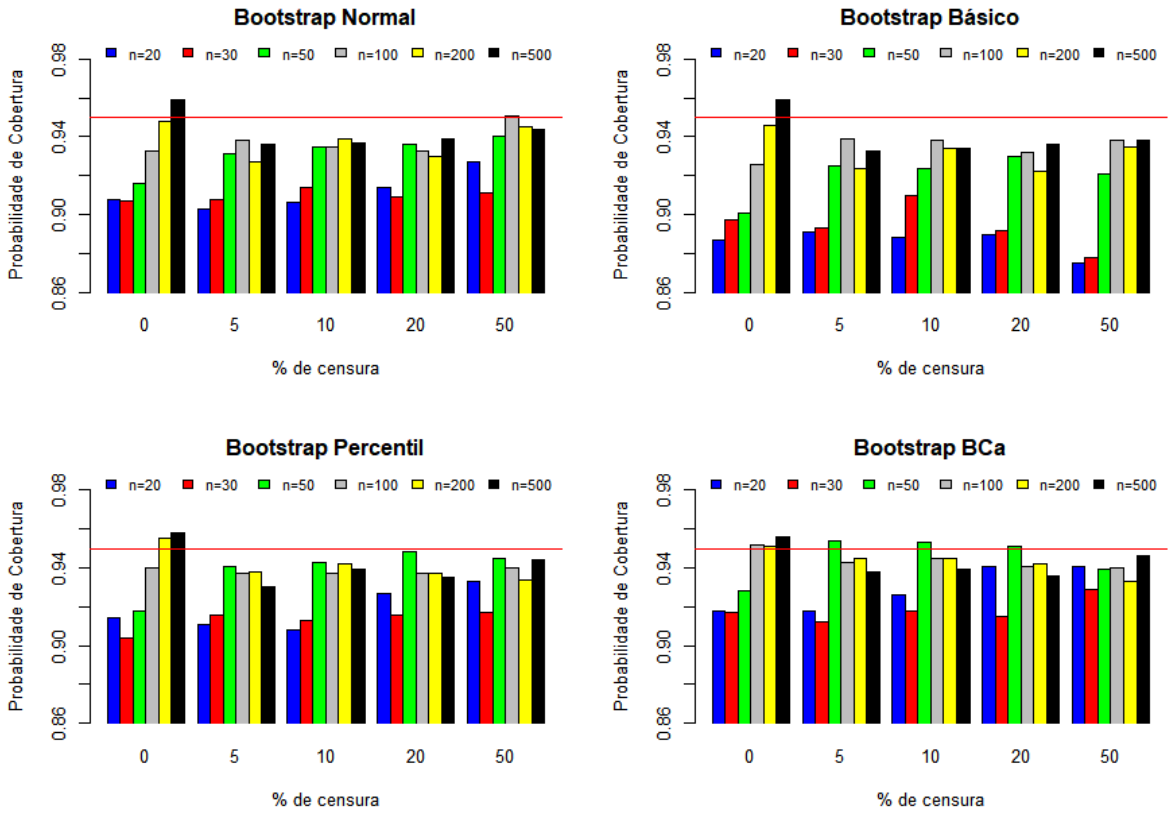

Figura 5.1: Probabilidades de cobertura para os Intervalos Bootstrap Normal, Básico, Percentil e BCa para o parâmetro da Exponencial para diversos tamanhos de amostra e % de censura.

Capítulo 6

Probabilidades de cobertura e DIV

Neste capítulo serão calculadas as probabilidades de cobertura para as estimativas não paramétricas em análise de sobrevivência, por meio de simulações. Para todos os casos, serão consideradas M=1000 réplicas de Monte Carlo, e B=1000 para as estimativas bootstrap. Da mesma forma que no Capítulo 5, será considerado que $T \sim Exp(\lambda)$, com λ=2, e, consequentemente, $C \sim Exp\left(\frac{p}{(1-p)}2\right)$, onde p é o percentual de censura desejado. Em todos os casos, a confiança adotada foi de 1-α = 0.95, com α=0.05. Para as estimativas de funções, como é o caso da função de sobrevivência, risco acumulado e vida média residual, foram considerados a pontos fixos, para a obtenção dos respectivos intervalos de confiança e probabilidades de cobertura. Esta técnica foi adotada pois em cada passo da reamostragem tem-se uma amostra diferente, e, portanto, diferentes pontos. Além disso, para estas três funções, foi calculada uma medida de divergência que não está presente na literatura e foi proposta neste trabalho, que indica o quão distante (em média) a probabilidade de cobertura está do nível de confiança do intervalo. Essa medida de divergência é definida por

$$DIV = \sum_{x=1}^{a} \frac{|\text{Probabilidade de cobertura} - (1-\alpha)|}{a}, \tag{6.1}$$

onde 1 - α é a confiança adotada e a é o número de pontos. Quanto mais próximo de 0 DIV estiver, maior é a precisão do intervalo.

Para a vida média e quantis, será calculada apenas a probabilidade de cobertura, como feito no capítulo 5.

6.1 DIV para a função de sobrevivência S(t)

Para a função de sobrevivência, risco acumulado e vida média residual, foram adotados a=15 pontos fixos para a obtenção dos intervalos de confiança e suas respectivas probabilidades de cobertura. Estes 15 pontos foram os quantis 0.05, 0.10, ..., 0.70, 0.75, de uma distribuição $T \sim Exp(2)$. Para efeito de ilustração da medida de divergência, a Tabela 6.1 apresenta os a=15 pontos, a sobrevivência verdadeira e seus respectivos valores de probabilidade de cobertura. Os valores foram obtidos a partir de uma $T \sim Exp(2)$, sem dados censurados, com n=100 e considerando as estimativas do Kaplan-Meier plano (Expressão 3.8).

Tabela 6.1: Probabilidade de cobertura e DIV para o intervalo KM plano, para a função de sobrevivência, com n=100 e 0% de censura.

Ponto t	**Sobrevivência verdadeira**	**Probabilidade de cobertura**
0.026	0.95	0.864
0.053	0.90	0.935
0.081	0.85	0.937
0.112	0.80	0.938
0.144	0.75	0.953
0.178	0.70	0.961
0.215	0.65	0.945
0.255	0.60	0.949
0.299	0.55	0.946
0.347	0.50	0.950
0.399	0.45	0.953
0.458	0.40	0.958
0.525	0.35	0.947
0.602	0.30	0.953
0.693	0.25	0.947
DIV		0.011

Valor de DIV = 0.011 indica que os valores da probabilidade de cobertura estão variando, em média, 1,1% em torno de 0.95, indicando um bom desempenho do intervalo.

Após isso, foi calculada a divergência DIV para os intervalos de Kaplan-Meier plano (Expressão 3.8), log (Expressão 3.11), log-log (Expressão 3.15), bootstrap percentil (Expressão 4.8) e bootstrap BCa (Expressão 4.10), para diferentes tamanhos de amostra e percentuais de censura. Os resultados são apresentados na Tabela 6.2 e Figura 6.1.

Tabela 6.2: Valores de DIV para os intervalos KM plano, KM log, KM log-log, bootstrap percentil e bootstrap BCa, para a função de sobrevivência, considerando diversos tamanhos de amostra e % de censura.

n	**% de censura**	**DIV**				
		KM plano	**KM log**	**KM log-log**	**Bootstrap Percentil**	**Bootstrap BCa**
10	**0%**	0.056	0.030	0.056	0.048	0.097
	10%	0.044	0.020	0.055	0.045	0.093
	20%	0.043	0.021	0.053	0.041	0.087
	50%	0.081	0.082	0.047	0.033	0.055
30	**0%**	0.021	0.018	0.012	0.023	0.024
	10%	0.015	0.012	0.007	0.019	0.022
	20%	0.016	0.012	0.007	0.015	0.018
	50%	0.036	0.030	0.009	0.011	0.019
50	**0%**	0.015	0.014	0.010	0.011	0.013
	10%	0.016	0.010	0.006	0.016	0.007
	20%	0.018	0.010	0.007	0.016	0.006
	50%	0.023	0.019	0.007	0.012	0.011
100	**0%**	0.011	0.013	0.009	0.015	0.018
	10%	0.013	0.010	0.007	0.009	0.010
	20%	0.014	0.012	0.006	0.007	0.009
	50%	0.019	0.015	0.009	0.004	0.011

Como era esperado, à medida em que o tamanho da amostra aumenta, a divergência DIV tende a diminuir (isto é, a probabilidade de cobertura se aproxima da verdadeira confiança do intervalo). Quando n=10, o valor DIV do intervalo log é menor que os demais, indicando maior precisão. Para os demais tamanho de amostra, o intervalo log-log se comporta melhor que todos os outros. É importante destacar também que o bootstrap percentil foi melhor quando comparado ao bootstrap BCa.

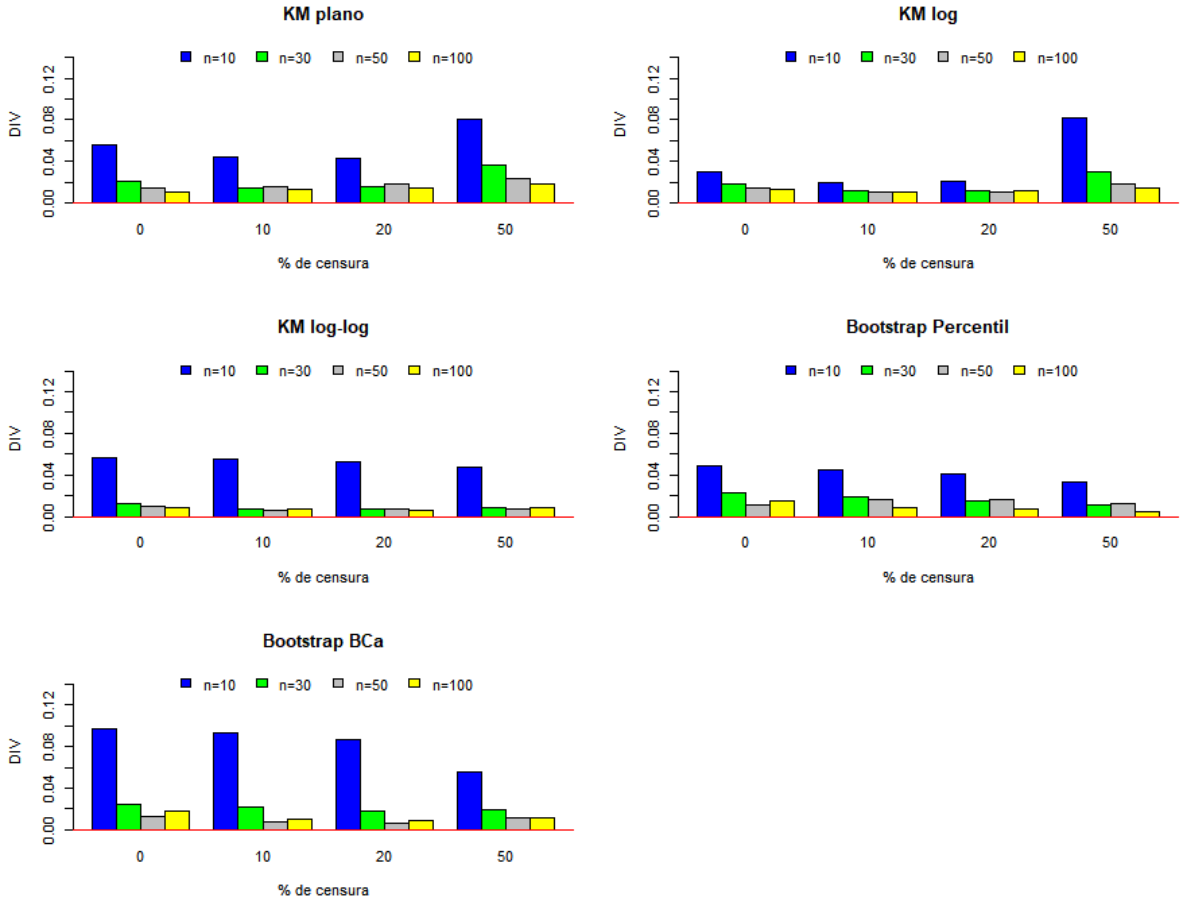

Figura 6.1: Valores de DIV para os intervalos KM plano, KM log, KM log-log, bootstrap percentil e bootstrap BCa, para a função de sobrevivência, considerando diversos tamanhos de amostra e $\%$ de censura.

