

UNIVERSIDADE DE BRASÍLIA  
INSTITUTO DE CIÊNCIAS EXATAS  
DEPARTAMENTO DE ESTATÍSTICA  
PROGRAMA DE MESTRADO EM ESTATÍSTICA

**Modelo Weibull discreto com fração de cura e excessos de zeros: uma aplicação sobre o tempo de sobrevivência de pacientes submetidos à intervenção coronária percutânea.**

**PRISCILA NASCIMENTO DE ALCÂNTARA GARCIA**

Brasília  
2015

**Modelo Weibull discreto com fração de cura e excessos de zeros: uma aplicação sobre o tempo de sobrevivência de pacientes submetidos à intervenção coronária percutânea.**

PRISCILA NASCIMENTO DE ALCÂNTARA GARCIA

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Universidade de Brasília  
Brasília  
2015

## RESUMO

O objetivo deste trabalho foi formular um modelo Weibull discreto com fração de cura e excessos de zeros para dados de sobrevivência. Procedimentos inferenciais do modelo proposto foram apresentados e as estimativas dos parâmetros foram obtidas através de procedimentos numéricos. O modelo proposto foi ilustrado em simulações e aplicado em um conjunto de dados reais sobre o tempo de sobrevivência de pacientes submetidos à intervenção coronária percutânea (ICP). Todas as simulações e estimativas foram obtidas através do software livre R.

**Palavras-chave:** Fração cura; distribuições com excessos de zeros; análise de sobrevivência; modelos discretos; mistura.

## ABSTRACT

The aim of this work was to formulate a zero-inflated discrete Weibull model with cure rate for survival data. Inferential procedures of the proposed model were presented and estimates of the parameters were numerically obtained. The proposed model was illustrated using simulated data and by an application on a real database of survival time of patients treated by a percutaneous coronary intervention (PCI). All simulations and estimates were performed by free software R.

**Keywords:** Cure rate, zero-Inflated distributions; survival analysis; discrete models; mixture.

## AGRADECIMENTOS

Primeiramente a Deus pela oportunidade.

Ao meu esposo, Phillipe Garcia, pelo amor, paciência e companheirismo.

A minha mãe, por todo investimento e incentivo.

Ao meu orientador Prof. Dr. Eduardo Yoshio Nakano, pelo ensinamento, dedicação e confiança depositada.

Aos meus amigos e familiares pela força, amizade e troca de conhecimentos.

Aos professores do Departamento de Estatística da UnB e a todos que propiciaram condições para a realização deste trabalho.

## ÍNDICE DE FIGURAS

Figura 1 – Forma típica da função de sobrevivência contínua .....	14
Figura 2 – Forma típica da função de sobrevivência discreta .....	14
Figura 3 - Função de sobrevivência imprópria.....	15
Figura 4– Exemplo de Funções de risco (crescente, decrescente e constante).....	16
Figura 5– Curvas de sobrevivência estimada por Kaplan-Meier e curvas de sobrevivência estimada dos modelos: Weibull e Exponencial .....	22
Figura 6 - Distribuição de probabilidades da Weibull discreta para diversos valores dos.....	26
Figura 7 - Função de sobrevivência da Weibull discreta para diversos valores dos parâmetros $q$ e $\beta$ . .....	27
Figura 8 - Função de risco da Weibull discreta para diversos valores dos parâmetros $q$ e $\beta$ . .....	28
Figura 9 - Representação do modelo por meio de uma mistura de três distribuições. ....	31
Figura 10 - Distribuição de probabilidades da Weibull discreta com fração de cura e excesso de zeros para diversos valores dos parâmetros $\phi, \rho, q$ e $\beta$ .....	33
Figura 11 - Função de sobrevivência da Weibull discreta com fração de cura e excesso de zeros para diversos valores dos parâmetros $\phi, \rho, q$ e $\beta$ .....	34
Figura 12 Função de risco da Weibull discreta com fração de cura e excesso de zeros para diversos valores dos parâmetros $\phi, \rho, q$ e $\beta$ .....	34
Figura 13 - Curvas de sobrevivência estimadas considerando o Cenário 1 .....	38
Figura 14 - Curvas de sobrevivência estimadas considerando o Cenário 2 .....	41
Figura 15 - Curvas de sobrevivência estimadas considerando o Cenário 2 (eixos reduzidos) .....	41
Figura 16 - Curvas de sobrevivência estimadas considerando o Cenário 3 .....	44
Figura 17 - Curvas de sobrevivência estimadas considerando o Cenário 3 (eixos reduzidos) .....	44
Figura 18 - Curvas de sobrevivência estimadas considerando o Cenário 4 .....	47
Figura 19- Curvas de sobrevivência estimadas considerando o Cenário 4 (eixos reduzidos) .....	47
Figura 20- Curvas de sobrevivência estimadas do tempo de sobrevivências (em semanas) de pacientes submetidos à ICP. ....	51
Figura 21 – Curvas de sobrevivência estimadas do tempo de sobrevivências (em semanas) de pacientes submetidos à ICP. (eixo $y$ reduzido).....	51

## ÍNDICE DE TABELAS

Tabela 1 - Quantis do teste de Kolmogorov-Smirnov.....	23
Tabela 2 - Cenários utilizados na simulação .....	37
Tabela 3 - Estimativas dos parâmetros dos modelos para o Cenário 1, $n=200$ e 10% de censura.....	37
Tabela 4 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 3. ....	38
Tabela 5 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 1. ....	39
Tabela 6 - Estimativas dos parâmetros dos modelos para o Cenário 2, $n=200$ e 10% de censura.....	40
Tabela 7 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 6. ....	40
Tabela 8 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 2. ....	42
Tabela 9 - Estimativas dos parâmetros dos modelos para o Cenário 3, $n=200$ e 10% de censura.....	43
Tabela 10 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 9. ....	43
Tabela 11 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 3. ....	45
Tabela 12 - Estimativas dos parâmetros dos modelos para o Cenário 4, $n=200$ e 10% de censura.....	46
Tabela 13 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 12. ....	46
Tabela 14 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 4. ....	48
Tabela 15 - Estimativas dos parâmetros dos modelos para os dados do tempo de sobrevivências de pacientes submetidos à ICP. ....	50
Tabela 16 - Tempos de falha e suas respectivas estimativas da função de sobrevivência. ...	52
Tabela 17 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 16. ....	52
Tabela 18 - Teste da Razão de Verossimilhança considerando como modelo completo o modelo WDFZ.....	53

## ÍNDICE DE ABREVIATURAS E SIGLAS

AIC: Critério de informação de Akaike

BIC: Critério de informação bayesiano

EQM: Erro quadrático médio

GEO: Geométrico

$h_{WD}$ : Função de risco da distribuição Weibull discreta

$h_{WDF}$ : Função de risco da distribuição Weibull discreta com fração de cura

$h_{WDFZ}$ : Função de risco da distribuição Weibull discreta com fração de cura e excesso de zeros

$h_{WDZ}$ : Função de risco da distribuição Weibull discreta com excesso de zeros

IAM: Infarto agudo do miocárdio

ICP: Intervenção coronária percutânea

KS: Kolmogorov-Smirnov

$L_{WD}$ : Função de verossimilhança para o modelo Weibull discreto

$L_{WDF}$ : Função de verossimilhança para o modelo Weibull discreto com fração de cura

$L_{WDFZ}$ : Função de verossimilhança para o modelo Weibull discreto com fração de cura e excesso de zeros

$L_{WDZ}$ : Função de verossimilhança para o modelo Weibull discreto com excesso de zeros

$P_{WD}$ : Distribuição de probabilidades da Weibull discreta

$P_{WDF}$ : Distribuição de probabilidades da Weibull discreta com fração de cura

$P_{WDF(Z)}$ : Distribuição de probabilidades Weibull discreto com excesso de zeros aplicado a casos com fração de cura

$P_{WDFZ}$ : Distribuição de probabilidades da Weibull discreta com fração de cura e excesso de zeros

$P_{WDZ}$ : Distribuição de probabilidades da Weibull discreta com excesso de zeros

$P_{WDZ(F)}$ : Distribuição de probabilidades Weibull discreto com fração de cura aplicado a casos com excesso de zeros

$\hat{S}_{KM}$ : Função de Sobrevivência estimada via Kaplan-Meier

$S_{WD}$ : Função de Sobrevivência da distribuição Weibull discreta

$S_{WDF}$ : Função de Sobrevivência da distribuição Weibull discreta com fração de cura

$S_{WDFZ}$ : Função de Sobrevivência da distribuição Weibull discreta com fração de cura e excesso de zeros



$S_{WDF(Z)}$ : Função de Sobrevivência da distribuição Weibull discreta com excesso de zeros aplicado a casos com fração de cura

$S_{WDZ}$ : Função de Sobrevivência da distribuição Weibull discreta com excesso de zeros

$S_{WDZ(F)}$ : Função de Sobrevivência da distribuição Weibull discreta com fração de cura aplicado a casos com excesso de zeros

TRV: Teste da razão de verossimilhanças

WD: Weibull discreto

WDF: Weibull discreto com fração de cura

WDFZ: Weibull discreto com fração de cura e excessos de zeros

WDZ: Weibull discreto com excessos de zeros

## SUMÁRIO

1. Introdução .....	10
2. Revisão Bibliográfica.....	12
2.1. Conceitos Básicos em Análise de Sobrevivência.....	12
2.1.1. Descrição do tempo de sobrevivência.....	13
2.1.1.1. Função de Sobrevivência.....	13
2.1.1.2. Função de Taxa de falha ou Função de Risco .....	15
2.1.2. Estimacão empírica da função de sobrevivência (Kaplan-Meier).....	17
2.1.3. Modelos Exponencial e Weibull .....	18
2.1.3.1. Modelo Exponencial.....	18
2.1.3.2. Modelo Weibull.....	19
2.1.4. Inferência dos modelos.....	19
2.1.4.1. Método de Máxima Verossimilhança Clássico .....	19
2.1.4.2. Intervalo de Confiança.....	20
2.1.4.3. Teste da razão de verossimilhanças.....	21
2.1.4.4. Escolha do Modelo Probabilístico.....	21
2.2. Fração de cura .....	23
2.3. Inflação de zeros ou excesso de zeros .....	24
3. Classe de modelos Weibull discreto .....	26
3.1. Modelo Weibull discreto.....	26
3.2. Modelo Weibull discreto com fração de cura (WDF).....	28
3.3. Modelo Weibull discreto com excessos de zeros.....	29
3.4. Modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ).....	31
4. Dados Simulados.....	37
4.1. Cenário 1 .....	37
4.2. Cenário 2 .....	40
4.3. Cenário 3 .....	43
4.4. Cenário 4.....	46
5. Estudo de caso.....	49
6. Considerações finais.....	54
BIBLIOGRAFIA .....	56
APÊNDICE: SCRIPT DESENVOLVIDO .....	58

## 1. Introdução

A Análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas décadas em razão ao desenvolvimento e aprimoramento de técnicas estatísticas combinado com computadores mais velozes. Essa técnica é adequada para situações de estudo em que o tempo de ocorrência de um evento é a variável resposta. A principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta (Colosimo e Giolo, 2006).

Por ser uma variável aleatória, o tempo até a ocorrência do evento de interesse pode ser representado por uma distribuição de probabilidades. Dentre uma infinidade de modelos probabilísticos utilizados para modelar dados de sobrevivência, a distribuição Weibull (Weibull, 1951) se destaca pela sua versatilidade e relativa simplicidade computacional.

Um exemplo do uso do modelo Weibull em dados de sobrevivência pode ser visto em Garcia (2013), que aplicou esse modelo para descrever um conjunto de dados referentes a pacientes submetidos à intervenção coronária percutânea (angioplastia transluminal coronária). A intervenção coronária percutânea é um tratamento não cirúrgico, porém invasivo, das obstruções das artérias coronárias por meio de cateter balão, com o objetivo de aumentar o fluxo de sangue para o coração. Nesse trabalho o óbito do paciente foi definido como o evento de interesse, com o tempo de sobrevivência medido em semanas.

Apesar do tempo discreto, Garcia (2013) obteve bons resultados ao ajustar um modelo de regressão Weibull contínuo com fração de cura nos dados dos pacientes submetidos à intervenção coronária percutânea. No entanto, Nakano e Carrasco (2006) estudaram as consequências do uso de um modelo contínuo em dados discretos e mostraram que há situações em que pode não ser razoável a utilização de um modelo contínuo em dados discretos. Além disso, dados sobre o tempo de sobrevivência de pacientes submetidos à intervenção coronária percutânea, preventivo e/ou tratamento contra o infarto agudo do miocárdio apresentam, em geral, duas características peculiares: uma parcela desses pacientes vem a óbito logo após a intervenção e outra parcela apresenta sucesso no tratamento. Essa peculiaridade sugere que, após a intervenção, existam três populações distintas: 1. Pacientes que morreram por complicações que surgiram devido à intervenção; 2. Pacientes cuja intervenção pode ter alterado seu tempo de sobrevivência padrão (mas não necessariamente foram curados) e; 3. Pacientes cuja intervenção foi bem-sucedida (curados).

Em se tratando da modelagem dos tempos de sobrevivência desses pacientes, aqueles pacientes que vierem a óbito devido à intervenção apresentarão tempo de sobrevivência zero (não sobreviveram por uma semana completa). Se essa intervenção é de alto risco, a taxa de mortes devido a intervenção pode ser maior do que o valor que o modelo probabilístico usual irá prever. Nesta situação, um modelo com excessos de zeros (ou modelos inflacionados de zeros) pode ser a solução. Por outro lado, aqueles pacientes cuja intervenção foi bem-sucedida, ou seja, foram curados, não morrerão devido a um infarto agudo do miocárdio. Nesse caso, esses pacientes nunca irão experimentar o evento de interesse e os modelos tradicionais utilizados em análise de sobrevivência podem não ser

adequados, pois os mesmos não assumem a existência de indivíduos curados. Os modelos com fração de cura preveem a existência de indivíduos curados na população e é uma opção da modelagem do tempo de sobrevivência nesta situação. O modelo de Berkson e Gage (1952) tem sido bastante utilizado na modelagem de tempos de sobrevivência com fração de cura.

Neste contexto, a proposta deste trabalho de dissertação foi formular um modelo de sobrevivência discreto que incorpore fração de cura e excesso de zeros e aplicá-lo no conjunto de dados de pacientes submetidos à intervenção coronária percutânea, preventivo e/ou tratamento contra o infarto agudo do miocárdio. Devido á sua importância na modelagem de dados de sobrevivência, adotou-se nesse trabalho uma versão discreta da distribuição Weibull contínua (Weibull, 1951) proposta por Nakagawa e Osaki (1975). O modelo Weibull discreto com fração de cura e excessos de zeros foi formulado dentro de um contexto de análise de sobrevivência e seus parâmetros foram estimados numericamente através do software R (R Core Team, 2015). O modelo proposto foi ilustrado em dados simulados e no conjunto de dados de pacientes submetidos à intervenção coronária percutânea, preventivo e/ou tratamento contra o infarto agudo do miocárdio.