6.2 DIV para o Risco Acumulado H(t)

Foi calculada a divergência DIV do risco acumulado para os intervalos de Kaplan-Meier plano (Expressão 3.8), log (Expressão 3.11), bootstrap percentil (Expressão 4.8) e bootstrap BCa (Expressão 4.10), para diferentes tamanhos de amostra e percentuais de censura. Os resultados são apresentados na Tabela 6.3 e Figura 6.2.

Pela Figura 6.2, percebe-se que quando n=10, a divergência DIV se comporta melhor com o bootstrap percentil e bootstrap BCa. Com n>10, o intervalo do tipo log se sobressai em comparação aos demais. Quando bootstrap percentil é comparado ao bootstrap BCa, nota-se

Tabela 6.3: Valores de DIV para os intervalos KM plano, KM log, bootstrap percentil e bootstrap BCa, para o risco acumulado, considerando diversos tamanhos de amostra e $\%$ de censura.

n	**% de censura**	**DIV**			
		KM plano	**KM log**	**Bootstrap Percentil**	**Bootstrap BCa**
10	**0%**	0.030	0.056	0.036	0.032
	10%	0.020	0.055	0.027	0.024
	20%	0.021	0.053	0.023	0.026
	50%	0.082	0.049	0.014	0.029
30	**0%**	0.018	0.012	0.023	0.018
	10%	0.012	0.007	0.018	0.019
	20%	0.012	0.007	0.015	0.016
	50%	0.030	0.009	0.013	0.020
50	**0%**	0.014	0.010	0.011	0.016
	10%	0.010	0.006	0.016	0.012
	20%	0.010	0.007	0.016	0.009
	50%	0.019	0.007	0.012	0.010
100	**0%**	0.013	0.009	0.015	0.019
	10%	0.010	0.007	0.009	0.012
	20%	0.012	0.006	0.007	0.010
	50%	0.015	0.009	0.004	0.009

que DIV diminui, à medida em que o tamanho da amostra aumenta.

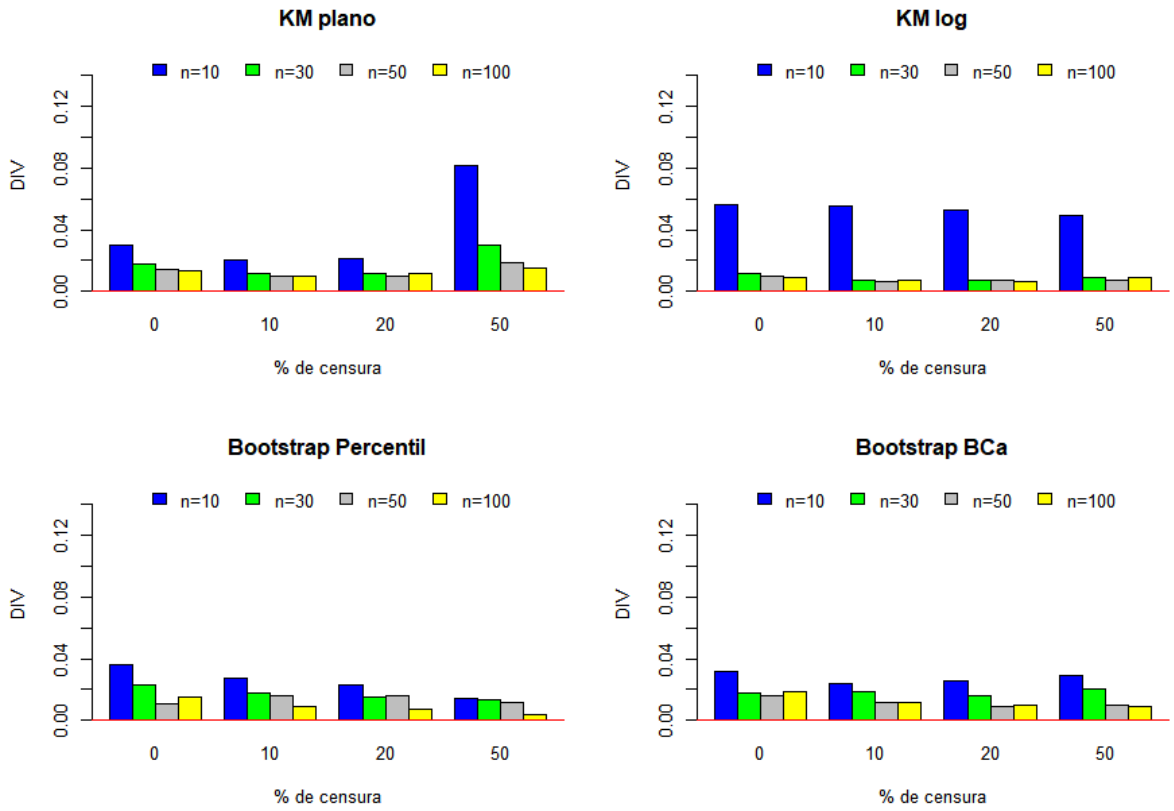

Figura 6.2: Valores de DIV para os intervalos KM plano, KM log, bootstrap percentil e bootstrap BCa, para o risco acumulado, considerando diversos tamanhos de amostra e $\%$ de censura.

6.3 Probabilidade de cobertura para o quantil t_p

A probabilidade de cobertura do quantil $t_{0.5}$ (mediana) para diferentes tamanhos de amostra, diferentes percentuais de censura e com 95% de confiança é descrita na Tabela 6.4 e Figura 6.3, considerando os intervalos plano, bootstrap percentil e bootstrap BCa.

Tabela 6.4: Probabilidade de cobertura para os intervalos bootstrap percentil e bootstrap BCa, para a mediana ($t_{0.5}$), considerando diversos tamanhos de amostra e $\%$ de censura.

n	**% de censura**	**Probabilidade de Cobertura**	
		Bootstrap Percentil	**Bootstrap BCa**
10	**0%**	0.911	0.895
	5%	0.898	0.894
	10%	0.890	0.883
	20%	0.888	0.882
	50%	0.858	—
20	**0%**	0.934	0.931
	5%	0.927	0.926
	10%	0.923	0.924
	20%	0.920	0.928
	50%	0.879	0.904
30	**0%**	0.928	0.927
	5%	0.936	0.932
	10%	0.932	0.933
	20%	0.937	0.944
	50%	0.909	0.939
50	**0%**	0.943	0.946
	5%	0.926	0.930
	10%	0.933	0.930
	20%	0.928	0.925
	50%	0.928	0.932
100	**0%**	0.947	0.947
	5%	0.956	0.960
	10%	0.958	0.958
	20%	0.961	0.963
	50%	0.959	0.957

Pela Figura 6.3, quando n=10, a probabilidade de cobertura do intervalo bootstrap percentil é melhor que a do intervalo bootstrap BCa. Quando o tamanho da amostra aumenta, esses valores tendem a se equilibrar, indicando maior precisão do intervalo bootstrap BCa. Não foi possível calcular a probabilidade de cobertura para n=10 e 50$\%$ de censura, para o intervalo bootstrap BCa.

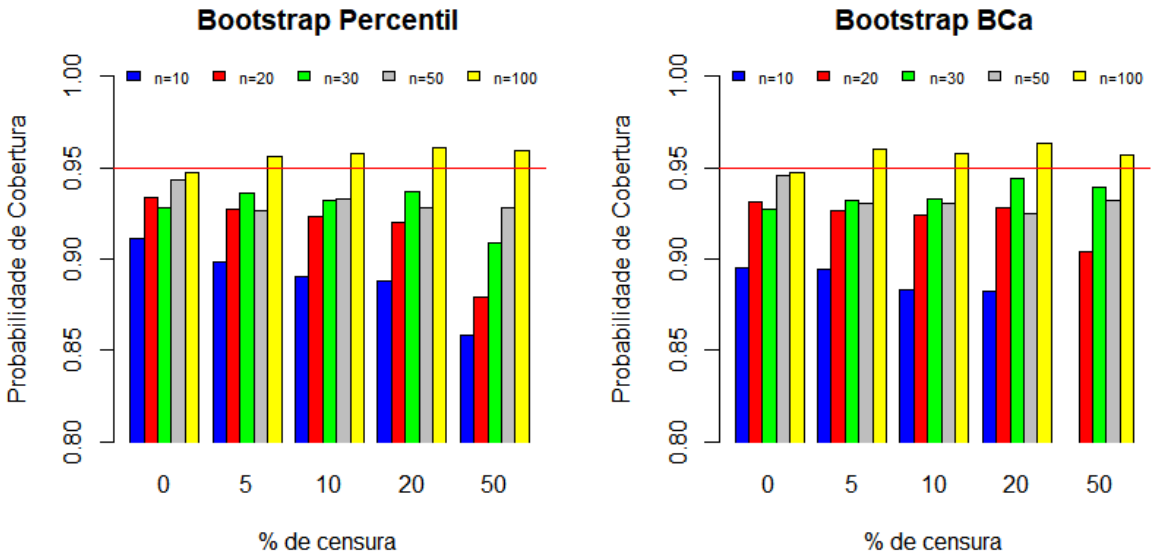

Figura 6.3: Probabilidade de cobertura para os intervalos bootstrap percentil e bootstrap BCa, para a mediana ($t_{0.5}$), considerando diversos tamanhos de amostra e $\%$ de censura.

6.4 Probabilidade de cobertura para a Vida Média

A probabilidade de cobertura da vida média para diferentes tamanhos de amostra, diferentes percentuais de censura e com 95$\%$ de confiança, é descrita na Tabela 6.5 e Figura 6.4, considerando os intervalos plano, bootstrap percentil e bootstrap BCa.

Tabela 6.5: Probabilidade de cobertura para os intervalos KM plano, bootstrap percentil e bootstrap BCa, para a vida média, considerando diversos tamanhos de amostra e % de censura.

n	% de censura	Probabilidade de cobertura		
		KM plano	Bootstrap Percentil	Bootstrap BCa
10	**0%**	0.874	0.863	0.880
	5%	0.851	0.853	0.859
	10%	0.838	0.840	0.852
	20%	0.806	0.826	0.842
	50%	0.637	0.609	0.653
20	**0%**	0.900	0.896	0.909
	5%	0.899	0.887	0.894
	10%	0.895	0.880	0.886
	20%	0.872	0.873	0.882
	50%	0.681	0.713	0.758
30	**0%**	0.919	0.907	0.916
	5%	0.907	0.920	0.933
	10%	0.901	0.913	0.926
	20%	0.886	0.885	0.903
	50%	0.718	0.721	0.769
50	**0%**	0.930	0.925	0.925
	5%	0.921	0.927	0.930
	10%	0.907	0.928	0.923
	20%	0.889	0.910	0.912
	50%	0.771	0.777	0.823
100	**0%**	0.943	0.943	0.952
	5%	0.935	0.942	0.936
	10%	0.924	0.940	0.941
	20%	0.916	0.932	0.934
	50%	0.774	0.806	0.841

Pela Tabela 6.5 e pela Figura 6.4, nota-se um certo equilíbrio entre os três intervalos com 0% e 5% de censura. Quando o tamanho da amostra e o percentual de censura aumentam, o intervalo bootstrap BCa se sobressai em relação ao plano e bootstrap percentil.