## 2. Revisão Bibliográfica

### 2.1. Conceitos Básicos em Análise de Sobrevida

A análise de sobrevivência é composta de um conjunto de métodos estatísticos utilizados na análise de dados para os quais a variável resposta é o tempo até a ocorrência de um evento de interesse. A unidade de estudo, na maioria das vezes, é o indivíduo. O tempo de início do estudo deve ser precisamente definido.

Por se tratar de uma observação temporal, a variável de interesse, em alguns casos, pode ter sua medição interrompida, seja pela ausência da observação do evento, perda de acompanhamento do indivíduo estudado; término do tempo de estudo; ou o indivíduo ser retirado do estudo por motivos alheios ao de interesse. Nesses casos, de observação parcial da resposta, os dados são denominados censurados. Fernandes (2014)

Mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser utilizados na análise estatística, pois mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida de pacientes. Além disso, a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viesadas.

De uma forma geral, as censuras são divididas em três tipos:

**Censura à direita:** ocorre quando o tempo exato de ocorrência do evento de interesse não é conhecido, mas sabe-se que é maior (se encontra à direita) do que o tempo registrado.

**Censura à esquerda:** ocorre quando o evento de interesse ocorreu antes do tempo registrado. Ele surge quando o evento de interesse já ocorreu quando o indivíduo foi observado. Assim na censura à esquerda, o tempo exato de ocorrência do evento de interesse não é conhecido, mas sabe-se que é menor (se encontra à esquerda) do que o tempo registrado.

**Censura intervalar:** ocorre quando o tempo exato de ocorrência do evento de interesse não é conhecido e a única informação disponível é que esse evento aconteceu em um determinado intervalo de tempo.

Em geral, o mecanismo de censura a direita é a situação frequentemente encontrada em estudos envolvendo dados biomédicos. Esse tipo de censura pode ainda ser classificado em três tipos:

- Censura (à direita) do Tipo I: quando o período de estudo é pré-fixado, portanto são censurados os indivíduos que não experimentaram o evento de interesse até esse tempo.
- Censura (à direita) do Tipo II: quando a quantidade de observações censuradas (ou não censuradas) é pré-fixada, portanto são censurados os indivíduos que não experimentaram o evento de interesse até a ocorrência da quantidade de observações pré-fixada.

- Censura (à direita) aleatória: é o tipo de censura mais geral e engloba as demais como casos particulares. Ocorre quando um ou mais indivíduos não serão acompanhados até o final do estudo, ou ainda, quando esses experimentaram o evento de interesse por motivos distintos do interesse do estudo (por exemplo, o desfecho é a morte por infarto agudo do miocárdio e o paciente morre em um acidente de trânsito). A censura aleatória ocorre de forma natural e aleatória, sem a interferência do pesquisador.

Em análise de sobrevivência, segundo Colosimo e Giolo (2006), o par  $(t_i, \delta_i)$  representa os dados relativos ao indivíduo  $i$  ( $i = 1, \dots, n$ ), sendo  $t_i$  o tempo registrado (tempo até o evento de interesse ou censura) e  $\delta_i$  a variável indicadora de falha/censura, em que:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo observado;} \\ 0, & \text{se } t_i \text{ é um tempo censurado.} \end{cases}$$

### 2.1.1. Descrição do tempo de sobrevivência

Os tempos de sobrevivência estão sujeitos a variações aleatórias e, como toda variável aleatória, possuem uma distribuição de probabilidade. Sendo  $T$ , uma variável aleatória representando o tempo transcorrido até a observação do evento de interesse, em análise de sobrevivência,  $T$  é geralmente descrita através de sua função de sobrevivência e risco, cujas definições são apresentadas a seguir.

#### 2.1.1.1. Função de Sobrevivência

A função de sobrevivência,  $S(t)$ , representa a probabilidade de um indivíduo não experimentar o evento de interesse até o tempo  $t$ , ou seja, é a probabilidade de um indivíduo sobreviver ao tempo  $t$ . No caso em que  $T$  é uma variável aleatória contínua, a função de sobrevivência é definida como:

$$S(t) = P(T > t) = 1 - F(t) = 1 - P(T \leq t) = \int_t^{\infty} f(u) du, t > 0 \quad (1)$$

em que  $F(\cdot)$  é a função de distribuição acumulada e  $f(\cdot)$  é função densidade de probabilidade de  $T$ . Quando  $T$  é uma variável aleatória contínua, a função de sobrevivência  $S(t)$  é absolutamente contínua e estritamente decrescente (LAWLESS, 1982).

Quando  $T$  é uma variável aleatória discreta, que assume os valores  $t = 0, 1, 2, \dots$ , a função de sobrevivência  $S(t)$  é definida por

$$S(t) = P(T > t) = \sum_{k=t+1}^{\infty} P(T = k), t = 0, 1, 2, \dots \quad (2)$$

A função de sobrevivência tem as propriedades

$$\lim_{t \rightarrow \infty} S(t) = 0 \quad \text{e} \quad \lim_{t \rightarrow 0} S(t) = 1, \quad (3)$$

que podem ser observadas nas Figuras 1 e 2, respectivamente.

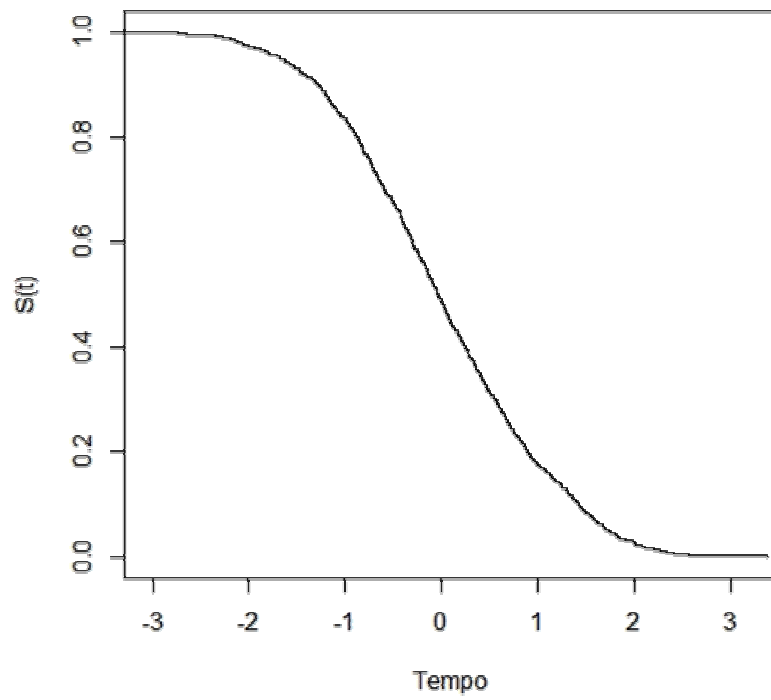


Figura 1 – Forma típica da função de sobrevivência contínua

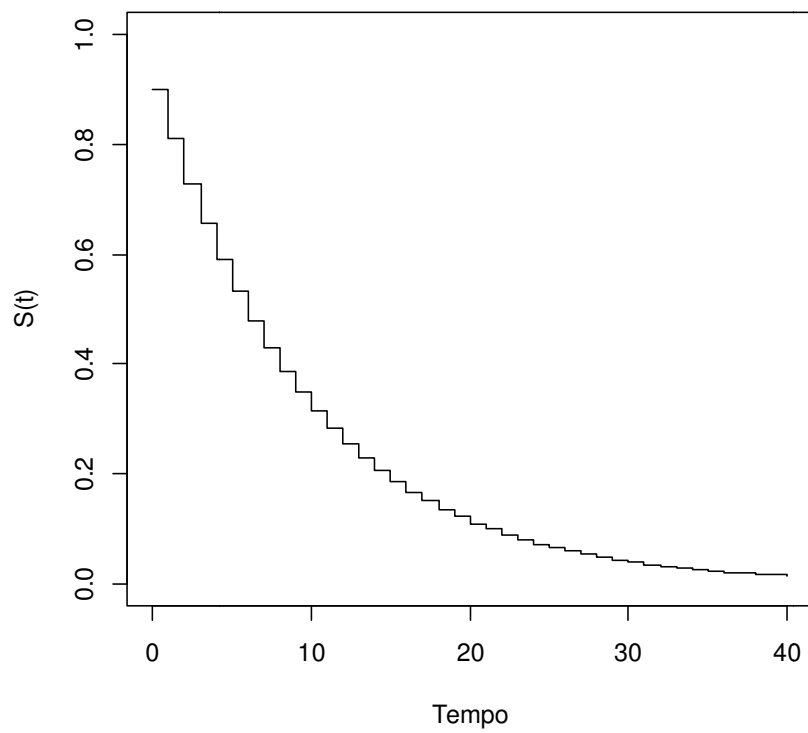


Figura 2 – Forma típica da função de sobrevivência discreta

Existem situações em que (3) não ocorre, sendo que o  $\lim_{t \rightarrow \infty} S(t) = p$ , em que  $p$  uma probabilidade qualquer diferente de 0, resultando em uma função de sobrevivência imprópria. Nessa situação utilizam-se modelos de sobrevivência com fração de cura para modelar os dados. Esse assunto será abordado na Seção 2.2. Um exemplo dessa situação pode ser observado na Figura 3.

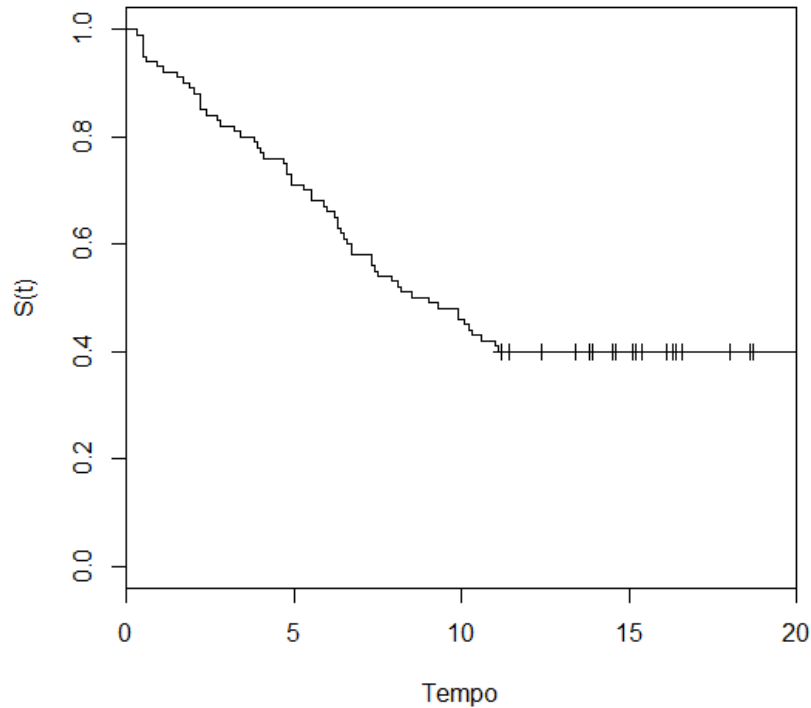


Figura 3 - Função de sobrevivência imprópria

#### 2.1.1.2. Função de Taxa de falha ou Função de Risco

Outra maneira de representar o tempo  $T$  é através da função taxa de falha ou função de risco. A função de risco representa o risco instantâneo de ocorrência do evento de interesse no tempo  $t$  condicional à sobrevivência até o tempo  $t$ .

Para uma variável contínua  $T$ , a função de risco é definida por:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}, \quad (4)$$

em que  $f(t)$  e  $S(t)$  são, respectivamente, a função densidade de probabilidades e a função de sobrevivência de  $T$ . No caso em que  $T$  é uma variável contínua,  $h(t)$  pode assumir qualquer valor real positivo.

Para o caso em que  $T$  é uma variável aleatória discreta, a função de risco é igual a zero, exceto nos pontos onde ocorrer falha. Por ser uma probabilidade, a função de risco é limitada no intervalo  $[0, 1]$  e é definida por:



$$h(t) = P(T = t | T \geq t) = \frac{P(T=t)}{P(T \geq t)} = \frac{P(T=t)}{P(T > t) + P(T=t)} = \frac{p(t)}{S(t) + p(t)}, \quad t=0, 1, \dots \quad (5)$$

em que  $p(t)$  e  $S(t)$  são, respectivamente a função (distribuição) de probabilidades e a função de sobrevivência de  $T$ .

A função de risco, definida em (4) e (5), é útil para descrever a distribuição do tempo de vida de pacientes. Ela representa a forma em que a taxa instantânea de falha, ou a força de mortalidade, muda com o tempo.

A função de risco é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente. Dessa forma, a modelagem da função de risco é um importante método para dados de sobrevivência (Colosimo e Giolo, 2006), pois pode ter forma crescente, decrescente, constante ou não monótona. Uma função crescente indica, por exemplo, que o risco de morte do paciente aumenta com o transcorrer do tempo, mostrando que existe um efeito gradual de envelhecimento. A função constante indica que a taxa de falha (ou risco) não se altera com o passar do tempo (como se um equipamento não envelhecesse). A função decrescente mostra que o risco diminui à medida que o tempo passa.

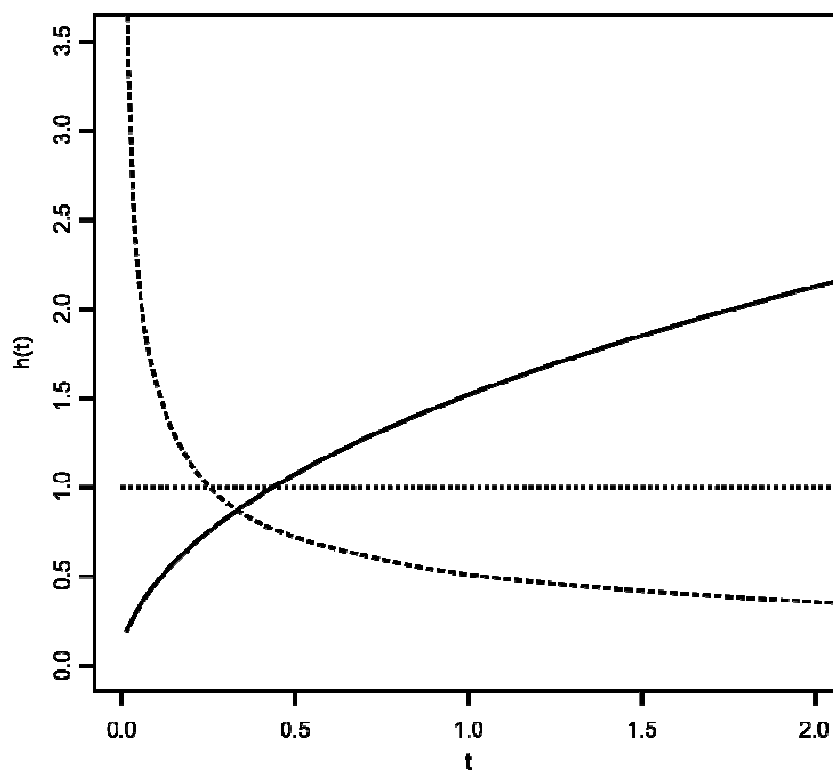


Figura 4— Exemplo de Funções de risco (crescente, decrescente e constante)

### 2.1.2. Estimação empírica da função de sobrevivência (Kaplan-Meier)

O procedimento inicial para dados com censura é encontrar a estimativa para a função de sobrevivência e a partir dela estimar as estatísticas de interesse. Para obter essa estimativa podem ser considerados três estimadores: Kaplan-Meier, Nelson-Aalen e Tabela de vida. Neste estudo será utilizado o estimador de Kaplan-Meier por ser o estimador de máxima verossimilhança da função de sobrevivência (Colosimo e Giolo, 2006).