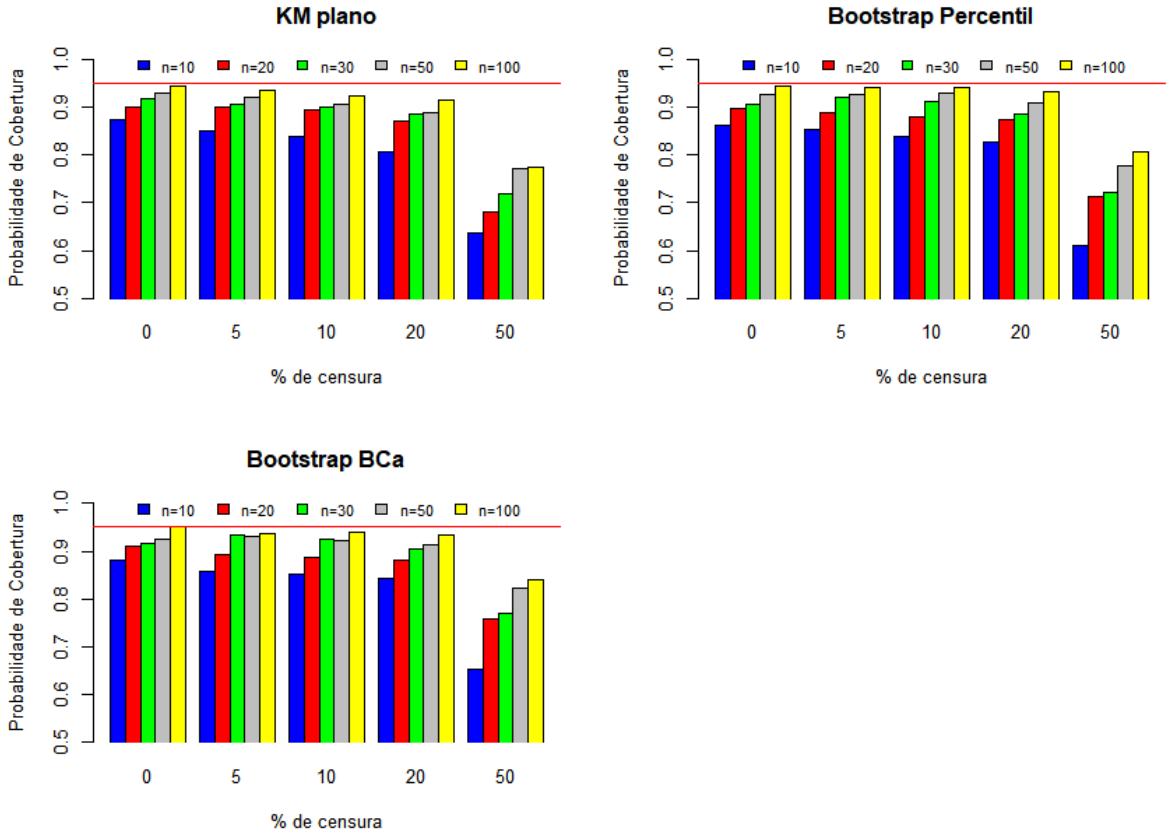

Figura 6.4: Probabilidade de cobertura para os intervalos KM plano, bootstrap percentil e bootstrap BCa, para a vida média, considerando diversos tamanhos de amostra e % de censura.

6.5 DIV para a Vida Média Residual v(t)

Foi calculada a divergência DIV do risco acumulado para os intervalos bootstrap percentil (Expressão 4.8) e bootstrap BCa (Expressão 4.10), para diferentes tamanhos de amostra e percentuais de censura. Não foi possível calcular valores da Probabilidade de Cobertura e, consequentemente, valores de DIV, para amostras de tamanho n=10. Os resultados são apresentados na Tabela 6.6 e Figura 6.5.

Como era esperado, quanto maior o tamanho da amostra, mais o valor de DIV diminui. Pela Tabela 6.6 e Figura 6.5, nota-se que o intervalo Bootstrap BCa foi mais preciso que o Percentil.

Tabela 6.6: Valores de DIV para os intervalos bootstrap percentil e bootstrap BCa, para a vida média residual, considerando diversos tamanhos de amostra e $\%$ de censura.

n	% de censura	DIV	
		Bootstrap Percentil	Bootstrap BCa
30	0%	0.067	0.053
	10%	0.076	0.057
	20%	0.110	0.091
	50%	0.370	0.357
50	0%	0.043	0.034
	10%	0.043	0.036
	20%	0.064	0.054
	50%	0.281	0.264
100	0%	0.020	0.015
	10%	0.028	0.021
	20%	0.049	0.043
	50%	0.235	0.215

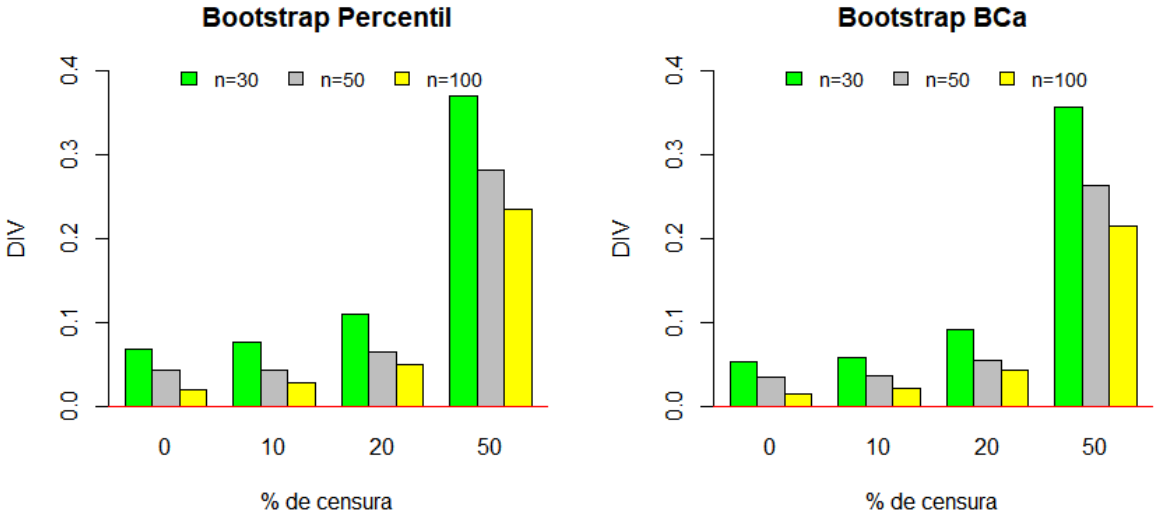

Figura 6.5: Valores de DIV para os intervalos bootstrap percentil e bootstrap BCa, para a vida média residual, considerando diversos tamanhos de amostra e $\%$ de censura.

Capítulo 7

Aplicação em Dados Reais

Este capítulo tem como objetivo aplicar, em dados reais, a técnica de reamostragem bootstrap para calcular o intervalo de confiança dos principais estimadores da análise de sobrevivência. Para todos os cálculos das estimativas bootstrap, será adotado B=1000 reamostras bootstrap.

7.1 Conjunto de Dados

Efron (1988) apresentou um estudo com 51 pacientes com câncer de pescoço e cabeça realizado pelo Grupo de Oncologia do Norte da Califórnia. Os dados são referentes ao tempo de sobrevivência em meses completos destes pacientes. Aqui, t=0 indica que o paciente faleceu antes de um mês de acompanhamento (Biazatti e Nakano, 2020). Os dados são apresentados na Tabela 7.1.

Tabela 7.1: Tempo em meses de 51 pacientes com câncer de pescoço e cabeça realizado pelo Grupo de Oncologia do Norte da Califórnia.

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

Fonte: Efron (1988)

$^+$: observações censuradas.

7.2 Função de Sobrevivência

A Tabela 7.2 apresenta a estimativa do tempo de sobrevivência a partir do estimador de Kaplan-Meier e da reamostragem via bootstrap.

Tabela 7.2: Estimativas de Kaplan-Meier e bootstrap da função de sobrevivência, para os dados da Tabela 7.1.

j	**Tempo** t_j	**Intervalo** $[t_{(j)}, t_{(j+1)}]$	d_j	n_j	$\hat{S}_{KM}(t)$	$\hat{S}_{BOOT}(t)$
1	0	[0;1)	1	51	0.980	0.980
2	1	[1;2)	2	50	0.941	0.941
3	2	[2;3)	5	48	0.843	0.843
4	3	[3;4)	2	42	0.803	0.803
5	4	[4;5)	8	40	0.642	0.642
6	5	[5;7)	7	32	0.502	0.502
7	7	[7;8)	3	24	0.439	0.439
8	8	[8;9)	2	21	0.397	0.397
9	9	[9;13)	2	19	0.356	0.356
10	13	[13;14)	3	15	0.284	0.284
11	14	[14;17)	1	12	0.261	0.261
12	17	[17;19)	1	11	0.237	0.237
13	19	[19;36)	2	9	0.184	0.184
14	36	[36;37)	1	7	0.158	0.158
15	37	[37;46)	1	5	0.126	0.126
16	46	[46,∞)	1	2	0.063	0.063

Pelos resultados da Tabela 7.2, é possível notar que os valores da estimativas pontuais da sobrevivência são iguais, tanto para o estimador de Kaplan-Meier, como para o estimador bootstrap.

A Tabela 7.3 e Figuras 7.1, 7.2 e 7.3 apresentam o intervalo plano, intervalo log, intervalo log-log, intervalo bootstrap percentil e intervalo bootstrap BCa, da função de sobrevivência, todos com 95% de confiança.

Tabela 7.3: Intervalos de confiança KM plano, KM log, KM log-log, bootstrap percentil e bootstrap BCa, com 95% de confiança, para a Função de Sobrevivência, para os dados da Tabela 7.1.

j	**Tempo** t_j	**KM plano**	**KM log**	**KM log-log**	**Bootstrap Percentil**	**Bootstrap BCa**
1	0	[0.942;1.000]	[0.943;1.000]	[0.869;0.997]	[0.941;1.000]	[0.902;1.000]
2	1	[0.877;1.000]	[0.879;1.000]	[0.829;0.981]	[0.863;1.000]	[0.784;0.980]
3	2	[0.743;0.943]	[0.749;0.949]	[0.711;0.918]	[0.745;0.941]	[0.686;0.902]
4	3	[0.694;0.912]	[0.701;0.920]	[0.665;0.889]	[0.686;0.902]	[0.680;0.900]
5	4	[0.510;0.775]	[0.523;0.790]	[0.494;0.758]	[0.510;0.765]	[0.496;0.765]
6	5	[0.363;0.640]	[0.381;0.661]	[0.357;0.630]	[0.361;0.642]	[0.359;0.642]
7	7	[0.301;0.577]	[0.321;0.602]	[0.299;0.570]	[0.308;0.578]	[0.308;0.578]
8	8	[0.261;0.534]	[0.282;0.560]	[0.262;0.529]	[0.268;0.533]	[0.261;0.530]
9	9	[0.221;0.490]	[0.244;0.518]	[0.226;0.488]	[0.226;0.488]	[0.235;0.490]
10	13	[0.155;0.414]	[0.181;0.448]	[0.165;0.416]	[0.167;0.424]	[0.170;0.428]
11	14	[0.134;0.387]	[0.161;0.423]	[0.145;0.392]	[0.149;0.404]	[0.151;0.409]
12	17	[0.114;0.360]	[0.141;0.399]	[0.127;0.367]	[0.113;0.371]	[0.115;0.371]
13	19	[0.069;0.300]	[0.099;0.345]	[0.086;0.311]	[0.076;0.311]	[0.078;0.315]
14	36	[0.048;0.268]	[0.079;0.317]	[0.068;0.282]	[0.052;0.264]	[0.058;0.280]
15	37	[0.023;0.230]	[0.056;0.288]	[0.046;0.249]	[0.031;0.229]	[0.042;0.255]
16	46	[0.000;0.165]	[0.013;0.317]	[0.007;0.214]	[0.000;0.179]	[0.000;0.192]