O estimador não paramétrico de Kaplan-Meier, também conhecido como estimador limite-produto, é baseado em um número de intervalos igual ao número de falhas distintos e se reduz a estimar  $\hat{q}_j$ , probabilidade de um indivíduo sofrer o evento de interesse no intervalo  $[t_{j-1}, t_j)$  sabendo que ele não falhou até  $t_{j-1}$  e considerando  $t_0 = 0$ , que é dado por:

$$\hat{q}_j = \frac{\text{n}^\circ \text{ de falhas em } t_{j-1}}{\text{n}^\circ \text{ de observações sob risco em } t_{j-1}} \quad (6)$$

para  $j = 1, \dots, k + 1$ , em que  $t_{k+1} = \infty$

O estimador de Kaplan-Meier incorpora a informação de todas as observações disponíveis, tanto censuradas quanto não censuradas. Sua expressão geral pode ser apresentada após estas considerações preliminares:

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

O estimador Kaplan-Meier é definido como:

$$\hat{S}_{KM}(t) = \prod_{j:t_{(j)} < t} \left( \frac{n_j - d_j}{n_j} \right). \quad (7)$$

Quando os dados apresentam algum valor repetido ou censurado, então  $k \leq n$ ; caso contrário,  $k = n$ .

Por ser um estimador de máxima verossimilhança,  $\hat{S}_{KM}(t)$  é um estimador fracamente consistente e não viciado para amostras grandes. Breslow e Crowley (1974) mostram a consistência e a normalidade assintótica de  $\hat{S}_{KM}(t)$ . Logo,  $\hat{S}_{KM}(t) \sim N\left(S(t), \text{Var}\left(\hat{S}_{KM}(t)\right)\right)$ .

Para que seja possível construir intervalos de confiança e realizar testes de hipóteses para  $S(t)$ , é necessário avaliar a precisão do estimador. A variância assintótica desse estimador é estimada pela fórmula de Greenwood expressa por:

$$\widehat{Var}(\hat{S}_{KM}(t)) = (\hat{S}_{KM}(t))^2 \sum_{j:t_{(j)} < t} \frac{d_j}{n_j(n_j - d_j)} \quad (8)$$

Um intervalo de confiança para  $S(t)$  considerando  $100(1 - \alpha)\%$  de confiança é dado por:

$$\hat{S}_{KM}(t) \pm z_{\alpha/2} \sqrt{\widehat{Var}[\hat{S}_{KM}(t)]}, \quad (9)$$

em que  $\alpha/2$  denota o  $\alpha/2$ -quantil da distribuição Normal padrão,  $\hat{S}_{KM}(t)$  é o estimador de Kaplan-Meier e  $\widehat{Var}[\hat{S}_{KM}(t)]$  é o estimador da variância do estimador de Kaplan-Meier.

Deve-se observar que o intervalo de confiança de  $S(t)$  não é muito indicado para amostras pequenas e, se  $t$  é um valor extremo, pode incluir valores fora do intervalo  $[0,1]$ . Uma alternativa é utilizar uma transformação para  $S(t)$ , como por exemplo,  $g(\hat{S}_{KM}(t)) = \log[-\log(\hat{S}_{KM}(t))]$ , sugerida por Kalbfleish e Prentice (2002).

### 2.1.3. Modelos Exponencial e Weibull

Embora exista uma série de modelos probabilísticos utilizados em análise de dados de sobrevivência, alguns deles ocupam uma posição de destaque por sua comprovada adequação a várias situações práticas. Entre estes modelos, é possível citar o Exponencial e o Weibull.

#### 2.1.3.1. Modelo Exponencial

A distribuição exponencial apresenta um único parâmetro e é a única distribuição contínua univariada que se caracteriza por ter uma função de risco de falha constante. Por isso é aplicável quando o tempo é curto. Quanto maior o risco, menor o tempo médio de sobrevivência e menor a variabilidade deste em torno da média.

A função densidade de probabilidades de uma variável  $T$  com distribuição exponencial é dada por:

$$f(t) = \frac{1}{\lambda} \exp\left\{-\frac{t}{\lambda}\right\}, \quad (10)$$

sendo que  $t \geq 0$  e  $\lambda > 0$  é o parâmetro de escala e a esperança de  $T$ .

As funções de sobrevivência e risco são dadas, respectivamente, por:

$$S(t) = \exp\left\{-\frac{t}{\lambda}\right\} \quad (11)$$

e

$$h(t) = \frac{1}{\lambda}, \quad (12)$$

com  $t \geq 0$  e  $\lambda > 0$ .

### 2.1.3.2. Modelo Weibull

Popular em aplicações práticas, a distribuição Weibull (Weibull, 1951) apresenta uma grande variedade de formas. Sua função de risco é monótona, isto é, ela é crescente, decrescente ou constante.

Uma variável aleatória  $T$  com distribuição Weibull tem a função densidade de probabilidade expressa por:

$$f(t) = \frac{\beta}{\lambda^\beta} t^{\beta-1} \exp\left\{-\left(\frac{t}{\lambda}\right)^\beta\right\}, \quad (13)$$

sendo que  $t \geq 0$ ,  $\lambda > 0$  é o parâmetro de escala e  $\beta > 0$  é o parâmetro de forma. As funções de sobrevivência e risco são, respectivamente, expressas por:

$$S(t) = \exp\left\{-\left(\frac{t}{\lambda}\right)^\beta\right\} \quad (14)$$

e

$$h(t) = \frac{\beta}{\lambda^\beta} t^{\beta-1} \quad (15)$$

com  $t \geq 0$ ,  $\lambda > 0$  e  $\beta > 0$ .

A função de risco,  $h(t)$ , da Weibull é estritamente crescente para  $\beta > 1$ , estritamente decrescente para  $\beta < 1$  e constante para  $\beta = 1$ , quando o modelo Weibull se reduz à Exponencial.

### 2.1.4. Inferência dos modelos

Em cada estudo envolvendo tempos de falha, os parâmetros devem ser estimados a partir das observações amostrais, para que o modelo fique determinado e, assim, seja possível responder às perguntas de interesse. É possível citar dois importantes métodos de estimação:

- **Método dos Mínimos Quadrados:** inapropriado para estudos de tempos de vida, pois não é capaz de incorporar a informação das observações censuradas no seu processo de estimação.
- **Método de Máxima Verossimilhança Clássico:** escolhe o parâmetro que melhor explique a amostra observada. Esse será o método utilizado nesse trabalho, pois ele é capaz de incorporar as observações censuradas.

#### 2.1.4.1. Método de Máxima Verossimilhança Clássico

Baseado nos resultados obtidos na amostra, tem-se como objetivo descobrir qual a distribuição e os parâmetros com maior possibilidade de ter gerado tal amostra. Usualmente é utilizada a função de máxima verossimilhança que é dada pela seguinte equação:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n f(t_i, \boldsymbol{\theta}). \quad (16)$$

No entanto, na equação acima o tempo de censura é desconsiderado. Portanto, é necessário utilizar uma variável indicadora.

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo observado;} \\ 0, & \text{se } t_i \text{ é um tempo censurado.} \end{cases}$$

A contribuição de cada elemento da amostra para a função de verossimilhança é dada por:

$$\begin{cases} f(t_i, \boldsymbol{\theta}), & \text{se } t_i \text{ é tempo de falha} \\ S(t_i, \boldsymbol{\theta}), & \text{se } t_i \text{ é tempo de censura} \end{cases}$$

Considerando todos os mecanismos de censura à direita, sob a suposição de que eles são não informativos (não carregam informação sobre os parâmetros), a função de verossimilhança, a menos de uma constante, é representada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^r [f(t_i, \boldsymbol{\theta})] \prod_{i=r+1}^n [S(t_i, \boldsymbol{\theta})] \quad (17)$$

ou, equivalentemente, por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n [f(t_i, \boldsymbol{\theta})]^{\delta_i} [S(t_i, \boldsymbol{\theta})]^{1-\delta_i}, \quad (18)$$

em que  $f(\cdot)$  é a função (densidade) de probabilidades e  $S(\cdot)$  a função de sobrevivência. Os estimadores de máxima verossimilhança são os valores de  $\boldsymbol{\theta}$  que maximizam  $L(\boldsymbol{\theta})$ .

#### 2.1.4.2. Intervalo de Confiança

Para a construção do intervalo de confiança é utilizada a distribuição assintótica do estimador de máxima verossimilhança  $\hat{\boldsymbol{\theta}}$ . Para grandes amostras, essa propriedade estabelece, sob certas condições de regularidade, que  $\hat{\boldsymbol{\theta}}$  tem distribuição Normal multivariada de média  $\boldsymbol{\theta}$  e matriz de variância-covariância  $V(\hat{\boldsymbol{\theta}})$ , isto é,

$$\hat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, \text{Var}(\hat{\boldsymbol{\theta}})),$$

sendo  $k$  a dimensão de  $\boldsymbol{\theta}$ .

Sob certas condições de regularidade (Casella, 2002), a matriz de variância-covariância é aproximadamente o negativo da inversa da matriz de informação de Fisher:

$$\text{Var}(\hat{\boldsymbol{\theta}}) \approx - [I_F(\boldsymbol{\theta})]^{-1},$$

em que  $I_F(\theta) = E \left[ \left( \frac{\partial \log L(\theta)}{\partial \theta} \right)^2 \right]$ .

#### 2.1.4.3. Teste da razão de verossimilhanças

Desde que o modelo completo apresente ajuste adequado, pode ser de interesse testar a hipótese para verificar se um modelo mais simples (com um número reduzido de parâmetros) pode ser utilizado para modelar os dados de sobrevivência. Neste caso, o problema se resume em testar a hipótese  $H_0: \theta = \theta_0$ , em que  $\theta$  pode ser um único parâmetro ou um vetor de parâmetros. Essa hipótese pode ser testada através do Teste da Razão de Verossimilhanças (TRV).

O TRV é realizado a partir dos seguintes ajustes: 1) modelo generalizado (modelo completo) e obtenção do valor do logaritmo de sua função de verossimilhança  $\log L(\hat{\theta}_C)$  e; 2) modelo de interesse (modelo reduzido) e obtenção do valor do logaritmo da sua função de verossimilhança  $\log L(\hat{\theta}_R)$ . A partir desses valores é possível calcular a estatística do teste da razão de verossimilhanças:

$$TRV = -2 \log \left[ \frac{L(\hat{\theta}_R)}{L(\hat{\theta}_C)} \right] = 2 [\log L(\hat{\theta}_C) - \log L(\hat{\theta}_R)], \quad (19)$$

que sob  $H_0$ , tem aproximadamente uma distribuição Qui-quadrado com  $p$  graus de liberdade, sendo  $p$  a diferença do número de parâmetros ( $\hat{\theta}_C$  e  $\hat{\theta}_R$ ) dos modelos comparados. A um nível de significância  $\alpha$ , valores do TRV superiores ao valor tabelado da distribuição  $X_{p,1-\alpha}^2$  indicam a rejeição de  $H_0$ . Aqui,  $X_{p,1-\alpha}^2$  é o quantil  $(1 - \alpha)$  de uma distribuição Qui-quadrado com  $p$  graus de liberdade.

#### 2.1.4.4. Escolha do Modelo Probabilístico

A escolha do modelo é extremamente importante, pois o método de máxima verossimilhança somente pode ser aplicado após a definição de um suposto modelo probabilístico adequado para os dados.

Uma proposta empírica consiste em ajustar os modelos probabilísticos conhecidos e, com base na comparação entre valores estimados e observados, decidir qual deles “melhor” explica os dados amostrais. A forma mais simples e eficiente de selecionar um modelo é por meio de técnicas gráficas. Entretanto teste de hipóteses com modelos encaixados (Cox e Hinkley, 1974) também podem ser utilizados.

O primeiro método gráfico consiste na comparação da função de sobrevivência do modelo proposto com a função de sobrevivência estimada pelo método de Kaplan-Meier. Para realizar este procedimento é necessário ajustar os modelos propostos ao conjunto de dados para obter as estimativas dos parâmetros e então estimar a função de sobrevivência

dos respectivos modelos. O modelo adequado é aquele em que sua curva de sobrevivência mais se aproximar da curva de sobrevivência estimada pelo método de Kaplan-Meier.

Para uma melhor visualização durante o processo de decisão, a curva de sobrevivência de Kaplan-Meier e a curva de sobrevivência estimada pelo modelo podem ser plotadas no mesmo gráfico como ilustra o exemplo da Figura 5.

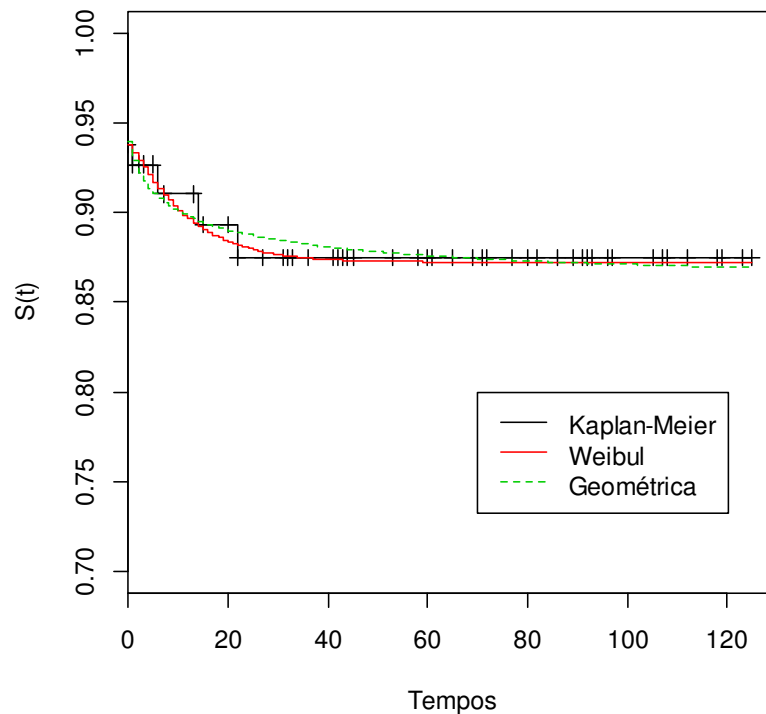


Figura 5– Curvas de sobrevivência estimada por Kaplan-Meier e curvas de sobrevivência estimada dos modelos: Weibull e Exponencial

A verificação da qualidade do ajuste pode também ser feita comparando as curvas de sobrevivência teórica e empírica através do teste não paramétrico de Kolmogorov-Smirnov, cuja estatística do teste é dada por:

$$KS = \underset{t: t \text{ não censurado}}{\text{Max}} |\hat{S}_{km}(t) - \hat{S}_{modelo\ proposto}(t)|$$

No teste de Kolmogorov-Smirnov, o ajuste do modelo é rejeitado, sob um nível de significância  $\alpha$ , se a estatística KS for maior que o quantil  $\alpha$  do teste de Kolmogorov-Smirnov. A Tabela 1 a seguir apresenta os quantis do teste de Kolmogorov-Smirnov.

Tabela 1 - Quantis do teste de Kolmogorov-Smirnov.

$\alpha$	0,80	0,90	0,95	0,98	0,99		0,80	0,90	0,95	0,98	0,99
$n' = 1$	,900	,950	,975	,990	,995	$n' = 21$	,226	,259	,287	,321	,344
2	,684	,776	,842	,900	,929	22	,221	,253	,281	,314	,337
3	,565	,636	,708	,785	,829	23	,216	,247	,275	,307	,330
4	,493	,565	,624	,689	,734	24	,212	,242	,269	,301	,323
5	,447	,509	,563	,627	,669	25	,208	,238	,264	,295	,317
6	,410	,468	,519	,577	,617	26	,204	,233	,259	,290	,311
7	,381	,436	,483	,538	,576	27	,200	,229	,254	,284	,305
8	,358	,410	,454	,507	,542	28	,197	,225	,250	,279	,300
9	,339	,387	,430	,480	,513	29	,193	,221	,246	,275	,295
10	,323	,369	,409	,457	,489	30	,190	,218	,242	,270	,290
11	,308	,352	,391	,437	,468	31	,187	,214	,238	,266	,285
12	,296	,338	,375	,419	,449	32	,184	,211	,234	,262	,281
13	,285	,325	,361	,404	,432	33	,182	,208	,231	,258	,277
14	,275	,314	,349	,390	,418	34	,179	,205	,227	,254	,273
15	,266	,304	,338	,377	,404	35	,177	,202	,224	,251	,269
16	,258	,295	,327	,366	,392	36	,174	,199	,221	,247	,265
17	,250	,286	,318	,355	,381	37	,172	,196	,218	,244	,262
18	,244	,279	,309	,346	,371	38	,170	,194	,215	,241	,258
19	,237	,271	,301	,337	,361	39	,168	,191	,213	,238	,255
20	,232	,265	,294	,329	,352	40	,165	,189	,210	,235	,252
						Aproxima	1,07	1,22	1,36	1,52	1,63
						ção para	$\frac{1,07}{\sqrt{n'}}$	$\frac{1,22}{\sqrt{n'}}$	$\frac{1,36}{\sqrt{n'}}$	$\frac{1,52}{\sqrt{n'}}$	$\frac{1,63}{\sqrt{n'}}$
						$n' > 40$					

**Nota:** Uma melhor aproximação pode ser alcançada para  $n' > 40$  se  $(n' + \sqrt{n'}/10)^{0,5}$  é usado no lugar de  $\sqrt{n'}$  no denominador. Aqui,  $n'$  é o número de observações não censuradas na amostra, isto é

$$n' = \sum_{i=1}^n \delta_i .$$

Fonte: (Siegel e Castellan, 2006)

## 2.2. Fração de cura

Os principais modelos de análise de sobrevivência têm como pressuposto o fato de que os indivíduos do estudo irão, em algum momento, experimentar o evento de interesse definido.

Como citado na Seção 2.1.1, a função de sobrevivência é considerada própria quando todos os indivíduos são suscetíveis ao evento de interesse. Entretanto, existem situações que, para uma proporção de indivíduos, o evento de interesse não ocorrerá. A função é considerada imprópria quando à medida que o tempo tende a infinito a função não tende a zero. Isso pode indicar que existe uma proporção de indivíduos curados ou imunes. Dessa forma, a função de sobrevivência, juntamente com suas propriedades, são muito importantes para a identificação de dados com a presença de indivíduos curados.



Nesse caso, considerar os modelos de sobrevivência usuais, que assumem que a função de sobrevivência converge para zero quando a variável tempo tende a infinito (função de sobrevivência própria), podem não ser adequados. Para modelar esse tipo de dados, modelos com fração de cura são mais apropriados (Fachini, 2011).

Os indivíduos não suscetíveis, definidos como curados, aparecem na base de dados como observações censuradas, visto que o evento de interesse não é observado.

A presença de indivíduos curados na base de dados é verificada ao construir o gráfico da função de sobrevivência empírica, estimada pelo método de Kaplan-Meier. Verifica-se no gráfico o comportamento da cauda direita e caso ela permaneça de maneira constante em um nível acima de zero por um período grande, conclui-se que há proporção de indivíduos curados (Fachini, 2011). A modelagem de fração de cura para o caso univariado pode ser abordada seguindo a metodologia introduzida por Berkson e Gage, (1952) que considera a construção de uma função de sobrevivência populacional na forma de mistura de duas populações.

Quando a metodologia de proporção de curados introduzida por Berkson e Gage (1952) é adotada para modelar, utiliza-se a função de sobrevivência populacional em forma de mistura, uma vez que a população é dividida em duas partes: indivíduos suscetíveis (NC: não curados) ao evento de interesse, cujo tempo de vida é representado por uma variável aleatória  $X$  e; indivíduos não suscetíveis (C: curados) ao evento de interesse, cujo tempo de sobrevivência é representado por uma variável aleatória  $Y$  degenerada no infinito (isto é, uma variável em que  $P(Y > y) = 1, \forall y$ ). Assim, a função de sobrevivência populacional em forma de mistura é dada por:

$$\begin{aligned} S(t) &= \phi S_Y(t) + (1 - \phi) S_X(t) \\ &= \phi + (1 - \phi) S_X(t), \end{aligned} \tag{20}$$

sendo  $S_X(t)$  a função de sobrevivência de  $X$  (função de sobrevivência própria), associada aos indivíduos não curados e;  $\phi$  é o parâmetro de fração de cura, que representa a probabilidade do indivíduo ser não suscetível (curado).

### 2.3. Inflação de zeros ou excesso de zeros

Se tratando de dados de sobrevivência discretos, é possível ter situações em que há uma ocorrência muito grande de observações iguais a zero. Nesta situação, um modelo com inflação de zeros será mais apropriado para o ajuste desses dados do que um modelo padrão. Uma aplicação do modelo Geométrico com inflação de zeros aplicados em dados de confiabilidade, foi discutido por Carrasco et al. (2012).

Uma distribuição inflacionada de zeros pode ser obtida através de uma mistura de duas variáveis, por exemplo,  $X$  e  $Z$ :  $X$  tem distribuição segundo o modelo que representa o tempo de sobrevivência de indivíduos suscetíveis ao evento de interesse e  $Z$  variável

degenerada no ponto zero (isto é, uma variável cuja probabilidade de assumir o valor zero é igual a 1).

Assim, assumindo que  $T$  é uma variável aleatória discreta com excessos de zeros, tem-se que a distribuição de probabilidades de  $T$  é dada por:

$$\begin{aligned}
 P(T = t) &= \rho P(Z = t) + (1 - \rho)P(X = t) \\
 &= \begin{cases} \rho + (1 - \rho)P(X = t), & \text{se } t = 0 \\ (1 - \rho)P(X = t), & \text{se } t = 1, 2, 3, \dots \end{cases} \quad (21)
 \end{aligned}$$

Que resulta na seguinte função de sobrevivência:

$$S_Z(t) = (1 - \rho)P(X > t) = (1 - \rho)S_X(t) \quad (22)$$

em que  $t \geq 0$  e  $0 \leq \rho \leq 1$  é parâmetro que modela o excesso de zeros e  $S_X(\cdot)$  é a função de sobrevivência de  $X$ .

### 3. Classe de modelos Weibull discreto

#### 3.1. Modelo Weibull discreto

A distribuição Weibull é uma distribuição bem consagrada para a modelagem de dados de sobrevivência. Porém, como mostrado por Nakano e Carrasco (2006), nem sempre é plausível usar um modelo contínuo quando os dados são discretos.

O modelo Weibull discreto proposto por Nakagawa e Osaki (1975) corresponde ao modelo Weibull contínuo, e sua distribuição de probabilidades é obtida através da variável aleatória  $T = [X]$ , em que  $[X]$  representa a parte inteira de  $X$ , isto é, o maior inteiro menor ou igual a  $X$  (Nakano e Carrasco, 2006). Se  $X \sim \text{Weibull}(\beta, \lambda)$ , com função de distribuição  $F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^\beta}$ , então a distribuição de probabilidades da Weibull discreta é dada por:

$$\begin{aligned} P_{WD}(t) &= P[T = t] = P[t \leq X < t + 1] \\ &= S_X(t) - S_X(t + 1) \\ &= e^{-\left(\frac{t}{\lambda}\right)^\beta} - e^{-\left(\frac{t+1}{\lambda}\right)^\beta} \\ &= q^{t^\beta} - q^{(t+1)^\beta}, \quad t=0, 1, 2, \dots \end{aligned} \quad (23)$$

onde  $q = e^{-\left(\frac{1}{\lambda^\beta}\right)}$ . Note que,  $0 < q < 1$ .

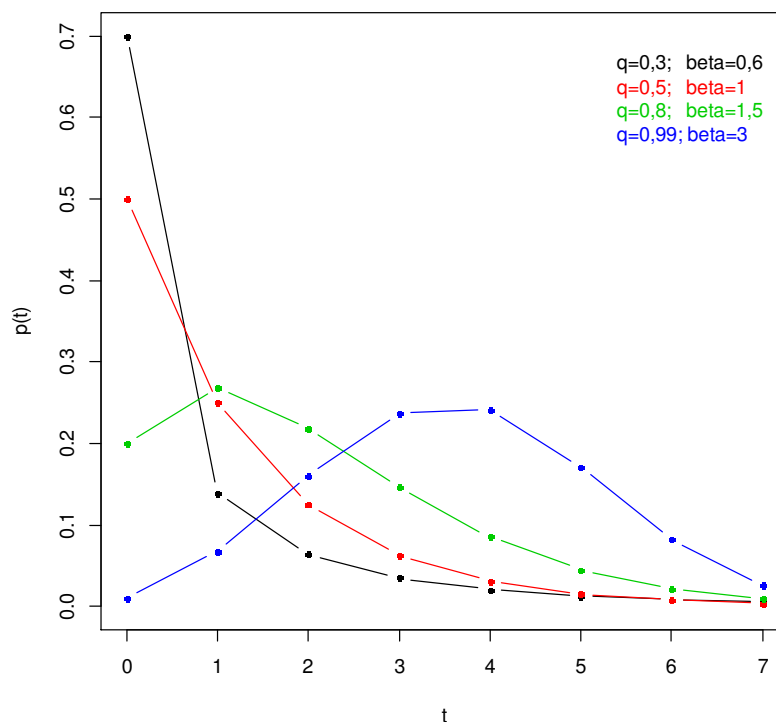


Figura 6 - Distribuição de probabilidades da Weibull discreta para diversos valores dos parâmetros  $q$  e  $\beta$ .

A função de sobrevivência é definida como:

$$\begin{aligned}
 S_{WD}(t) &= P[T > t] = \sum_{k=t+1}^{\infty} P_{WD}(t) \\
 &= \sum_{k=t+1}^{\infty} q^{t^\beta} - q^{(t+1)^\beta} \\
 &= (q^{(t+1)^\beta} - q^{(t+2)^\beta}) + (q^{(t+2)^\beta} - q^{(t+3)^\beta}) + \dots \\
 &= q^{(t+1)^\beta}, \quad t=0, 1, 2, \dots
 \end{aligned} \tag{24}$$

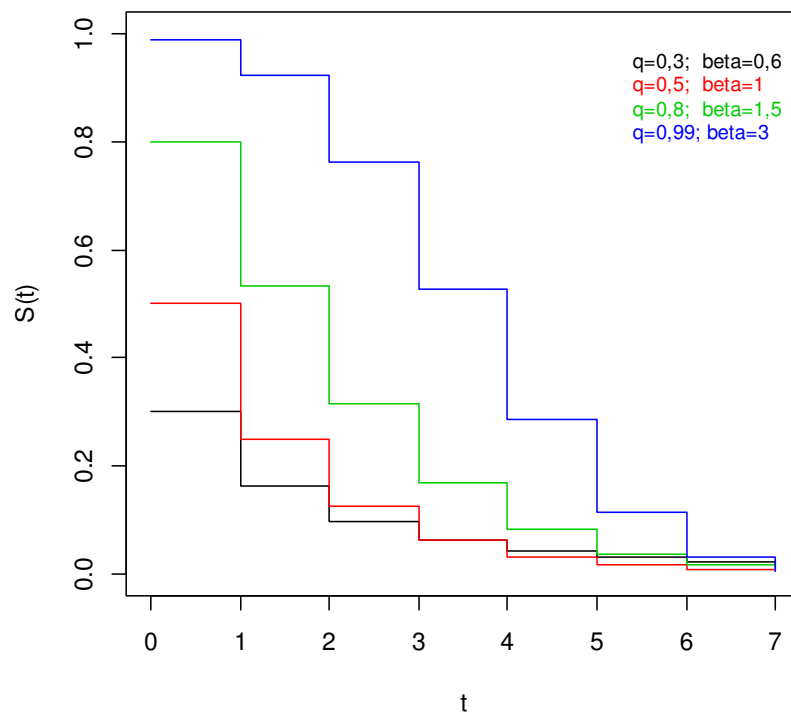


Figura 7 - Função de sobrevivência da Weibull discreta para diversos valores dos parâmetros  $q$  e  $\beta$ .