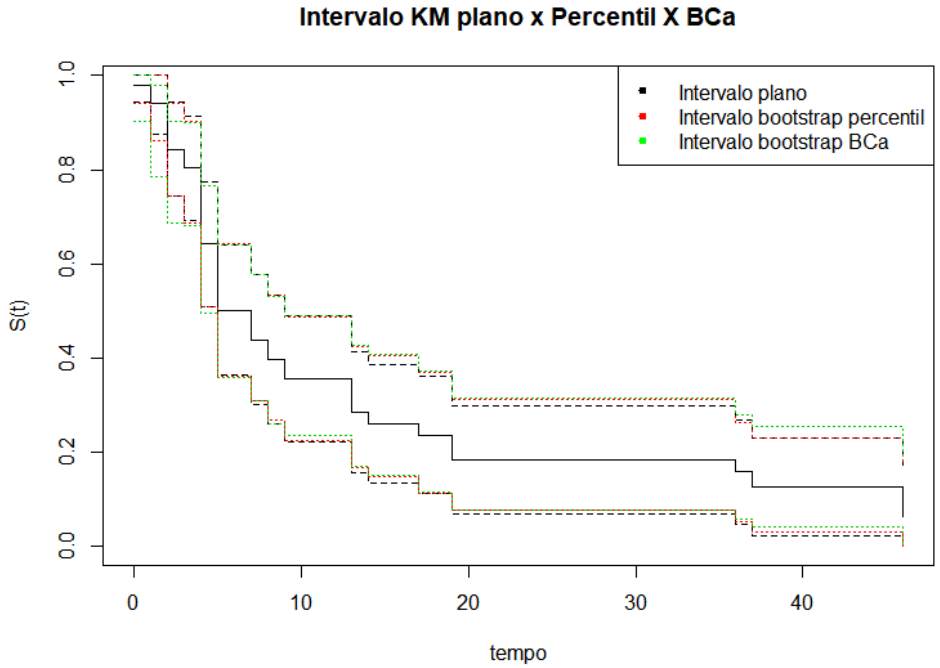

Figura 7.1: Intervalo KM plano vs Intervalo bootstrap percentil vs Intervalo Bootstrap BCa, para os dados da Tabela 7.1.

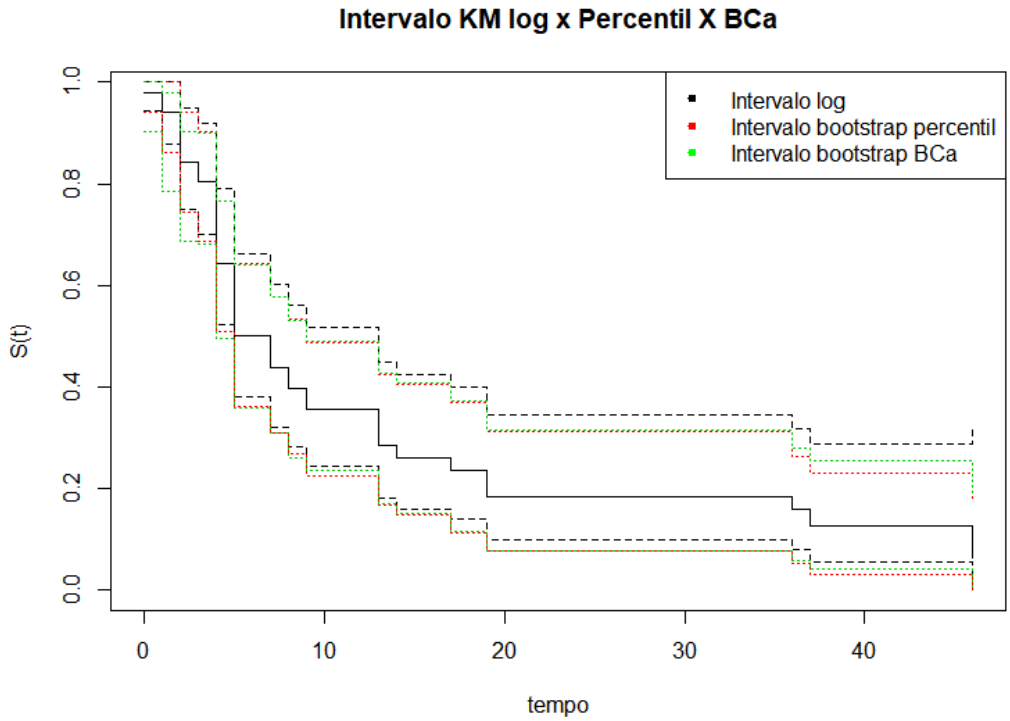

Figura 7.2: Intervalo KM log vs Intervalo bootstrap percentil vs Intervalo Bootstrap BCa, para os dados da Tabela 7.1.

Pelas Figuras 7.1 a 7.3, nota-se que o intervalo KM log-log está mais pareado com os intervalos Bootstrap Percentil e Bootstrap BCa. Já o intervalo KM log, apresenta limites inferiores e superiores sempre maiores que os intervalos Bootstrap Percentil e Bootstrap BCa. Como foi notado no capítulo 6, o intervalo log-log pode ser a melhor escolha para o cálculo de intervalos de confiança para a função de sobrevivência, com o Bootstrap Percentil e BCa logo em seguida.

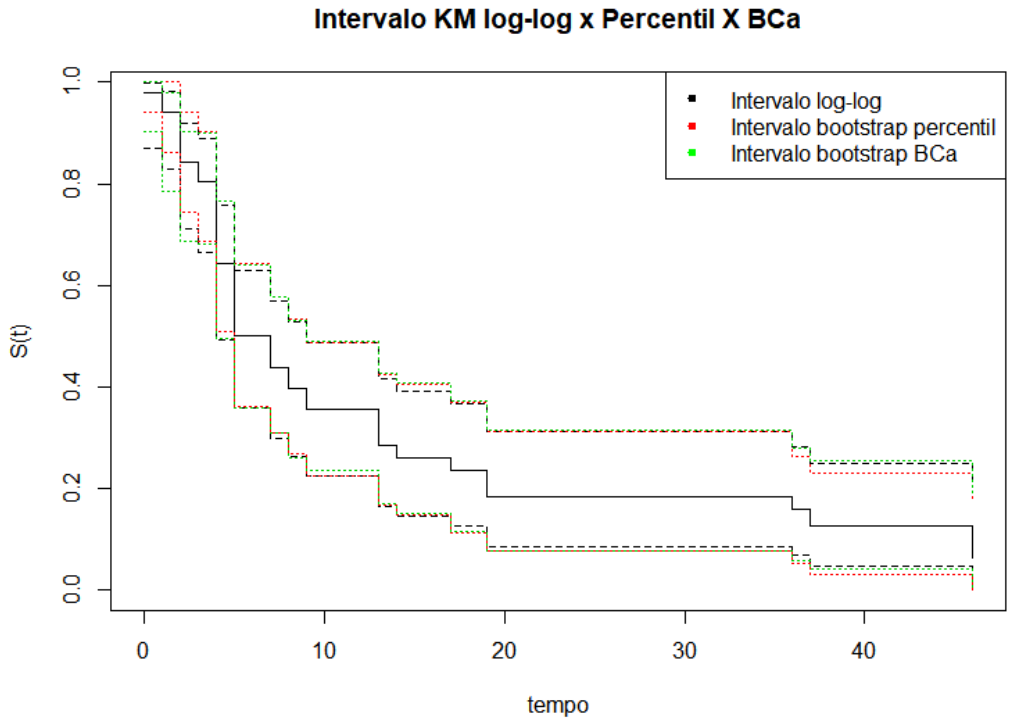

Figura 7.3: Intervalo KM log-log vs Intervalo bootstrap percentil vs Intervalo Bootstrap BCa, para os dados da Tabela 7.1.

7.3 Risco Acumulado

A Tabela 7.4 e Figuras 7.4 e 7.5 apresentam a estimativa do Risco Acumulado a partir do estimador de Kaplan-Meier e da reamostragem via bootstrap.

Pela Tabela 7.4, nota-se que os valores da estimativa pontual do risco acumulado é igual ao valor da estimativa via bootstrap. Estes valores se diferenciariam se fossem consideradas várias casas decimais.

A Tabela 7.5 apresenta o intervalo plano, intervalo log, intervalo bootstrap percentil e intervalo bootstrap BCa, todos com 95% de confiança, para o Risco Acumulado.

Tabela 7.4: Estimador de Kaplan-Meier e estimador bootstrap do Risco Acumulado, para os dados da Tabela 7.1.

j	**Tempo** t_j	**Intervalo** $[t_{(j)}, t_{(j+1)}]$	d_j	n_j	$\hat{H}_{KM}(t)$	$\hat{H}_{BOOT}(t)$
1	0	[0;1)	1	51	0.020	0.020
2	1	[1;2)	2	50	0.061	0.061
3	2	[2;3)	5	48	0.171	0.171
4	3	[3;4)	2	42	0.219	0.219
5	4	[4;5)	8	40	0.443	0.443
6	5	[5;7)	7	32	0.689	0.689
7	7	[7;8)	3	24	0.823	0.823
8	8	[8;9)	2	21	0.923	0.923
9	9	[9;13)	2	19	1.034	1.034
10	13	[13;14)	3	15	1.257	1.257
11	14	[14;17)	1	12	1.344	1.344
12	17	[17;19)	1	11	1.440	1.440
13	19	[19;36)	2	9	1.691	1.691
14	36	[36;37)	1	7	1.845	1.845
15	37	[37;46)	1	5	2.068	2.068
16	46	[46,∞)	1	2	2.761	2.761

Tabela 7.5: Intervalos de confiança KM plano, KM log, bootstrap percentil e bootstrap BCa, com 95% de confiança, para o Risco Acumulado, para os dados da Tabela 7.1.

j	**Tempo** t_j	**KM plano**	**KM log**	**Bootstrap Percentil**	**Bootstrap BCa**
1	0	[0.000;0.059]	[0.003;0.141]	[0.000;0.061]	[0.000;0.061]
2	1	[0.000;0.129]	[0.020;0.188]	[0.000;0.148]	[0.020;0.171]
3	2	[0.052;0.289]	[0.085;0.341]	[0.061;0.294]	[0.082;0.332]
4	3	[0.083;0.356]	[0.118;0.408]	[0.103;0.376]	[0.105;0.383]
5	4	[0.236;0.649]	[0.278;0.706]	[0.268;0.674]	[0.268;0.701]
6	5	[0.413;0.965]	[0.462;1.029]	[0.444;1.018]	[0.443;1.016]
7	7	[0.508;1.138]	[0.561;1.206]	[0.548;1.177]	[0.548;1.177]
8	8	[0.579;1.267]	[0.636;1.340]	[0.630;1.315]	[0.633;1.342]
9	9	[0.657;1.411]	[0.718;1.489]	[0.717;1.489]	[0.713;1.447]
10	13	[0.803;1.711]	[0.876;1.804]	[0.858;1.790]	[0.848;1.773]
11	14	[0.859;1.829]	[0.937;1.928]	[0.906;1.903]	[0.894;1.891]
12	17	[0.920;1.959]	[1.003;2.066]	[0.993;2.181]	[0.990;2.163]
13	19	[1.065;2.317]	[1.168;2.449]	[1.168;2.575]	[1.154;2.547]
14	36	[1.150;2.541]	[1.266;2.690]	[1.331;2.953]	[1.274;2.833]
15	37	[1.246;2.890]	[1.390;3.078]	[1.474;3.250]	[1.389;3.090]
16	46	[1.150;4.373]	[1.541;4.949]	[1.641;3.523]	[2.002;3.754]