E a função risco pode ser expressa da seguinte forma:

$$\begin{aligned}
 h_{WD}(t) &= \frac{P_{WD}(t)}{S_{WD}(t) + P_{WD}(t)} \\
 &= \frac{q^{t^\beta} - q^{(t+1)^\beta}}{[q^{(t+1)^\beta}] + [q^{t^\beta} - q^{(t+1)^\beta}]} \\
 &= \frac{q^{t^\beta} - q^{(t+1)^\beta}}{q^{t^\beta}} \\
 &= 1 - q^{(t+1)^\beta - t^\beta}, \quad t=0, 1, 2, \dots
 \end{aligned} \tag{25}$$

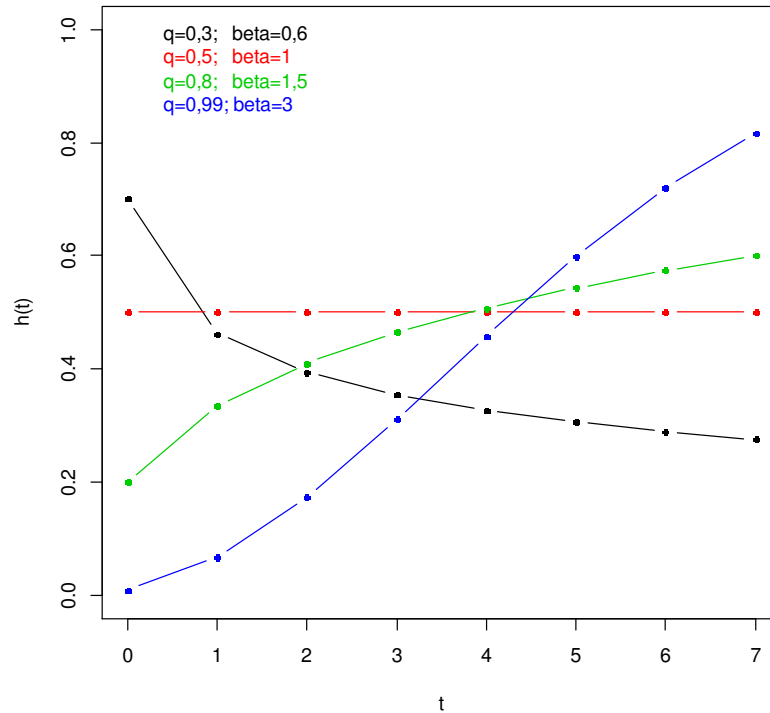


Figura 8 - Função de risco da Weibull discreta para diversos valores dos parâmetros  $q$  e  $\beta$ .

Note que, se  $\beta = 1$  o modelo se reduz à distribuição geométrica, que tem função de risco constante e igual a  $(1 - q)$ .

Em que  $\theta = (q, \beta)$  o vetor de parâmetros e  $T = (t_1, \dots, t_n)$  o vetor de valores observados com seus respectivos indicadores de censuras  $\delta = (\delta_1, \dots, \delta_n)$ , a função de verossimilhança para o modelo Weibull discreto (WD) é definida por:

$$\begin{aligned}
 L_{WD}(q, \beta | T, \delta) &= \prod_{i=1}^n [P_{WD}(t_i)]^{\delta_i} [S_{WD}(t_i)]^{1-\delta_i} \\
 &= \prod_{i=1}^n \left\{ q^{t_i^\beta} - q^{(t_i+1)^\beta} \right\}^{\delta_i} \left\{ q^{(t_i+1)^\beta} \right\}^{1-\delta_i}, \quad (26)
 \end{aligned}$$

em que  $0 < q < 1$  e  $\beta > 0$  são os parâmetros a serem estimados e  $t_i$  são os tempos observados com seus respectivos indicadores de censura  $\delta_i, i = 1, 2, \dots, n$ .

Brunello e Nakano (2015) apresentaram a inferência dos parâmetros do modelo Weibull discreto (WD) dentro de um contexto bayesiano.

### 3.2. Modelo Weibull discreto com fração de cura (WDF)

Assuma que o tempo de sobrevivência dos indivíduos não curados segue uma distribuição Weibull discreta dada por  $P_{WD}(t)$ . A partir de (20) e (24), tem-se que a função de sobrevivência do modelo Weibull discreto com fração de cura (WDF) é dada por:

$$S_{WDF}(t) = \phi + (1 - \phi)S_{WD}(t) = \phi + (1 - \phi)q^{(t+1)^\beta}, t=0, 1, 2, \dots \quad (27)$$

Assim, a distribuição de probabilidade de  $T$  é:

$$\begin{aligned} P_{WDF}(t) &= S_{WDF}(t-1) - S_{WDF}(t) \\ &= \left[ \phi + (1 - \phi)q^{(t)^\beta} \right] - \left[ \phi + (1 - \phi)q^{(t+1)^\beta} \right] \\ &= (1 - \phi) \left[ q^{t^\beta} - q^{(t+1)^\beta} \right] \end{aligned} \quad (28)$$

que resulta na seguinte função de risco:

$$h_{WDF} = \frac{P_{WDF}(t)}{S_{WDF}(t) + P_{WDF}(t)} = \frac{(1 - \phi) \left[ q^{t^\beta} - q^{(t+1)^\beta} \right]}{\phi + (1 - \phi)q^{t^\beta}} \quad (29)$$

Note que se  $\phi = 0$ , o modelo Weibull discreto com fração de cura (WDF) se reduz ao modelo Weibull discreto dado pela equação (23).

Sendo  $\theta = (q, \beta, \phi)$  o vetor de parâmetros e  $\mathbf{t} = (t_1, \dots, t_n)$  o vetor de valores observados com seus respectivos indicadores de censuras  $\delta_i = (\delta_1, \dots, \delta_n)$ , a função de verossimilhança para o modelo Weibull discreto com fração de cura (WDF) é definida por:

$$\begin{aligned} L_{WDF}(q, \beta, \phi | T, \delta) &= \prod_{i=1}^n [P_{WDF}(t_i)]^{\delta_i} [S_{WDF}(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^n \left\{ (1 - \phi) \left[ q^{t_i^\beta} - q^{(t_i+1)^\beta} \right] \right\}^{\delta_i} \\ &\quad \times \left\{ \phi + (1 - \phi)q^{(t_i+1)^\beta} \right\}^{1-\delta_i} \end{aligned} \quad (30)$$

onde  $0 < q < 1$ ,  $\beta > 0$  e  $0 \leq \phi \leq 1$  são os parâmetros a serem estimados e  $t_i$  são os tempos observados com seus respectivos indicadores de censura  $\delta_i$ ,  $i = 1, 2, \dots, n$ .

Fernandes (2014) apresentou a inferência dos parâmetros do modelo Weibull discreto com fração de cura (WDF) dentro de um contexto bayesiano.

### 3.3. Modelo Weibull discreto com excessos de zeros

A distribuição Weibull discreta com inflação de zeros é uma generalização da distribuição Weibull discreta, sendo uma combinação dessa distribuição com uma distribuição de uma variável degenerada no zero (probabilidade de zero igual a 1). A partir

de (21) e (23), tem-se que a distribuição de probabilidades do modelo Weibull discreto com excessos de zeros é dada por:

$$\begin{aligned} P_{WDZ}(t) &= (1 - \rho)P_{WD}(t) + \rho I_{(t)}\{0\} \\ &= (1 - \rho) \left( q^{t^\beta} - q^{(t+1)^\beta} \right) + \rho I_{(t)}\{0\}, \quad t = 0, 1, 2, \dots \end{aligned} \quad (31)$$

em que  $I_{(t)}\{0\}$  é uma função indicadora, que vale um quando  $t=0$  e zero quando  $t>0$ .

Note que, se  $\rho = 0$  o modelo Weibull discreto com excessos de zeros se reduz ao modelo Weibull discreto (WD) dado pela equação (23).

A partir de (22) e (24), tem-se que as funções de sobrevivência e de risco podem ser escritas, respectivamente, por

$$S_{WDZ}(t) = (1 - \rho)S_{WD}(t) = (1 - \rho)q^{(t+1)^\beta}, \quad t = 0, 1, 2, \dots \quad (32)$$

e

$$\begin{aligned} h_{WDZ}(t) &= \frac{P_{WDZ}(t)}{S_{WDZ}(t) + P_{WDZ}(t)} \\ &= \frac{(1 - \rho) \left( q^{t^\beta} - q^{(t+1)^\beta} \right) + \rho I_{(t)}\{0\}}{(1 - \rho)q^{t^\beta} + \rho I_{(t)}\{0\}}, \quad t = 0, 1, 2, \dots \end{aligned} \quad (33)$$

Sendo  $\boldsymbol{\theta} = (q, \beta, \rho)$  o vetor de parâmetros e  $\boldsymbol{T} = (t_1, \dots, t_n)$  o vetor de valores observados com seus respectivos indicadores de censuras  $\boldsymbol{\delta}_i = (\delta_1, \dots, \delta_n)$ , a função de verossimilhança para o modelo Weibull discreto com excessos de zeros é definida por:

$$\begin{aligned} L_{WDZ}(q, \beta, \rho | \boldsymbol{T}, \boldsymbol{\delta}) &= \prod_{i=1}^n [P_{WDZ}(t_i)]^{\delta_i} [S_{WDZ}(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^n \left\{ (1 - \rho) \left[ q^{t_i^\beta} - q^{(t_i+1)^\beta} \right] + \rho I_{(t_i)}\{0\} \right\}^{\delta_i} \\ &\quad \times \left\{ (1 - \rho) q^{(t_i+1)^\beta} \right\}^{1-\delta_i} \end{aligned} \quad (34)$$

onde  $0 < q < 1$ ,  $\beta > 0$  e  $0 \leq \rho \leq 1$  são os parâmetros a serem estimados e  $t_i$  são os tempos observados com seus respectivos indicadores de censura  $\delta_i$ ,  $i = 1, 2, \dots, n$ .

### 3.4. Modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ)

De forma análoga a construção dos modelos com fração de cura ou excessos de zeros, o modelo Weibull com fração de cura e excessos de zeros pode ser formulado através de uma mistura de três variáveis aleatórias, considerando que os indivíduos podem fazer parte de três populações distintas (Figura 9).

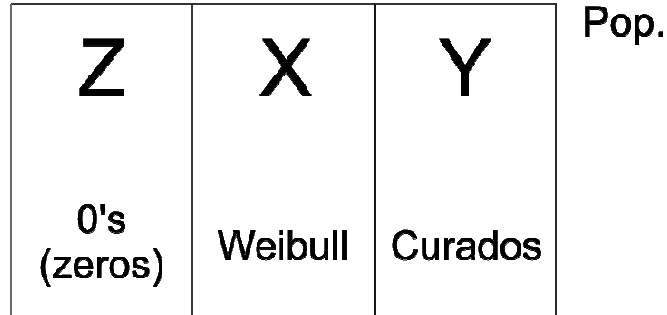


Figura 9 - Representação do modelo por meio de uma mistura de três distribuições.

Na Figura 9,  $X$  é uma variável aleatória com distribuição Weibull discreta e representa o tempo de sobrevivência dos indivíduos não curados;  $Y$  é uma variável degenerada no infinito, que representa o tempo de vida dos indivíduos curados (isto é,  $P(Y > y) = 1$ ) e;  $Z$  é uma variável degenerada no tempo zero, isto é,  $P(Z = 0) = 1$ .

A distribuição da variável  $T$ , resultado da mistura das três variáveis,  $X$ ,  $Y$  e  $Z$ , descritas acima, é dada por:

$$P_T(t) = (1 - \phi - \rho)P(X = t) + \phi P[Y = t] + \rho P[Z = t]. \quad (35)$$

Como  $P(Y = t) = 0 \forall t$  e  $P[Z = t] = I_{\{t\}}\{0\}$ , então a distribuição de probabilidades do modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) é dada por:

$$\begin{aligned} P_{WDFZ}(t) &= (1 - \phi - \rho)P(X = t) + \phi P[Y = t] + \rho P[Z = t] \\ &= (1 - \phi - \rho)P_{WD}(t) + \rho I_{\{t\}}(0) \\ &= (1 - \phi - \rho) \left[ q^{t\beta} - q^{(t+1)\beta} \right] + \rho I_{\{t\}}(0), t = 0, 1, 2 \dots \end{aligned} \quad (36)$$

em que  $I_{\{t\}}\{0\}$  é uma função indicadora, que vale um quando  $t = 0$  e zero quando  $t > 0$  e  $0 \leq \phi \leq 1$ ,  $0 \leq \rho \leq 1$  (restrito à  $\phi + \rho \leq 1$ ) são os parâmetros de mistura representando, respectivamente, a proporção de curados e a proporção de excessos de zeros na população.



Note que se  $\phi = 0$ , o modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) se reduz ao modelo Weibull discreto com excessos de zeros dado pela equação (31). Se  $\rho = 0$ , o modelo (36) se reduz ao modelo Weibull discreto com fração de cura (WDF) dado pela equação (28). Também, se  $\phi = 0$  e  $\rho = 0$ , o modelo (36) se reduz ao modelo Weibull discreto (WD) dado pela equação (23).

A função de sobrevivência do modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) é representada por:

$$\begin{aligned}
S_{WDFZ}(t) &= \sum_{u=t+1}^{\infty} P_{WDFZ}(u) \\
&= \phi I_{(t)}\{\infty\} + \sum_{u=t+1}^{\infty} [(1 - \phi - \rho)P_{WD}(t) + \rho I_{\{t\}}(0)] \\
&= \phi + \sum_{u=t+1}^{\infty} (1 - \phi - \rho)P_{WD}(u) \\
&= \phi + (1 - \phi - \rho)S_{WD}(t) \\
&= \phi + (1 - \phi - \rho)q^{(t+1)^\beta}, \quad t = 0, 1, 2, \dots
\end{aligned} \tag{37}$$

Que resulta na seguinte função de risco:

$$\begin{aligned}
h_{WDFZ}(t) &= \frac{P_{WDFZ}(t)}{S_{WDFZ}(t) + P_{WDFZ}(t)} \\
&= \frac{(1 - \phi - \rho) [q^{t^\beta} - q^{(t+1)^\beta}] + \rho I_{\{t\}}(0)}{\phi + (1 - \phi - \rho)q^{t^\beta} + \rho I_{\{t\}}(0)}, \quad t = 0, 1, 2, \dots
\end{aligned} \tag{38}$$

Sendo  $\theta = (q, \beta, \phi, \rho)$  o vetor de parâmetros e  $\mathbf{T} = (t_1, \dots, t_n)$  o vetor de valores observados com seus respectivos indicadores de censuras  $\delta_i = (\delta_1, \dots, \delta_n)$ , a função de verossimilhança para o modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) é definida por:

$$L_{WDF}(q, \beta, \phi, \rho | \mathbf{T}, \delta) = \prod_{i=1}^n [P_{WDFZ}(t_i)]^{\delta_i} [S_{WDFZ}(t_i)]^{1-\delta_i}$$

$$\begin{aligned}
&= \prod_{i=1}^n \left\{ (1 - \phi - \rho) \left[ q^{t_i^\beta} - q^{(t_i+1)^\beta} \right] + \rho I_{(t_i)} \{0\} \right\}^{\delta_i} \\
&\quad \times \left\{ \phi + (1 - \phi - \rho) q^{(t_i+1)^\beta} \right\}^{1 - \delta_i}
\end{aligned} \tag{39}$$

onde  $0 < q < 1$ ,  $\beta > 0$ ,  $0 \leq \phi \leq 1$  e  $0 \leq \rho \leq 1$  (restrito à  $\phi + \rho \leq 1$ ) são os parâmetros a serem estimados e  $t_i$  são os tempos observados com seus respectivos indicadores de censura  $\delta_i$ ,  $i = 1, 2, \dots, n$ .