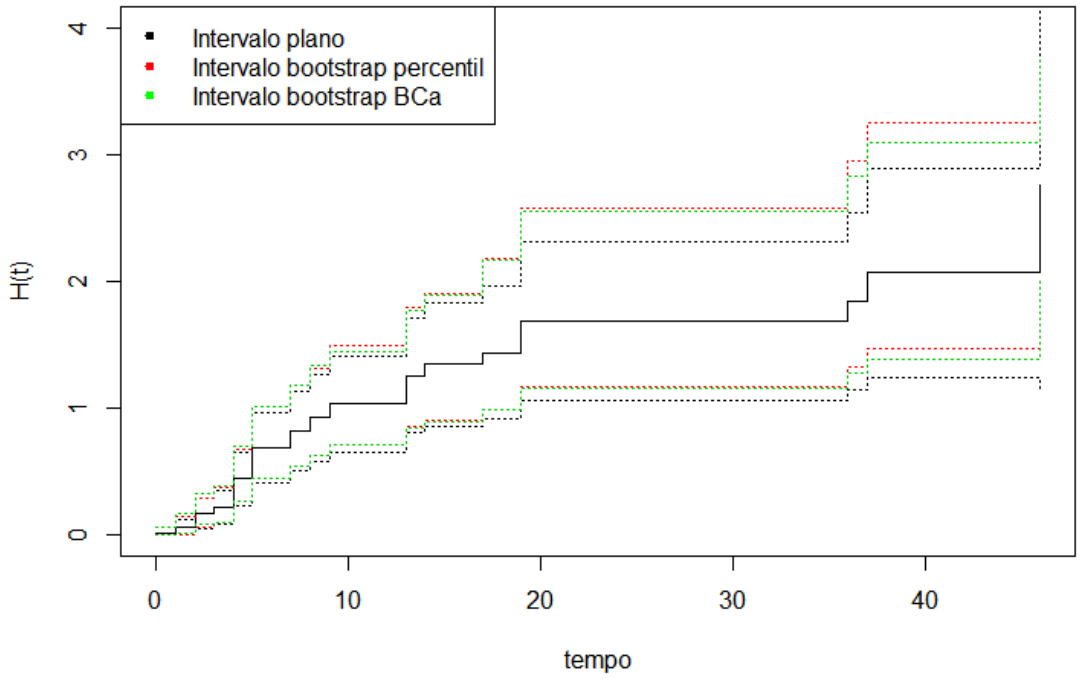

Figura 7.4: Intervalo KM plano vs Intervalo bootstrap percentil vs Intervalo Bootstrap BCa.

Pelas Figuras 7.4 e 7.5, nota-se que o intervalo KM log está mais pareado com os intervalos Bootstrap Percentil e Bootstrap BCa. Já o intervalo KM plano apresenta limites inferiores e superiores sempre menores que os intervalos Bootstrap Percentil e Bootstrap BCa. Como foi notado no Capítulo 6, o intervalo log pode ser a melhor escolha para o cálculo de intervalos de confiança para o risco acumulado, com o Bootstrap Percentil e BCa logo em seguida.

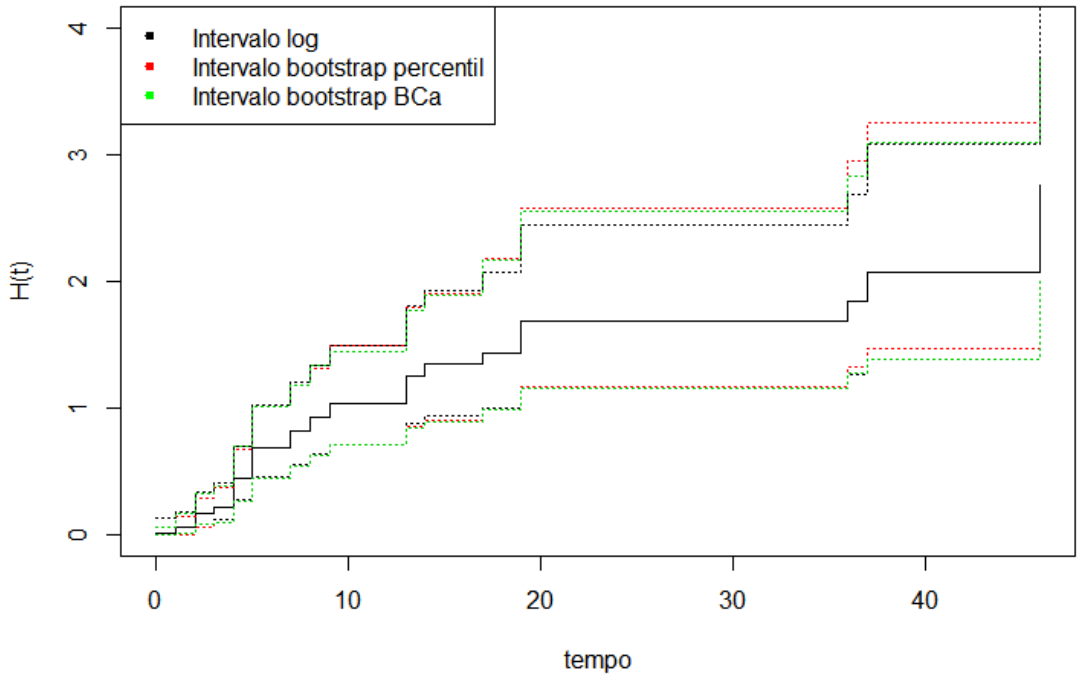

Figura 7.5: Intervalo KM log vs Intervalo bootstrap percentil vs Intervalo Bootstrap BCa.

7.4 Quantis e Vida Média

A Tabela 7.6 apresenta as estimativas pontuais dos quantis 0.25, 0.50 e 0.75, e da vida média, via bootstrap. Também são apresentados seus respectivos intervalos de 95% de confiança, considerando o bootstrap percentil e o bootstrap BCa.

Tabela 7.6: Estimativa da média, quartis, e intervalos com 95% de confiança, para os dados da Tabela 7.1.

Estimador	**Est. pontual via bootstrap**	**IC KM**	**IC bootstrap percentil**	**IC bootstrap BCa**
Média	13.442	[9.076;17.807]	[10.507;17.212]	[9.554;18.276]
$t_{0.25}$	3.330	—	[1.958;4.142]	[1.964;4.153]
$t_{0.5}$	5.060	—	[4.050;8.498]	[4.083;8.723]
$t_{0.75}$	15.353	—	[8.461;36.146]	[9.045;37.427]

Pela Tabela 7.6, nota-se que o intervalo Boostrap Percentil apresenta amplitude menor em

relação aos demais, tanto para a média como para os quantis.

7.5 Vida Média Residual

Para o cálculo da vida média residual, foram considerados intervalos de tempo de 1 em 1, incluindo os pontos censurados da amostra, ou seja, a vida média residual foi calculada para t = $\{0, 1, 2, 3, ...46\}$.

A Tabela 7.7 e Figura 7.6 apresentam a estimativa pontual da vida média residual ($\hat{v}(t)$), a estimativa pontual via bootstrap, e os intervalos bootstrap percentil (Expressão 4.8) e bootstrap BCa (Expressão 4.10), ambos com 95% de confiança:

Tabela 7.7: Estimativa da vida média residual, vida média residual via bootstrap, e intervalos percentil e BCa com 95% de confiança, para os dados da Tabela 7.1.

t	$\hat{v}(t)$	$\hat{v}_{BOOT}(t)$	**Bootstrap Percentil**	**Bootstrap BCa**
0	13.442	13.442	[9.350;18.126]	[9.579;18.239]
1	13.240	13.240	[9.085;17.946]	[9.130;17.991]
2	13.664	13.664	[9.090;18.628]	[9.208;18.781]
3	13.297	13.297	[8.514;18.500]	[8.626;18.613]
4	15.371	15.371	[9.883;21.256]	[10.048;21.384]
5	18.395	18.395	[12.010;25.062]	[12.516;25.609]
15	20.808	20.808	[12.760;27.666]	[11.554;27.039]
20	23.029	23.029	[19.143;26.000]	[18.936;26.000]
25	18.029	18.029	[14.143;21.000]	[13.817;21.000]
35	8.029	8.029	[4.143;11.000]	[3.857;11.000]
36	8.200	8.200	[4.000;10.000]	[1.000;10.000]
37	9.000	9.000	—	—
46	0	0	—	—

Os valores das estimativas pontuais e via bootstrap são iguais e mudam somente quando são consideradas várias casas decimais. É importante notar que $v(0)$ é igual a média, e que a vida média residual pode ser calculada para qualquer ponto t que não esteja na amostra, como foi feito, por exemplo, para os pontos 15 e 20. Outra observação importante: é possível calcular o intervalo de confiança até o antepenúltimo tempo não censurado da amostra, que neste caso é t=36.

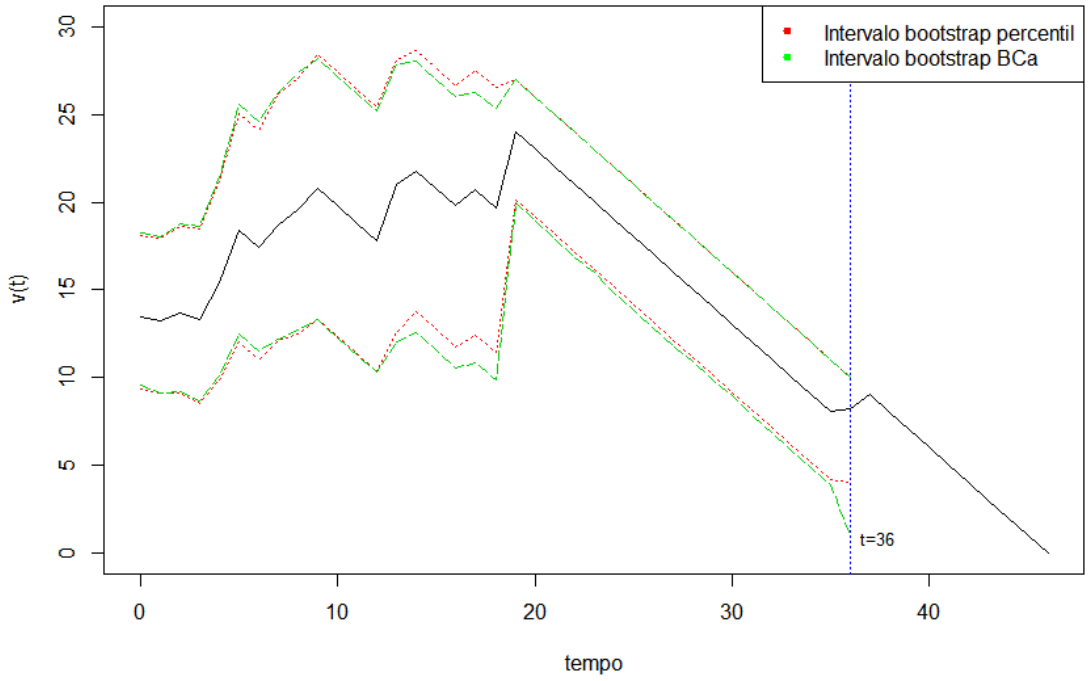

Figura 7.6: Estimativa da vida média residual, vida média residual via bootstrap, e intervalos percentil e BCa com 95% de confiança, para os dados da Tabela 7.1.