As Figuras 10, 11 e 12 apresentam, respectivamente, a distribuição de probabilidades, sobrevivência e a função de risco do modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ).

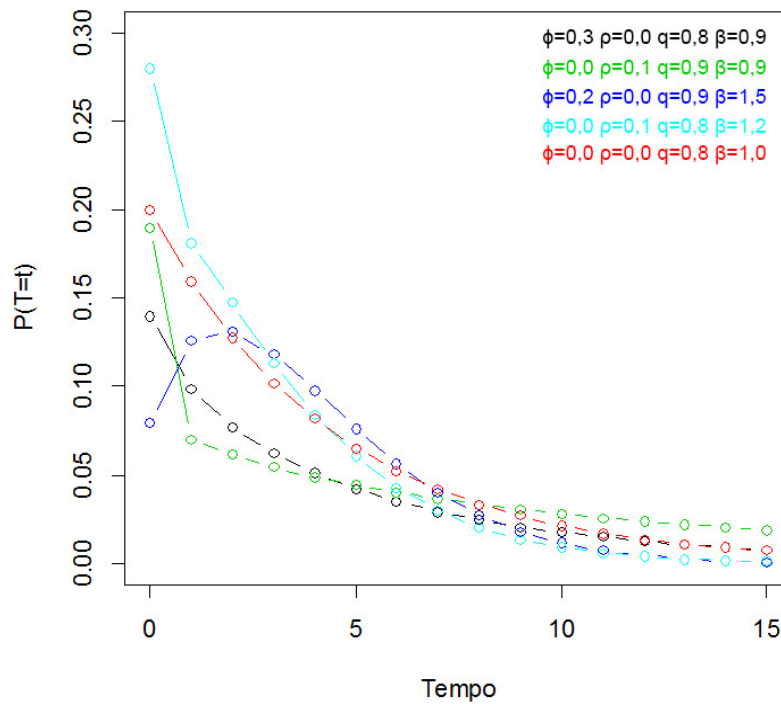


Figura 10 - Distribuição de probabilidades da Weibull discreta com fração de cura e excesso de zeros para diversos valores dos parâmetros  $\phi$ ,  $\rho$ ,  $q$  e  $\beta$

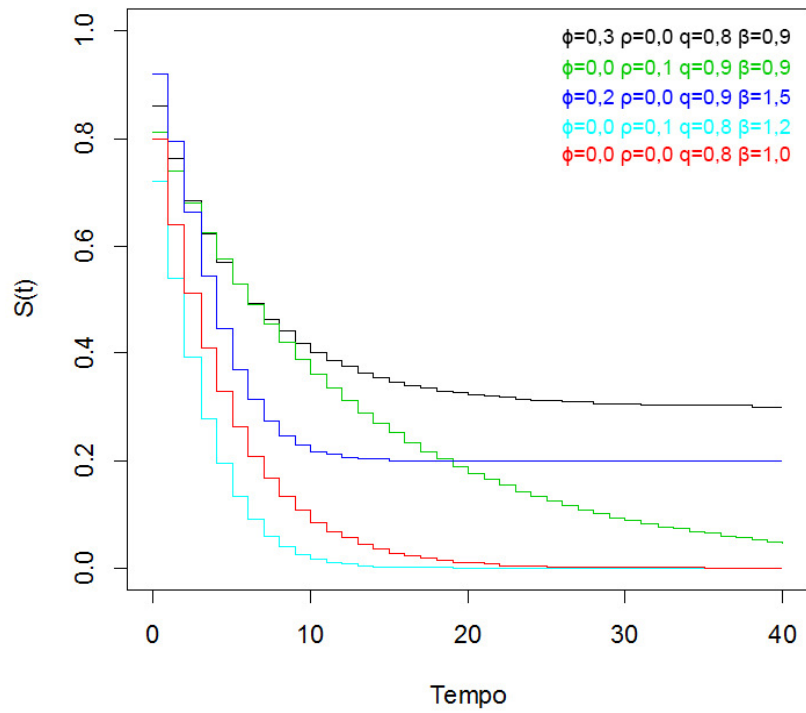


Figura 11 - Função de sobrevivência da Weibull discreta com fração de cura e excesso de zeros para diversos valores dos parâmetros  $\phi$ ,  $\rho$ ,  $q$  e  $\beta$

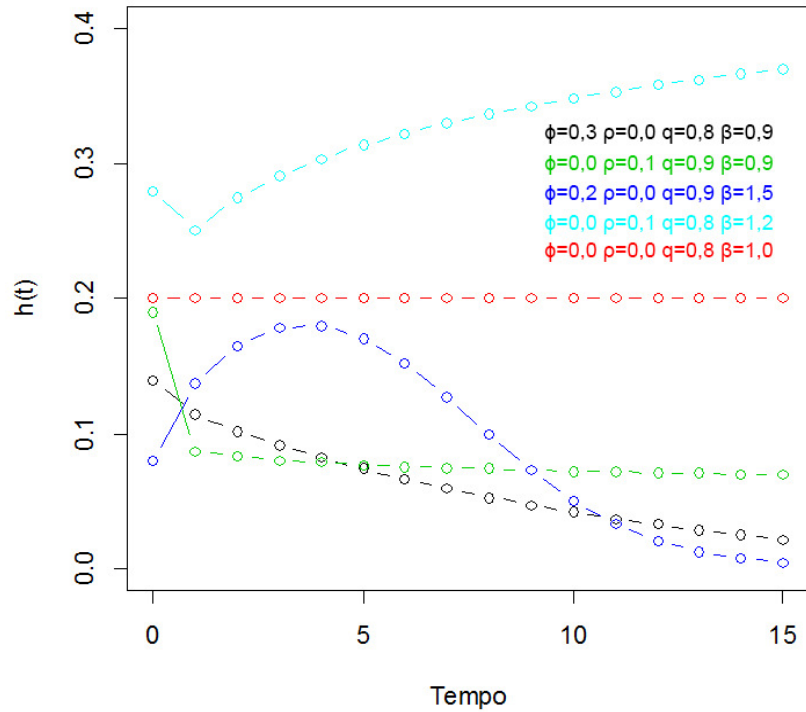


Figura 12 Função de risco da Weibull discreta com fração de cura e excesso de zeros para diversos valores dos parâmetros  $\phi$ ,  $\rho$ ,  $q$  e  $\beta$

Note que a distribuição de probabilidades e, conseqüentemente, a função de sobrevivência do modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) apresentada por (36) e (37) é diferente de uma distribuição de probabilidades do modelo Weibull discreto com fração de cura aplicado a casos com excesso de zeros. As equações (40) e (41) apresentam, respectivamente, a função de sobrevivência e a distribuição de probabilidades para o modelo Weibull discreto com fração de cura aplicado a casos com excesso de zeros. Esse modelo considera que a população é dividida em duas partes: curados e não curados, sendo que os indivíduos não curados são divididos em indivíduos que falharam no tempo zeros e não-zeros.

$$\begin{aligned}
S_{WDF(Z)}(t) &= \phi + (1 - \phi)S_{WDZ}(t) \\
&= \phi + (1 - \phi)(1 - \rho)S_{WD}(t) \\
&= \phi + (1 - \phi)(1 - \rho)q^{(t+1)^\beta}, t=0, 1, 2, \dots
\end{aligned} \tag{40}$$

$$\begin{aligned}
P_{WDF(Z)}(t) &= (1 - \phi)P_{WDZ}(t) \\
&= (1 - \phi)[(1 - \rho)P_{WD}(t) + \rho I_{(t)}\{0\}] \\
&= (1 - \phi)[(1 - \rho)(q^{t^\beta} - q^{(t+1)^\beta}) + \rho I_{(t)}\{0\}], t=0, 1, 2, \dots
\end{aligned} \tag{41}$$

Ainda, as funções (36) e (37) do modelo proposto neste trabalho também são diferentes da distribuição de probabilidades e da função de sobrevivência do modelo Weibull discreto com excesso de zeros aplicado a casos com fração de cura. As equações (42) e (43) apresentam, respectivamente, a função de sobrevivência e a distribuição de probabilidades para o modelo Weibull discreto com excesso de zeros aplicado a casos com fração de cura. Esse modelo considera que a população é dividida em duas partes: indivíduos que falharam no tempo zeros e não-zeros, sendo que os indivíduos que não falharam no tempo zero são divididos em curados e não curados.

$$\begin{aligned}
S_{WDZ(F)}(t) &= (1 - \rho)S_{WDF}(t) \\
&= (1 - \rho)[\phi + (1 - \phi)S_{WD}(t)] \\
&= (1 - \rho)[\phi + (1 - \phi)q^{(t+1)^\beta}], t=0, 1, 2, \dots
\end{aligned} \tag{42}$$

$$\begin{aligned}
P_{WDZ(F)}(t) &= (1 - \rho)P_{WDF}(t) + \rho I_{(t)}\{0\} \\
&= (1 - \rho)[(1 - \phi)P_{WD}(t)] + \rho I_{(t)}\{0\} \\
&= (1 - \rho)[(1 - \phi)(q^{t^\beta} - q^{(t+1)^\beta})] + \rho I_{(t)}\{0\}, t=0, 1, \dots
\end{aligned} \tag{43}$$

É interessante ressaltar que a diferença entre o modelo central proposto nesse trabalho (WDFZ) e os dois acima (WDF(Z) e WDZ(F)) está basicamente no controle da proporção de indivíduos curados e de zeros. Para os novos modelos, apresentados por (40) e (42), têm-se as restrições  $0 \leq \phi \leq 1$  e  $0 \leq \rho \leq 1$ , enquanto que o modelo WDFZ apresenta a restrição adicional de  $0 \leq \phi + \rho \leq 1$ . No entanto, todos os três modelos, (36), (40) e (42) se reduzem ao modelo Weibull discreto com excessos de zeros (31) se  $\phi = 0$ , ao modelo Weibull discreto com fração de cura (28) se  $\rho=0$  e ao modelo Weibull discreto (23) se  $\phi = 0$  e  $\rho=0$ .

#### 4. Dados Simulados

Neste capítulo são descritas as simulações computacionais realizadas via *software* R (R CORE TEAM, 2015) juntamente com os resultados numéricos obtidos. As simulações têm por objetivo gerar dados de sobrevivência com fração de curados e excessos de zeros a fim de testar o ajuste de modelos derivados da distribuição Weibull Discreta.

Foram calculadas as estimativas para os modelos WDFZ, WDZ, WDF, WD e GEO para diferentes cenários (combinações distintas dos parâmetros), como descrito na Tabela 2. Para o cálculo das estimativas dos parâmetros foi utilizada uma amostra de tamanho  $n=200$  com percentual de censura (de indivíduos não curados) igual a 10%. Através do método de inversão da função de distribuição acumulada, foram geradas amostras a partir de uma distribuição Weibull discreta com fração de cura e excessos de zeros. Um modelo multinomial foi considerado para gerar a mistura das distribuições e as censuras foram geradas, independentemente do tempo de sobrevivência, através de uma distribuição Bernoulli com probabilidade de sucesso igual à proporção de indivíduos não censurados. A programação utilizada para a geração dos dados e obtenção das estimativas dos parâmetros dos modelos se encontra no Apêndice deste trabalho.

O vício e do EQM dos estimadores dos parâmetros do modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) foram calculados a partir de 1000 réplicas de Monte Carlo considerando amostras de tamanho  $n = 50, 100, 200, 400$  e  $800$  e diferentes percentuais de censura apenas para a distribuição Weibull Discreta com fração de cura e excessos de zeros.

Tabela 2 - Cenários utilizados na simulação

Cenário	$\phi$	$\rho$	$q$	$\beta$
1	0,05	0,05	0,9	0,9
2	0,35	0,3	0,9	0,9
3	0,15	0,5	0,98	1,5
4	0,15	0,2	0,95	1,9

##### 4.1. Cenário 1

Para o Cenário 1, descrito pela Tabela 2, com o percentual de censura igual a 10% e uma amostra de tamanho 200, foram estimados os parâmetros dos modelos WDFZ, WDZ, WDF, WD e GEO; apresentados na Tabela 3.

Tabela 3 - Estimativas dos parâmetros dos modelos para o Cenário 1,  $n=200$  e 10% de censura.

Parâmetro	Modelo				
	WDFZ	WDZ	WDF	WD	GEO
$\phi$	0,06077	-	0,03381	-	-
$\rho$	0,10294	0,07229	-	-	-
$q$	0,9496	0,9217	0,88366	0,8791	0,94901
$\beta$	1,044	0,83886	0,76797	0,72138	-

A Tabela 4 apresenta os valores obtidos pelo Critério de informação de Akaike – AIC (AKAIKE, 1974) e o Critério de informação Bayesiano – BIC (SCHWARS, 1978) para a simulação do Cenário 1. Segundo o AIC, o melhor modelo seria o WDFZ (menor AIC) e segundo o critério de informação BIC, o melhor modelo seria o WD (menor BIC). O critério BIC penaliza mais aqueles modelos com muitos parâmetros. Provavelmente esse foi o motivo do modelo WD ter apresentado menor BIC. Esse resultado é esperado, visto o pequeno percentual de curados e excessos de zeros considerados no Cenário 1.

Tabela 4 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 3.

Medidas	Modelos				
	WDFZ	WDZ	WDF	WD	GEO
AIC	1313,484	1317,701	1320,148	1319,095	1345,065
BIC	1326,677	1327,596	1330,043	1325,692	1348,363

A Figura 13 apresenta o gráfico de Kaplan-Meier para os dados simulados de acordo com o Cenário 1 e o ajuste de cada modelo analisado. É possível verificar que o modelo WDFZ aparentemente se ajusta melhor a curva de Kaplan-Meier.