Capítulo 8

Conclusão

Levando-se em conta o que foi observado neste trabalho, os intervalos Bootstrap Percentil e BCa se mostraram como uma ótima alternativa no cálculo de intervalos de confiança para as principais estimativas não paramétricas da análise de sobrevivência.

Os resultados apresentados nesse trabalho são de extrema utilidade principalmente nos casos em que metodologias para a obtenção de estimativas intervalares não paramétricas são inexistentes, como é o caso da vida média residual ou algumas definições da mediana, por exemplo.

Ademais, poder contar com mais um procedimento para a obtenção de estimativas intervalares, que seja de simples entendimento e de fácil implementação pode incentivar a popularização das técnicas de análise de sobrevivência, principalmente nas pesquisas aplicadas.

Por fim, como sugestão futura, outros trabalhos podem estudar o comportamento dos intervalos bootstrap por meio de estudos de simulações abrangendo a geração de dados a partir de outras distribuições de probabilidade (com riscos monótonos ou não monótonos), variação do nível de confiança nominal dos intervalos ou até a inclusão de covariáveis (como por exemplo o modelo semi-paramétrico de riscos proporcionais de Cox).

Capítulo 9

ANEXO - Programação em R

```
##### CAPITULO 7
##### APLICACAO EM DADOS REAIS
##### FUNCAO DE SOBREVIVENCIA

library(survival)
library(bootstrap)
library(boot)

t <-c(0,rep(1,2),rep(2,6),3,3,rep(4,8),rep(5,7),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)
censura <- c(rep(1,5),0,rep(1,20),0,rep(1,6),0,1,0,rep(1,5),0,1,1,1,0,1,0,0,0,1)
dados <- data.frame(t,censura)

k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")

if(k_m$time[1]!=0){t1 <- c(0,k_m$time)
}else {t1 <- k_m$time}

theta.boot_sobrev <- function(data,idx){
  censura  <- data[idx,1]
  t        <- data[idx,2]
  k_m <- survfit(Surv(t,censura) ~ 1, data=data[idx,], type="kaplan-meier")

  if (k_m$time[1] !=0){
    sobrev <- rep(1,length(t1))
    for (i in 2:length(t1)) {
      if (sum(t1[i] == k_m$time) > 0)
        sobrev[i] <- k_m$surv[which(k_m$time == t1[i])]
      else if (sum(t1[i] == k_m$time) == 0)
        sobrev[i] <- sobrev[i-1]
    }
    return(sobrev[1:length(t1)])
  }else{
    sobrev <- numeric()
    for (i in 1:length(t1)) {
```

```
      if (sum(t1[i] == k_m$time) > 0)
        sobrev[i] <- k_m$surv[which(k_m$time == t1[i])]
      else if (sum(t1[i] == k_m$time) == 0)
        sobrev[i] <- sobrev[i-1]
    }
    return(sobrev[1:length(t1)])
  }
}

data <- data.frame(censura,t)
data[order(data$t),]

set.seed(777)
bootobject <- boot(data,statistic=theta.boot_sobrev,R= 1000)

#Percentil
ci_percentil<- do.call(rbind, lapply(1:length(bootobject$t0),  function(i)
                      {boot.ci(bootobject,type="perc",conf=.95,index=i)$percent}))

s_boot <- bootobject$t0

li_sboot <-  c(rep(1,sum(s_boot==1)),ci_percentil[,4])
ls_sboot <-  c(rep(1,sum(s_boot==1)),ci_percentil[,5])
unique(cbind(round(li_sboot,3),round(ls_sboot,3)))



#BCa
ci_bca<- do.call(rbind, lapply(1:length(bootobject$t0), function(i)
               {boot.ci(bootobject,type="bca",conf=.95,index=i)$bca}))

li_sboot_bca <-  c(rep(1,sum(s_boot==1)),ci_bca[,4])
ls_sboot_bca <-  c(rep(1,sum(s_boot==1)),ci_bca[,5])
unique(cbind(round(li_sboot_bca,3),round(ls_sboot_bca,3)))

par(mfrow=c(1,3))
####### Intervalo plano x percentil x Bca
k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
plot(k_m, ylab = "S(t)", xlab="tempo", main="Intervalo KM plano x Percentil X BCa")

points(t1,li_sboot,type="s",lty=3,col=2)
points(t1,ls_sboot,type="s",lty=3,col=2)

teste_bca <- t1[1:length(li_sboot_bca)]

points(teste_bca,li_sboot_bca,type="s",lty=3,col=3)
points(teste_bca,ls_sboot_bca,type="s",lty=3,col=3)
legend("topright", c("Intervalo plano","Intervalo bootstrap percentil", "Intervalo bootstrap BCa"),
       col = c("black","red","green"), lwd = 2, lty = c(0,0), pch = c(20,20,20) )

####### Intervalo log x percentil x Bca
km_log <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="log")
plot(km_log, ylab = "S(t)", xlab="tempo", main="Intervalo KM log x Percentil X BCa")
summary(km_log)

points(t1,li_sboot,type="s",lty=3,col=2)
```

```
points(t1,ls_sboot,type="s",lty=3,col=2)

teste_bca <- t1[1:length(li_sboot_bca)]

points(teste_bca,li_sboot_bca,type="s",lty=3,col=3)
points(teste_bca,ls_sboot_bca,type="s",lty=3,col=3)
legend("topright", c("Intervalo log", "Intervalo bootstrap percentil","Intervalo bootstrap BCa"),
       col = c("black","red","green"), lwd = 2, lty = c(0,0), pch = c(20,20,20) )

####### Intervalo log log x percentil x Bca
km_log_log <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="log-log")
plot(km_log_log, ylab = "S(t)", xlab="tempo", main="Intervalo KM log-log x Percentil X BCa")

points(t1,li_sboot,type="s",lty=3,col=2)
points(t1,ls_sboot,type="s",lty=3,col=2)

teste_bca <- t1[1:length(li_sboot_bca)]

points(teste_bca,li_sboot_bca,type="s",lty=3,col=3)
points(teste_bca,ls_sboot_bca,type="s",lty=3,col=3)
legend("topright", c("Intervalo log-log",  "Intervalo bootstrap percentil", "Intervalo bootstrap BCa"),
       col = c("black","red","green"), lwd = 2,  lty = c(0,0), pch = c(20,20,20))

############################
##### RISCO ACUMULADO ##########
k_m_plano <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="log")
d_plano <- summary(k_m_plano)
d_plano$H_t <- -log(d_plano$surv)
LI_plano <- - log(d_plano$upper)
LS_plano <- - log(d_plano$lower)
cbind(round(-log(d_plano$surv),3), round(LI_plano,3),round(LS_plano,3))

k_m_log <- survfit(Surv(t,censura) ~ 1, data=dados,  type="kaplan-meier",conf.type="log-log")
d_log <- summary(k_m_log)
d_log$H_t <- -log(d_log$surv)
LI_log <- - log(d_log$upper)
LS_log <- - log(d_log$lower)
cbind(round(LI_log,3),round(LS_log,3))

###################### IC plano #########
########################################
k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
d <- summary(k_m)

if(k_m$time[1]!=0){
  t1 <- c(0,k_m$time)
}else {t1 <- k_m$time}

theta.boot_risco_ac <- function(data,idx){
  censura  <- data[idx,1]
  t        <- data[idx,2]
  k_m <- survfit(Surv(t,censura) ~ 1, data=data[idx,],
                 type="kaplan-meier")

  k_m$surv <-  k_m$surv[ k_m$surv != 0]
```

```
  if (k_m$time[1] !=0){
    sobrev <- rep(1,length(t1))
    for (i in 2:length(t1)) {
      if (sum(t1[i] == k_m$time) > 0)
        sobrev[i] <- k_m$surv[which(k_m$time == t1[i])]

      else if (sum(t1[i] == k_m$time) == 0)
        sobrev[i] <- sobrev[i-1]
    }
    return(-log(sobrev[1:length(t1)]))
  }else{
    sobrev <- numeric()
    for (i in 1:length(t1)) {
      if (sum(t1[i] == k_m$time) > 0)
        sobrev[i] <- k_m$surv[which(k_m$time == t1[i])]

      else if (sum(t1[i] == k_m$time) == 0)
        sobrev[i] <- sobrev[i-1]
    }
    return(-log(sobrev[1:length(t1)]))
  }
}

data <- data.frame(censura,t)
data[order(data$t),]

set.seed(777)
bootobject <- boot(data,statistic=theta.boot_risco_ac,R= 1000)
bootobject

#Percentil
ci_percentil<- do.call(rbind, lapply(1:length(bootobject$t0), function(i)
                     {boot.ci(bootobject,type="perc",conf=.95,index=i)$percent}))
s_boot <- bootobject$t0
li_sboot_perc <-  c(rep(1,sum(s_boot==1)),ci_percentil[,4])
ls_sboot_perc <-  c(rep(1,sum(s_boot==1)),ci_percentil[,5])
unique(cbind(round(li_sboot_perc,3),round(ls_sboot_perc,3)))


ci_bca     <- do.call(rbind, lapply(1:length(bootobject$t0), function(i)
                   {boot.ci(bootobject,type="bca",conf=.95,index=i)$bca}))
li_sboot_bca <-  c(rep(0,sum(s_boot==0)),ci_bca[,4])
ls_sboot_bca <-  c(rep(0,sum(s_boot==0)),ci_bca[,5])
unique(cbind(round(li_sboot_bca,3),round(ls_sboot_bca,3)))


par(mfrow=c(1,2))
####### Intervalo plano x percentil x Bca
plot(d_plano$time,d_plano$H_t,type = "s",ylab = "H(t)",
     xlab="tempo",main="Intervalo KM plano x Percentil X BCa", ylim = c(0,4))
points(d_log$time,LI_plano,type = "s",lty=3)
points(d_log$time,LS_plano,type = "s",lty=3)

points(t1,li_sboot_perc,type="s",lty=3,col=2)
```

```
points(t1,ls_sboot_perc,type="s",lty=3,col=2)

teste_bca <- t1[1:length(li_sboot_bca)]

points(teste_bca,li_sboot_bca,type="s",lty=3,col=3)
points(teste_bca,ls_sboot_bca,type="s",lty=3,col=3)
legend("topleft", c("Intervalo plano", "Intervalo bootstrap percentil","Intervalo bootstrap BCa"),
       col = c("black","red","green"),lwd=2,lty = c(0,0), pch = c(20,20,20),cex=0.8 )


####### Intervalo log x percentil x Bca
plot(d_log$time,d_log$H_t,type = "s",ylab = "H(t)",
     xlab="tempo",main="Intervalo KM log x Percentil X BCa", ylim = c(0,4))
points(d_plano$time,LI_log,type = "s",lty=3)
points(d_plano$time,LS_log,type = "s",lty=3)

points(t1,li_sboot_perc,type="s",lty=3,col=2)
points(t1,ls_sboot_perc,type="s",lty=3,col=2)

teste_bca <- t1[1:length(li_sboot_bca)]

points(teste_bca,li_sboot_bca,type="s",lty=3,col=3)
points(teste_bca,ls_sboot_bca,type="s",lty=3,col=3)
legend("topleft", c("Intervalo log","Intervalo bootstrap percentil","Intervalo bootstrap BCa"),
       col = c("black","red","green"), lwd = 2, lty = c(0,0), pch = c(20,20,20),cex=0.8 )


################################
#### ESTIMATIVA DA MEDIA ####
################################

############ Codigo Media #################
km<-survfit(Surv(t,censura) ~ 1, data=dados,
            type="kaplan-meier",conf.type="plain")

if(k_m$time[1]!=0){
  n<-length(t)
  t<-km$time
  t<-c(0,t,max(t))
  intervalo<-diff(t,1)
  tj<-t[-(length(t)-1)]
  nj<-c(km$n.risk[1],km$n.risk)
  dj<-c(0,km$n.event)
  sobrev<-km$surv
  sobrev<-c(1,sobrev)
  intXsobrev<-intervalo*sobrev
  k<-length(tj)
  A<-rep(0,k)
  for (j in 1:k-1){
    A[j]<-sum(intXsobrev[(j):k])
  }
  media<-sum(intXsobrev)
  d<-sum(censura)
  YY<- A^2*dj/(nj*(nj-dj))
  YY<-YY[-length(YY)]
```

```
  variancia<-d/(d-1)*sum(YY)
  LI<-media-qnorm(.975)*variancia^.5
  LS<-media+qnorm(.975)*variancia^.5
  X<-c(media,variancia,LI,LS)
  X} else{
    n<-length(t)
    t<-km$time
    t<-c(t,max(t))
    intervalo<-diff(t,1)
    tj<-t[-(length(t)-1)]
    nj<-c(km$n.risk)
    dj<-km$n.event
    sobrev<-km$surv
    intXsobrev<-intervalo*sobrev
    k<-length(tj)
    A<-rep(0,k)
    for (j in 1:k){
      A[j]<-sum(intXsobrev[(j):k])
    }
    media<-sum(intXsobrev)
    d<-sum(censura)
    YY<- A^2*dj/(nj*(nj-dj))
    YY<-YY[-length(YY)]
    variancia<-d/(d-1)*sum(YY)
    LI<-media-qnorm(.975)*variancia^.5
    LS<-media+qnorm(.975)*variancia^.5
    X<-c(media,variancia,LI,LS)
    X
  }


theta.boot_media <- function(data,idx){
  censura  <- data[idx,1]
  t        <- data[idx,2]
  k_m <- survfit(Surv(t,censura) ~ 1, data=data[idx,], type="kaplan-meier")

  if(k_m$time[1]!=0){
    n<-length(t)
    t<-k_m$time
    t<-c(0,t,max(t))
    intervalo<-diff(t,1)
    tj<-t[-(length(t)-1)]
    nj<-c(k_m$n.risk[1],k_m$n.risk)
    dj<-c(0,k_m$n.event)
    sobrev<-k_m$surv
    sobrev<-c(1,sobrev)
    intXsobrev<-intervalo*sobrev
    k<-length(tj)
    A<-rep(0,k)
    for (j in 1:k-1){
      A[j]<-sum(intXsobrev[(j):k])
    }
    media<-sum(intXsobrev)
    return(media)
  } else{
```

```
    n<-length(t)
    t<-km$time
    t<-c(t,max(t))
    intervalo<-diff(t,1)
    tj<-t[-(length(t)-1)]
    nj<-c(km$n.risk)
    dj<-km$n.event
    sobrev<-km$surv
    intXsobrev<-intervalo*sobrev
    k<-length(tj)
    A<-rep(0,k)
    for (j in 1:k){
      A[j]<-sum(intXsobrev[(j):k])
    }
    media<-sum(intXsobrev)
    return(media)
  }
}

data <- data.frame(censura,t)

set.seed(777)
bootobject_media<-boot(data,statistic=theta.boot_media,R=1000)

#Percentil
boot_ci_perc <- boot.ci(bootobject_media, type = "perc")
LI_perc <- boot_ci_perc$percent[4]
LS_perc <- boot_ci_perc$percent[5]

#BCa
set.seed(777)
boot_ci_bca <- boot.ci(bootobject_media, type = "bca")
LI <- boot_ci_bca$bca[4]
LS <- boot_ci_bca$bca[5]

################################################
#### ESTIMATIVA QUANTIL 0.25 ####
########################################
library(survival)
library(bootstrap)
library(boot)

t <-c(0,rep(1,2),rep(2,6),3,3,rep(4,8),rep(5,7),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)
censura <- c(rep(1,5),0,rep(1,20),0,rep(1,6), 0,1,0,rep(1,5),0,1,1,1,0,1,0,0,0,1)
dados <- data.frame(t,censura)

k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
summary(k_m)
km <- summary(k_m)
sobrev  <- km$surv
t1      <- km$time
dados_tp <- cbind(sobrev,t1)
p <- 0.25
t_u_tmp <- subset(dados_tp,sobrev>1-p)
nlin <- nrow(t_u_tmp)
```

```
dados_tp1 <- dados_tp[c(nlin,nlin+1),]

t_u    <- dados_tp1[1,2]
t_u_1  <- dados_tp1[2,2]
S_u    <- dados_tp1[1,1]
S_u_1  <- dados_tp1[2,1]

t_p <- t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1))

theta.boot_q025 <- function(data,idx){

  censura  <- data[idx,1]
  t        <- data[idx,2]
  k_m      <- survfit(Surv(t,censura) ~ 1, data=data[idx,], type="kaplan-meier")
  p <- 0.25
  km <- summary(k_m)
  sobrev  <- km$surv
  t1      <- km$time

  nlin <- max(which(sobrev>1-p))

  if (nlin<length(sobrev)){
    dados_tp1<-data.frame(sobrev[c(nlin,nlin+1)],t1[c(nlin,nlin+1)])
    t_u    <- dados_tp1[1,2]
    t_u_1  <- dados_tp1[2,2]
    S_u    <- dados_tp1[1,1]
    S_u_1  <- dados_tp1[2,1]
    return(t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1)))

  }
  else{
    dados_tp1 <- data.frame(sobrev[nlin],t1[nlin])
    t_u    <- dados_tp1[1,2]
    t_u_1  <- max(t)
    S_u    <- dados_tp1[1,1]
    S_u_1  <- 0
    return(t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1)))
  }
}

data <- data.frame(censura,t)

set.seed(777)
bootobject_q025 <- boot(data,statistic=theta.boot_q025,R=1000)
bootobject_q025

#Percentil
boot_ci_q025 <- boot.ci(bootobject_q025, type = "perc")
LI_q025 <- boot_ci_q025$percent[4]
LS_q025 <- boot_ci_q025$percent[5]
LI_q025
LS_q025

#BCA
boot_ci_q025_bca <- boot.ci(bootobject_q025, type = "bca")
```

```
LI_q025_bca <- boot_ci_q025_bca$bca[4]
LS_q025_bca <- boot_ci_q025_bca$bca[5]
LI_q025_bca
LS_q025_bca

plot(bootobject_q025)

################################
#### ESTIMATIVA QUANTIL 0.50 (MEDIANA) ####

t <-c(0,rep(1,2),rep(2,6),3,3,rep(4,8),rep(5,7),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)
censura <- c(rep(1,5),0,rep(1,20),0,rep(1,6),0,1,0,rep(1,5),0,1,1,1,0,1,0,0,0,1)
dados <- data.frame(t,censura)

k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
summary(k_m)
km <- summary(k_m)
sobrev  <- km$surv
t1      <- km$time
dados_tp <- cbind(sobrev,t1)
p <- 0.5
t_u_tmp <- subset(dados_tp,sobrev>1-p)
nlin <- nrow(t_u_tmp)
dados_tp1 <- dados_tp[c(nlin,nlin+1),]

t_u    <- dados_tp1[1,2]
t_u_1  <- dados_tp1[2,2]
S_u    <- dados_tp1[1,1]
S_u_1  <- dados_tp1[2,1]

t_p <- t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1))

theta.boot1 <- function(data,idx){

  censura  <- data[idx,1]
  t        <- data[idx,2]
  k_m      <- survfit(Surv(t,censura) ~ 1, data=data[idx,], type="kaplan-meier")
  p <- 0.5
  km <- summary(k_m)
  sobrev  <- km$surv
  t1      <- km$time

  nlin <- max(which(sobrev>1-p))

  if (nlin<length(sobrev)){
    dados_tp1<-data.frame(sobrev[c(nlin,nlin+1)],t1[c(nlin,nlin+1)])
    t_u    <- dados_tp1[1,2]
    t_u_1  <- dados_tp1[2,2]
    S_u    <- dados_tp1[1,1]
    S_u_1  <- dados_tp1[2,1]
    return(t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1)))

  }
  else{
    dados_tp1 <- data.frame(sobrev[nlin],t1[nlin])
```

```
    t_u    <- dados_tp1[1,2]
    t_u_1  <- max(t)
    S_u    <- dados_tp1[1,1]
    S_u_1  <- 0
    return(t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1)))
  }
}

data <- data.frame(censura,t)

#Percentil
set.seed(777)
bootobject <- boot(data, statistic =  theta.boot1 , R= 1000)
bootobject

#Percentil
boot.ci(bootobject, type = "perc")

#BCa
boot.ci(bootobject, type = "bca")

################################################
#### ESTIMATIVA QUANTIL 0.75 ####
################################################
t <-c(0,rep(1,2),rep(2,6),3,3,rep(4,8),rep(5,7),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)
censura <- c(rep(1,5),0,rep(1,20),0,rep(1,6), 0,1,0,rep(1,5),0,1,1,1,0,1,0,0,0,1)
dados <- data.frame(t,censura)

k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
summary(k_m)
km <- summary(k_m)
sobrev  <- km$surv
t1      <- km$time
dados_tp <- cbind(sobrev,t1)
p <- 0.75
t_u_tmp <- subset(dados_tp,sobrev>1-p)
nlin <- nrow(t_u_tmp)
dados_tp1 <- dados_tp[c(nlin,nlin+1),]

t_u    <- dados_tp1[1,2]
t_u_1  <- dados_tp1[2,2]
S_u    <- dados_tp1[1,1]
S_u_1  <- dados_tp1[2,1]
t_p <- t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1))

theta.boot_q075 <- function(data,idx){
  censura  <- data[idx,1]
  t        <- data[idx,2]
  k_m      <- survfit(Surv(t,censura) ~ 1, data=data[idx,], type="kaplan-meier")
  p <- 0.75
  km <- summary(k_m)
  sobrev  <- km$surv
  t1      <- km$time

  nlin <- max(which(sobrev>1-p))
```

```
  if (nlin<length(sobrev)){
    dados_tp1 <- data.frame(sobrev[c(nlin,nlin+1)],t1[c(nlin,nlin+1)])
    t_u    <- dados_tp1[1,2]
    t_u_1  <- dados_tp1[2,2]
    S_u    <- dados_tp1[1,1]
    S_u_1  <- dados_tp1[2,1]
    return(t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1)))

  }
  else{
    dados_tp1 <- data.frame(sobrev[nlin],t1[nlin])
    t_u    <- dados_tp1[1,2]
    t_u_1  <- max(t)
    S_u    <- dados_tp1[1,1]
    S_u_1  <- 0
    return(t_u + ((t_u_1 - t_u)*(S_u-(1-p))/(S_u-S_u_1)))
  }
}

data <- data.frame(censura,t)

set.seed(777)
bootobject_q075 <- boot(data,statistic=theta.boot_q075,R=1000)
bootobject_q075

#Percentil
boot_ci_q075 <- boot.ci(bootobject_q075, type = "perc")
LI_q075 <- boot_ci_q075$percent[4]
LS_q075 <- boot_ci_q075$percent[5]
LI_q075
LS_q075

#BCA
boot_ci_q075_bca <- boot.ci(bootobject_q075, type = "bca")
LI_q075_bca <- boot_ci_q075_bca$bca[4]
LS_q075_bca <- boot_ci_q075_bca$bca[5]
LI_q075_bca
LS_q075_bca

################################
##### VIDA MEDIA RESIDUAL
k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
d_original <- summary(k_m)

d_time_original  <- d_original$time

T_max <- max(t)

dados <- data.frame(t,censura)

k_m <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
d <- summary(k_m)

if(k_m$time[1]!=0){
```

```
  sobrev  <- d$surv
  t1      <- d$time
  sobrev1 <- c(1,sobrev)
  t2      <- c(0,t1,T_max)
  deltat  <- diff(t2,1)
  deltat_sobrev1 <- deltat*sobrev1
} else{
  sobrev  <- d$surv
  t1      <- d$time
  sobrev1 <- sobrev
  t2      <- c(t1,T_max)
  deltat  <- diff(t2,1)
  deltat_sobrev1 <- deltat*sobrev1
}

dados2<-data.frame(cbind(seq(deltat_sobrev1),deltat_sobrev1))
dados2$id <- dados2[,1]
dados3 <- dados2[order(dados2$id,decreasing = TRUE),]
dados3$Aj <- rev(cumsum(dados3$deltat_sobrev1))

dados3$t3 <- t2[1:length(dados3$Aj)]
dados3$surv <- sobrev1[1:length(dados3$Aj)]
dados3$tempo <- t2[1:length(t2)-1]
dados4 <- dados3[c("tempo","surv","Aj")]
dados4$vmr_tmp <- 1/dados4$surv*((0)*dados4$surv+dados4$Aj)

if(k_m$time[1]!=0){
  dados4$vmr_tmp <- 1/dados4$surv*((0)*dados4$surv+dados4$Aj)
} else{
  dados4$vmr_tmp[1] <- dados4$Aj[1]
}

k <- nrow(dados4)

tempo<-data.frame(sort(c(seq(0,T_max,1),d_time_original,0,T_max)))

tempo <- unique(tempo)

tempo$tempo <- tempo[,1]
tempo <- tempo[2]

dados5 <- merge(tempo,dados4,by='tempo',all.x = T)
k2 <- nrow(dados5)

dados5$t_u <- rep(0,k2)

for (i in 2:k2) {
  for(j in 2:k){
    if((dados5$tempo[i] <= dados4$tempo[j]) &&
       ((dados5$tempo[i] > dados4$tempo[j-1]))){
      dados5$vmr_tmp[i] <-  dados4$vmr_tmp[j]
      dados5$Aj[i]      <-  dados4$Aj[j]
      dados5$t_u[i] <-  dados4$tempo[j]
    }
    if((dados5$tempo[i] < dados4$tempo[j]) &&
```

```
       ((dados5$tempo[i] >= dados4$tempo[j-1]))){
      dados5$surv[i] <-  dados4$surv[j-1]}
  }
}

dados5$vmr_final <- 1/dados5$surv *
  ((dados5$t_u - dados5$tempo)*dados5$surv + dados5$Aj)

if(k_m$time[1]!=0){
  dados5$vmr_final <- 1/dados5$surv * ((dados5$t_u - dados5$tempo)*dados5$surv + dados5$Aj)
} else{
  dados5$vmr_final[1] <- dados5$vmr_tmp[1]
}

# Estimativa pontual da Vida Media Residual
dados5$vmr_final[which(is.na(dados5$vmr_final))]<-T_max - dados5$tempo[which(is.na(dados5$vmr_final))]
plot(dados5$tempo,dados5$vmr_final,type = "l")
cbind(dados5$tempo,round(dados5$vmr_final,3))


# Estimativa via Bootstrap
t <-c(0,rep(1,2),rep(2,6),3,3,rep(4,8),rep(5,7),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)
censura <- c(rep(1,5),0,rep(1,20),0,rep(1,6), 0,1,0,rep(1,5),0,1,1,1,0,1,0,0,0,1)
dados <- data.frame(t,censura)

k_m_original <- survfit(Surv(t,censura) ~ 1, data=dados, type="kaplan-meier",conf.type="plain")
d_original <- summary(k_m_original)
d_time_original  <- d_original$time

if(k_m_original$time[1]!=0){
  tempos <- sort(c(seq(0,max(k_m_original$time),1), d_time_original,max(k_m_original$time)))
} else {
  tempos <- sort(c(seq(min(d_time_original), max(k_m_original$time),1),d_time_original,max(k_m_original$time)))
}

t_distintos <- unique(tempos)

T_max <- max(t_distintos)

theta.boot_vmr <- function(data,idx){
  censura  <- data[idx,1]
  t        <- data[idx,2]
  T_max    <- max(data[idx,3])
  km <- survfit(Surv(t,censura) ~ 1, data=data[idx,], type="kaplan-meier")
  d <- summary(km)

  if(km$time[1]!=0){
    sobrev  <- d$surv
    t1      <- d$time
    sobrev1 <- c(1,sobrev)
    t2      <- c(0,t1,T_max)
    deltat  <- diff(t2,1)
    deltat_sobrev1 <- deltat*sobrev1
  } else{
    sobrev  <- d$surv
```

```
  t1      <- d$time
  sobrev1 <- sobrev
  t2      <- c(t1,T_max)
  deltat  <- diff(t2,1)
  deltat_sobrev1 <- deltat*sobrev1
}

dados2 <- data.frame(cbind(seq(deltat_sobrev1),deltat_sobrev1))
dados2$id <- dados2[,1]
dados3 <- dados2[order(dados2$id,decreasing = TRUE),]
dados3$Aj <- rev(cumsum(dados3$deltat_sobrev1))

dados3$t3 <- t2[1:length(dados3$Aj)]
dados3$surv <- sobrev1[1:length(dados3$Aj)]
dados3$tempo <- t2[1:length(t2)-1]
dados4 <- dados3[c("tempo","surv","Aj")]
dados4$vmr_tmp <- 1/dados4$surv * ( (0)*dados4$surv + dados4$Aj)

if(km$time[1]==0){
  dados4$vmr_tmp[1] <- dados4$Aj[1]}

k <- nrow(dados4)
tempo <- data.frame(sort(c(seq(0,T_max,1), d_time_original,0,T_max)))
tempo <- unique(tempo)
tempo$tempo <- tempo[,1]
tempo <- tempo[2]

dados5 <- merge(tempo,dados4,by='tempo',all.x = T)
k2 <- nrow(dados5)

dados5$t_u <- rep(0,k2)

for (i in 2:k2) {
  for(j in 2:k){

    if((dados5$tempo[i] <= dados4$tempo[j]) &&
       ((dados5$tempo[i] > dados4$tempo[j-1]))){
      dados5$Aj[i]      <-  dados4$Aj[j]
      dados5$t_u[i] <-  dados4$tempo[j]
    }
    if((dados5$tempo[i] < dados4$tempo[j]) &&
       ((dados5$tempo[i] >= dados4$tempo[j-1]))){
      dados5$surv[i] <-  dados4$surv[j-1]}
  }
}

dados5$vmr_final <- 1/dados5$surv * ((dados5$t_u - dados5$tempo)*dados5$surv + dados5$Aj)

if(km$time[1]!=0){
  dados5$vmr_final <- 1/dados5$surv * ((dados5$t_u - dados5$tempo)*dados5$surv + dados5$Aj)
} else{
  dados5$vmr_final[1] <- dados5$vmr_tmp[1]
}

dados5$vmr_final[which(is.na(dados5$vmr_final))]<-T_max - dados5$tempo[which(is.na(dados5$vmr_final))]
```

```
  vmr <- dados5$vmr_final

  return(vmr[1:length(t_distintos)])
}

data <- data.frame(censura,t,rep(T_max,length(censura)))

set.seed(777)
boot_vmr <- boot(data, statistic =  theta.boot_vmr , R= 1000)
boot_vmr

#Percentil
ci_percentil<- do.call(rbind, lapply(1:length(boot_vmr$t0), function(i)
                     {boot.ci(boot_vmr,type="perc",conf=.95,index=i)$percent}))
ci_percentil <- subset(ci_percentil,ci_percentil[,4] != 0)
li_sboot_perc <-  ci_percentil[,4]
ls_sboot_perc <-  ci_percentil[,5]
cbind(seq(0,36),round(li_sboot_perc,3),round(ls_sboot_perc,3))

#BCa
ci_bca     <- do.call(rbind, lapply(1:nrow(ci_percentil), function(i)
                   {boot.ci(boot_vmr,type="bca",conf=.95,index=i)$bca}))
li_sboot_bca <-  ci_bca[,4]
ls_sboot_bca <-  ci_bca[,5]
cbind(seq(0,36),round(li_sboot_bca,3),round(ls_sboot_bca,3))

####### Est. pontual x percentil x Bca
plot(dados5$tempo[1:nrow(ci_bca)], dados5$vmr_final[1:nrow(ci_bca)],ylim=c(0,30),xlab="Tempo",ylab="v(t)",type = "l")

plot(dados5$tempo,dados5$vmr_final,ylim=c(0,30), xlab="tempo",ylab="v(t)",type = "l")

#points(dados5$tempo,vmr_boot,type="l",col=2)
points(dados5$tempo[1:nrow(ci_bca)], li_sboot_perc,type="l",lty=3,col=2)
points(dados5$tempo[1:nrow(ci_bca)], ls_sboot_perc,type="l",lty=3,col=2)

points(dados5$tempo[1:nrow(ci_bca)], li_sboot_bca,type="l",lty=5,col=3)
points(dados5$tempo[1:nrow(ci_bca)], ls_sboot_bca,type="l",lty=5,col=3)
abline(v=36, col = "blue",lty=3)
legend("topright", c("Intervalo bootstrap percentil", "Intervalo bootstrap BCa"), col = c("red","green"),
            lwd = 2, lty = c(0,0), pch = c(20,20,20) )
legend(35,2,"t=36",bty="n",cex=0.8)
```