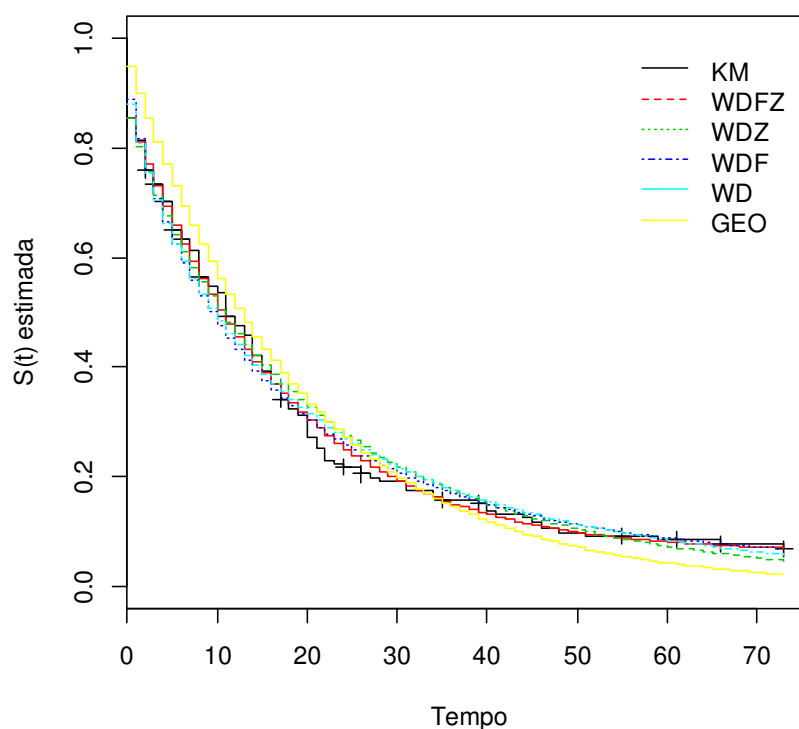


Figura 13 - Curvas de sobrevivência estimadas considerando o Cenário 1

A Tabela 5 apresenta o Vício e o Erro quadrático médio (EQM) para cada parâmetro dentro do Cenário 1 proposto. Para isso foram utilizadas apenas as estimativas do modelo

WDFZ replicadas 1000 vezes para diferentes tamanhos de amostra e porcentagens de censuras.

Tabela 5 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 1.

<i>n</i>	Parâmetro	Cenário 1							
		$\phi = 0,05$		$\rho = 0,05$		$q = 0,9$		$\beta = 0,9$	
		Censuras*							
		0%		10%		20%			
		Vício	EQM	Vício	EQM	Vício	EQM		
50	$\phi$	-0,02280	0,00052	-0,01080	0,00012	-0,00284	0,00001		
	$\rho$	-0,00140	0,00000	-0,00539	0,00003	-0,00486	0,00002		
	$q$	-0,00714	0,00005	-0,00507	0,00003	0,00015	0,00000		
	$\beta$	-0,01783	0,00032	-0,01881	0,00035	-0,01286	0,00017		
100	$\phi$	-0,01878	0,00035	-0,00938	0,00009	-0,00155	0,00000		
	$\rho$	-0,00426	0,00002	-0,00595	0,00004	-0,00484	0,00002		
	$q$	-0,00762	0,00006	-0,00422	0,00002	0,00123	0,00000		
	$\beta$	-0,02252	0,00051	-0,01933	0,00037	-0,01197	0,00014		
200	$\phi$	-0,01513	0,00023	-0,00725	0,00005	0,00058	0,00000		
	$\rho$	-0,00599	0,00004	-0,00587	0,00003	-0,00477	0,00002		
	$q$	-0,00800	0,00006	-0,00329	0,00001	0,00216	0,00000		
	$\beta$	-0,02489	0,00062	-0,01876	0,00035	-0,01131	0,00013		
400	$\phi$	-0,01214	0,00015	-0,00502	0,00003	0,00306	0,00001		
	$\rho$	-0,00599	0,00004	-0,00573	0,00003	-0,00450	0,00002		
	$q$	-0,00747	0,00006	-0,00236	0,00001	0,00315	0,00001		
	$\beta$	-0,02338	0,00055	-0,01728	0,00030	-0,00962	0,00009		
800	$\phi$	-0,01005	0,00010	-0,00297	0,00001	0,00547	0,00003		
	$\rho$	-0,00572	0,00003	-0,00527	0,00003	-0,00409	0,00002		
	$q$	-0,00689	0,00005	-0,00141	0,00000	0,00417	0,00002		
	$\beta$	-0,02168	0,00047	-0,01536	0,00024	-0,00764	0,00006		

Nota: Vício e EQM calculados a partir de  $m=1000$  replicações de Monte Carlo.

\* Percentual de censuras nas observações não curadas.

De modo geral, os resultados obtidos para o Cenário 1, que apresenta pequeno percentual de censuras e zeros, resultaram em valores referentes ao vício dos estimadores e EQM pequenos, mesmo para amostras de tamanho razoável ( $n=50$ ). Esse fato confirma a acurácia do modelo proposto.



## 4.2. Cenário 2

Para o Cenário 2, com o percentual de censura igual a 10% e uma amostra de tamanho 200, foram estimados os parâmetros dos modelos WDFZ, WDZ, WDF, WD e GEO cujos valores são apresentados na Tabela 6.

Tabela 6 - Estimativas dos parâmetros dos modelos para o Cenário 2,  $n=200$  e 10% de censura.

Parâmetro	Modelo				
	WDFZ	WDZ	WDF	WD	GEO
$\phi$	0,28763	0,10882	-	-	-
$\rho$	0,33971	-	0,15881	-	-
$q$	0,87299	0,74615	0,60861	0,66775	0,97645
$\beta$	0,76436	0,27736	0,25847	0,22698	-

A Tabela 7 apresenta as medidas de seleção AIC e BIC. Segundo o critério de informação AIC, o melhor modelo seria o WDFZ. Segundo o critério de informação BIC, o melhor modelo seria o WDZ. O critério BIC penaliza mais aqueles modelos com muitos parâmetros. Provavelmente esse foi o motivo do modelo WD ter apresentado menor BIC.

Tabela 7 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 6.

Medidas	Modelos				
	WDFZ	WDZ	WDF	WD	GEO
AIC	871,8066	875,0793	875,1492	873,3097	1205,195
BIC	884,9999	884,9743	885,0441	879,9064	1208,494

A Figura 14 apresenta o gráfico de Kaplan-Meier para os dados simulados de acordo com o Cenário 2 e o ajuste de cada modelo analisado. Para uma melhor visualização do ajuste dos modelos propostos, também foi analisada a Figura 15 que é semelhante à Figura 14, porém com o eixo y reduzido. Através de uma análise visual, é possível notar que o modelo WDFZ possui o melhor ajuste.

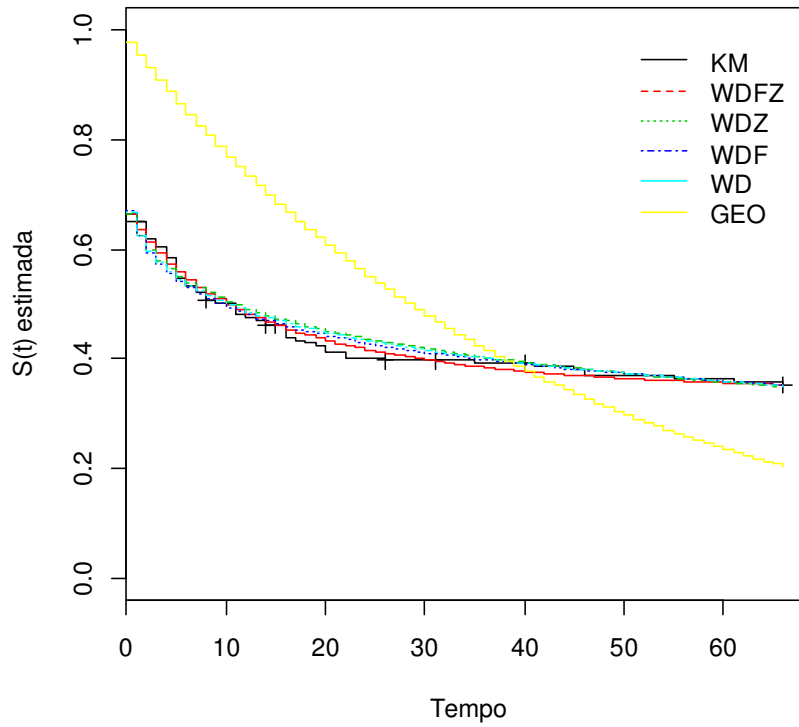


Figura 14 - Curvas de sobrevivência estimadas considerando o Cenário 2

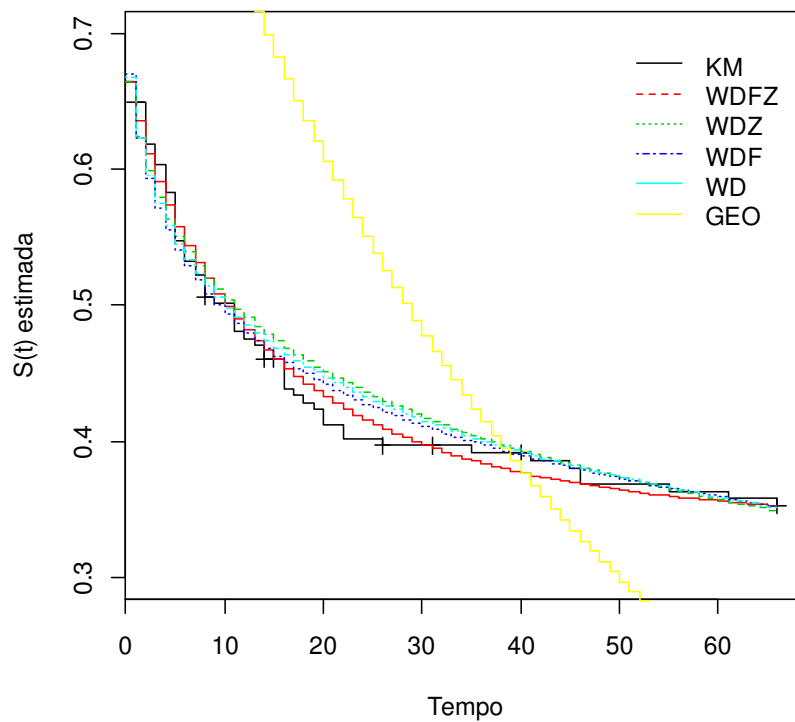


Figura 15 - Curvas de sobrevivência estimadas considerando o Cenário 2 (eixos reduzidos)

A Tabela 8 apresenta o Vício e o Erro quadrático médio (EQM) para cada parâmetro dentro do Cenário 2. Para isso foram utilizadas apenas as estimativas do modelo WDFZ replicadas 1000 vezes para diferentes tamanhos de amostra e porcentagens de censuras.

Veja que em situações como no Cenário 2, que considera altos percentuais de curados e zeros, os estimadores apresentam vício maior do que os percebidos em situações com baixo percentual de curados e zeros (veja Tabela 5). Como esperado, esse vício é maior quando a amostra é pequena e diminui à medida que o tamanho da amostra aumenta. Além disso, os baixos valores encontrados para o EQM sinalizam a acurácia do modelo WDFZ.

Tabela 8 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 2.

		Cenário 2							
		$\phi = 0,35$		$\rho = 0,3$		$q = 0,9$		$\beta = 0,9$	
		Censuras*							
n	Parâmetro	0%		10%		20%			
		Vício	EQM	Vício	EQM	Vício	EQM	Vício	EQM
50	$\phi$	-0,13422	0,01801	-0,05584	0,00312	-0,03704	0,00137		
	$\rho$	-0,05525	0,00305	-0,02847	0,00081	-0,02467	0,00061		
	$q$	-0,04789	0,00229	-0,03107	0,00097	-0,02637	0,00070		
	$\beta$	-0,15794	0,02494	-0,09838	0,00968	-0,08375	0,00701		
100	$\phi$	-0,09592	0,00920	-0,05271	0,00278	-0,03481	0,00121		
	$\rho$	-0,04495	0,00202	-0,02912	0,00085	-0,02512	0,00063		
	$q$	-0,04641	0,00215	-0,03169	0,00100	-0,02681	0,00072		
	$\beta$	-0,14733	0,02171	-0,10077	0,01016	-0,08646	0,00748		
200	$\phi$	-0,06813	0,00464	-0,04473	0,00200	-0,02956	0,00087		
	$\rho$	-0,03520	0,00124	-0,02746	0,00075	-0,02421	0,00059		
	$q$	-0,03919	0,00154	-0,03094	0,00096	-0,02631	0,00069		
	$\beta$	-0,12083	0,01460	-0,09787	0,00958	-0,08516	0,00725		
400	$\phi$	-0,05250	0,00276	-0,03746	0,00140	-0,02403	0,00058		
	$\rho$	-0,02811	0,00079	-0,02523	0,00064	-0,02280	0,00052		
	$q$	-0,03316	0,00110	-0,02891	0,00084	-0,02478	0,00061		
	$\beta$	-0,10217	0,01044	-0,09169	0,00841	-0,08053	0,00649		
800	$\phi$	-0,04252	0,00181	-0,03135	0,00098	-0,01899	0,00036		
	$\rho$	-0,02310	0,00053	-0,02301	0,00053	-0,02146	0,00046		
	$q$	-0,02812	0,00079	-0,02649	0,00070	-0,02304	0,00053		
	$\beta$	-0,08658	0,00750	-0,08441	0,00712	-0,07547	0,00570		

Nota: Vício e EQM calculados a partir de  $m=1000$  replicações de Monte Carlo.

\* Percentual de censuras nas observações não curadas.

### 4.3. Cenário 3

Para o Cenário 3, com o percentual de censura igual a 10% e uma amostra de tamanho 200, foram estimados os parâmetros dos modelos WDFZ, WDZ, WDF, WD e GEO cujos valores são apresentados na Tabela 9.

Tabela 9 - Estimativas dos parâmetros dos modelos para o Cenário 3,  $n=200$  e 10% de censura.

Parâmetro	Modelo				
	WDFZ	WDZ	WDF	WD	GEO
$\phi$	0,14942	-	0,00000	-	-
$\rho$	0,48573	0,46741	-	-	-
$q$	0,97459	0,94831	0,53940	0,52966	0,92172
$\beta$	1,35505	0,87691	0,27654	0,27248	-

A Tabela 10 apresenta as medidas de seleção AIC e BIC. Segundo os critérios de informação apresentados, o melhor modelo seria o WDFZ. É interessante salientar que apenas do BIC penalizar modelos com muitos parâmetros, nesse cenário tanto o AIC como o BIC escolheram o modelo mais completo, WDFZ.

Tabela 10 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 9.

Medidas	Modelos				
	WDFZ	WDZ	WDF	WD	GEO
AIC	843,9043	848,6033	871,2252	869,1488	1152,3740
BIC	857,0976	858,4982	881,1202	875,7455	1155,6720

A Figura 16 apresenta o gráfico de Kaplan-Meier para os dados simulados de acordo com o Cenário 3 e o ajuste de cada modelo analisado. Para uma melhor visualização do ajuste dos modelos propostos, a Figura 17 é semelhante à Figura 16, porém com os eixos reduzidos. Através de uma análise visual, é possível notar que o modelo WDFZ possui o melhor ajuste.