Referências Bibliográficas

Aalen, O. (1978). *Nonparametric inference for a family of counting processes*. Annals of Statistics, vol. 6, n.4, p. 701-726.

Aalen, O. e Johansen, S. (1978). *An empirical transition matrix for non-homogeneous markov chains based on censored observations*. Scandinavian Journal of Statistics, vol. 5, n.3, p. 141-150.

Biazatti, E. C. e Nakano, E. Y. (2020). Uma proposta de orientação para o uso de modelos contínuos em dados de sobrevivência discretos. *REMAT: Revista Eletrônica daMatemática, vol.6, n.2, e4002*.

Carrasco, C. e Nakano, E. (2016). Estimação intervalar para os parâmetros do modelo exponencial discreto: uma aplicação para dados de sobrevivência. *Nucleus, vol.13, n.1, p. 41-52.*

Carrasco, C., Tutia, M., e Nakano, E. (2012). Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. *TEMA (São Carlos), vol 13, n.3, p.247-255*.

Colosimo, E. e Giolo, S. (2006). *Análise de sobrevivência aplicada*. São Paulo: Edgard Blucher Ltda.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, vol.7, n.1, p.1-26*.

Efron, B. (1988). Logistic regression,survival analysis and the kaplan-meier curve. *Journal of the American Statistical Association, vol.83, n.402, p.414-425*.

Efron, B. e Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman e Hall/CRC Monographs on Statistics and Applied Probability.

Greenwood, M. (1926). *The natural duration of cancer*. Reports on Public Health and Medical subjects , p.1-26.

James, B. (2015). *Probabilidade: um curso em nível intermediário*. Rio de Janeiro: IMPA, p. 299, 4ed.

Kalbfleisch, J. e Prentice, R. (2002). *The Statistical Analysis of Failure Time*. John Wiley and Sons, New York.

Kaplan, E. e Meier, P. (1958). *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, vol.53, n.282, p.457-481.

Meyer, P. (1983). *Probabilidade - Aplicações à Estatística*. Livros Técnicos e Científicos Editora S.A, p. 444.

Nakano, E. Y. (2017). Um curso de análise de sobrevivência. *Departamento de Estatística, Universidade de Brasília, Brasília.*

Nelson, W. (1972). *Theory and applications of hazard plotting for censored failure data*. Technometrics, vol.14, n.4, p.945-966.

Oliveira, F. A. P. (2021). Procedimentos de geração de dados de sobrevivência com censura à direita. *Dissertação - Mestrado em Estatística, Universidade de Brasília*.