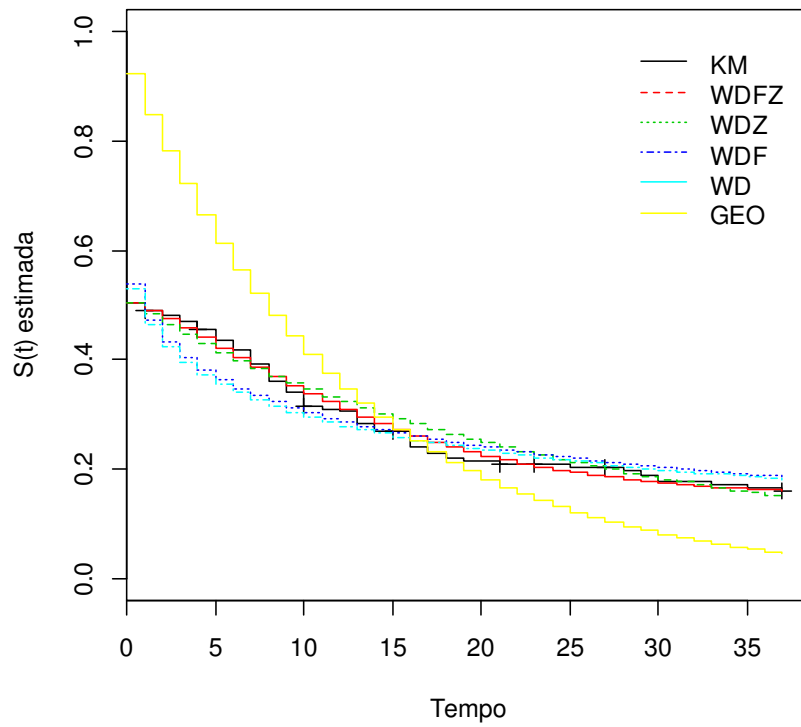


Figura 16 - Curvas de sobrevivência estimadas considerando o Cenário 3

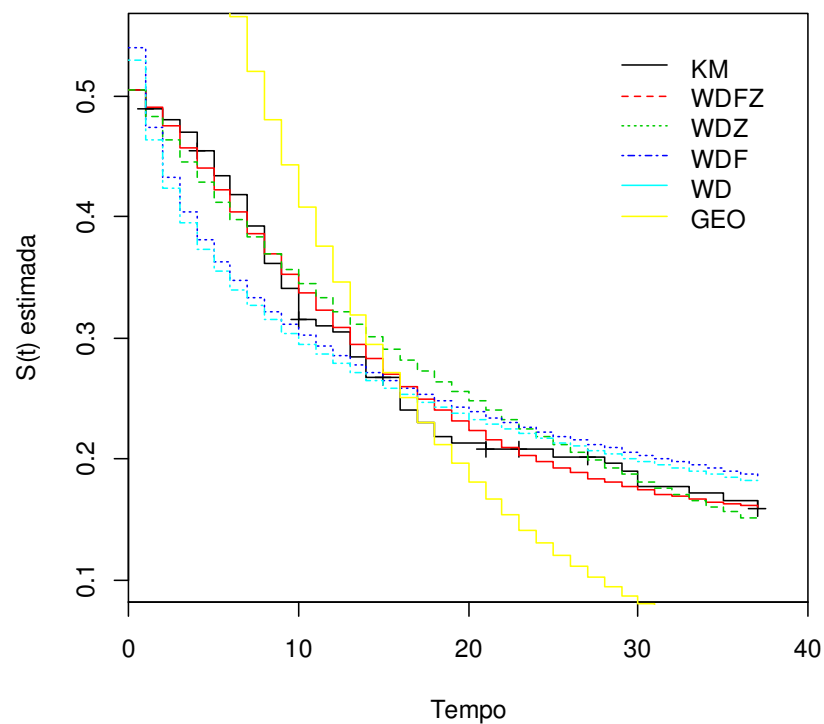


Figura 17 - Curvas de sobrevivência estimadas considerando o Cenário 3 (eixos reduzidos)

A Tabela 11 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 3. apresenta o Vício e o Erro quadrático médio (EQM) para

cada parâmetro dentro do Cenário 3. Para isso foram utilizadas apenas as estimativas do modelo WDFZ replicadas 1000 vezes para cada tamanho de amostra e porcentagens de censuras.

Veja que em situações, como no Cenário 3, que consideram altos percentuais de excessos de zeros, os estimadores apresentam vício maior do que os percebidos em situações com baixo percentual de excessos de zeros (veja Tabela 5). Como esperado, esse vício é maior quando a amostra é pequena e diminui à medida que o tamanho da amostra aumenta. É importante ressaltar que os valores dos EQM encontrados para o cenário 3 estão próximos de zero, o que confirma a acurácia do modelo.

Tabela 11 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 3.

		<b>Cenário 3</b>							
		$\phi = 0,15$		$\rho = 0,5$		$q = 0,98$		$\beta = 1,5$	
		<b>Censuras*</b>							
<i>N</i>	<b>Parâmetro</b>	<b>0%</b>		<b>10%</b>		<b>20%</b>			
		<b>Vício</b>	<b>EQM</b>	<b>Vício</b>	<b>EQM</b>	<b>Vício</b>	<b>EQM</b>	<b>Vício</b>	<b>EQM</b>
50	$\phi$	-0,06306	0,00398	-0,02569	0,00066	-0,01449	0,00021		
	$\rho$	-0,01058	0,00011	-0,00306	0,00001	-0,00294	0,00001		
	$q$	-0,01461	0,00021	-0,00763	0,00006	-0,00612	0,00004		
	$\beta$	-0,11532	0,01330	-0,07499	0,00562	-0,06224	0,00387		
100	$\phi$	-0,04531	0,00205	-0,02365	0,00056	-0,01283	0,00016		
	$\rho$	-0,00725	0,00005	-0,00349	0,00001	-0,00319	0,00001		
	$q$	-0,01198	0,00014	-0,00765	0,00006	-0,00628	0,00004		
	$\beta$	-0,11284	0,01273	-0,07696	0,00592	-0,06465	0,00418		
200	$\phi$	-0,03340	0,00112	-0,01953	0,00038	-0,00963	0,00009		
	$\rho$	-0,00482	0,00002	-0,00319	0,00001	-0,00327	0,00001		
	$q$	-0,00986	0,00010	-0,00722	0,00005	-0,00599	0,00004		
	$\beta$	-0,09686	0,00938	-0,07443	0,00554	-0,06296	0,00396		
400	$\phi$	-0,02576	0,00066	-0,01553	0,00024	-0,00622	0,00004		
	$\rho$	-0,00394	0,00002	-0,00307	0,00001	-0,00311	0,00001		
	$q$	-0,00812	0,00007	-0,00660	0,00004	-0,00554	0,00003		
	$\beta$	-0,08289	0,00687	-0,06962	0,00485	-0,05880	0,00346		
800	$\phi$	-0,02088	0,00044	-0,01221	0,00015	-0,00311	0,00001		
	$\rho$	-0,00326	0,00001	-0,00276	0,00001	-0,00295	0,00001		
	$q$	-0,00684	0,00005	-0,00595	0,00004	-0,00507	0,00003		
	$\beta$	-0,07132	0,00509	-0,06342	0,00402	-0,05408	0,00292		

Nota: Vício e EQM calculados a partir de m=1000 replicações de Monte Carlo.

\* Percentual de censuras nas observações não curadas.

## 4.4. Cenário 4

Para o Cenário 4, com o percentual de censura igual a 10% e uma amostra de tamanho 200, foram estimados os parâmetros dos modelos WDFZ, WDZ, WDF, WD e GEO; cujos valores são apresentados na Tabela 12.

Tabela 12 - Estimativas dos parâmetros dos modelos para o Cenário 4,  $n=200$  e 10% de censura.

Parâmetro	Modelo				
	WDFZ	WDZ	WDF	WD	GEO
$\phi$	0,21731	-	0,00000	-	-
$\rho$	0,22109	0,16033	-	-	-
$q$	0,94843	0,89321	0,77053	0,77060	0,85839
$\beta$	1,76262	1,03552	0,71583	0,71585	-

A Tabela 13 apresenta as medidas de seleção AIC e BIC. Segundo os critérios de informação apresentados, o melhor modelo seria o WDFZ. É interessante salientar que apenas do BIC penalizar modelos com muitos parâmetros, nesse cenário tanto o AIC como o BIC escolheram o modelo mais completo, WDFZ.

Tabela 13 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 12.

Medidas	Modelos				
	WDFZ	WDZ	WDF	WD	GEO
AIC	819,8093	825,7322	830,348	828,348	843,0393
BIC	833,0026	835,6271	840,243	834,9447	846,3377

A Figura 18 apresenta o gráfico de Kaplan-Meier para os dados simulados de acordo com o Cenário 4 e o ajuste de cada modelo analisado. Para uma melhor visualização do ajuste dos modelos propostos, a Figura 18 é semelhante à Figura 19, porém com os eixos reduzidos.

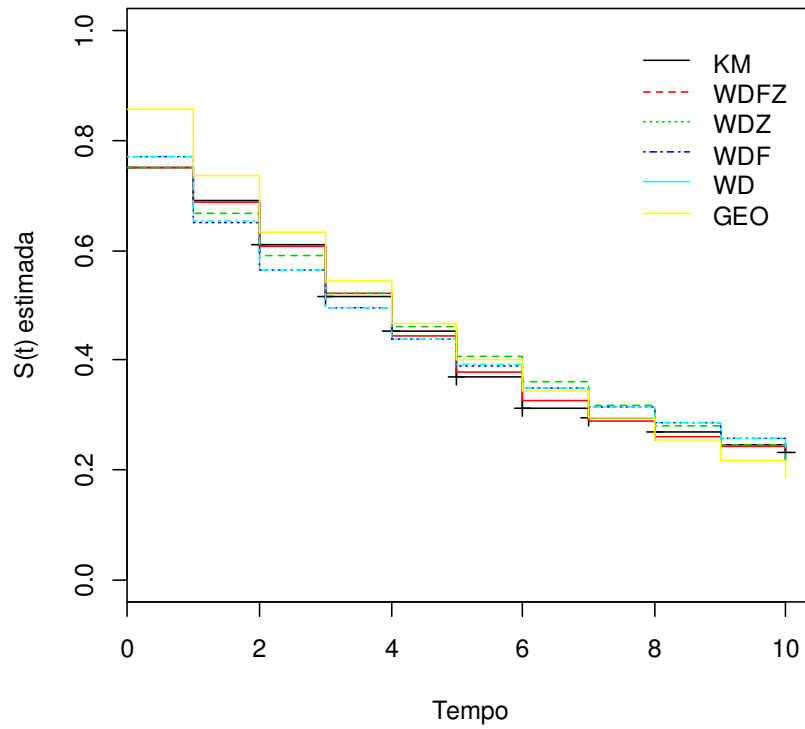


Figura 18 - Curvas de sobrevivência estimadas considerando o Cenário 4

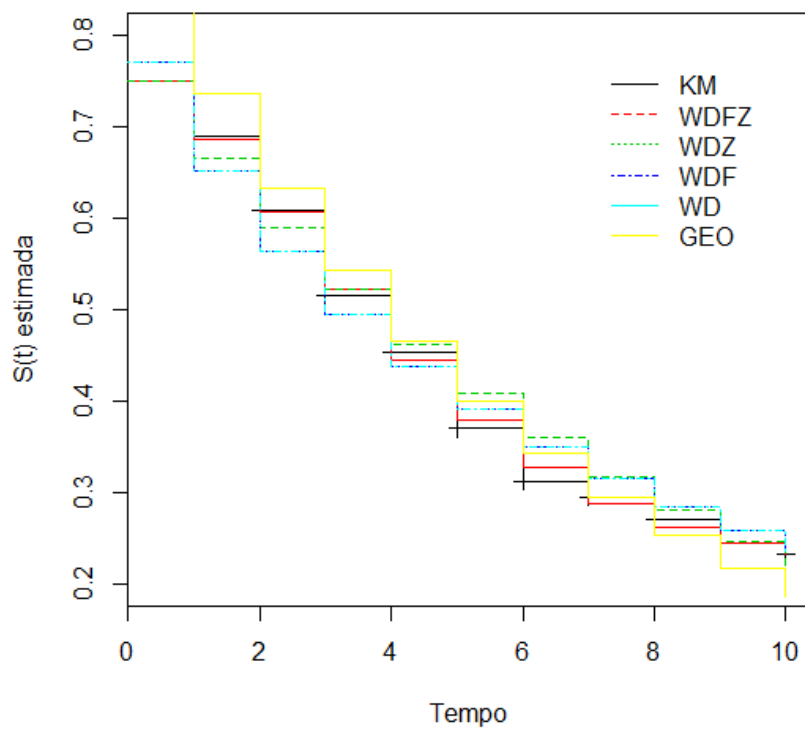


Figura 19- Curvas de sobrevivência estimadas considerando o Cenário 4 (eixos reduzidos)

A Tabela 14 apresenta o Vício e o Erro quadrático médio (EQM) para cada parâmetro dentro do Cenário 4. Para isso foram utilizadas apenas as estimativas do modelo



WDFZ replicadas 1000 vezes para cada tamanho de amostra e porcentagens de censuras. O valores encontrados para o vício e para o EQM foram próximos de zero, independentemente do tamanho da amostra e da porcentagem de censura. Esse fato confirma a acurácia do modelo proposto, WDFZ.

Tabela 14 - Vício e Erro Quadrático Médio (EQM) para diversos tamanhos de amostras e percentuais de censura do Cenário 4.

		<b>Cenário 4</b>							
		$\phi = 0,15$		$\rho = 0,2$		$q = 0,95$		$\beta = 1,9$	
		<b>Censuras*</b>							
<i>n</i>	Parâmetro	0%		10%		20%			
		Vício	EQM	Vício	EQM	Vício	EQM	Vício	EQM
50	$\phi$	-0,03196	0,00102	-0,01067	0,00011	0,00407	0,00002		
	$\rho$	-0,02200	0,00048	-0,01036	0,00011	-0,00788	0,00006		
	$q$	-0,02304	0,00053	-0,01028	0,00011	-0,00690	0,00005		
	$\beta$	-0,13617	0,01854	-0,06885	0,00474	-0,05009	0,00251		
100	$\phi$	-0,02278	0,00052	-0,00735	0,00005	0,00753	0,00006		
	$\rho$	-0,01583	0,00025	-0,01005	0,00010	-0,00775	0,00006		
	$q$	-0,01728	0,00030	-0,00989	0,00010	-0,00651	0,00004		
	$\beta$	-0,10809	0,01168	-0,06780	0,00460	-0,04861	0,00236		
200	$\phi$	-0,01695	0,00029	-0,00342	0,00001	0,01148	0,00013		
	$\rho$	-0,01234	0,00015	-0,00894	0,00008	-0,00734	0,00005		
	$q$	-0,01359	0,00018	-0,00880	0,00008	-0,00578	0,00003		
	$\beta$	-0,08762	0,00768	-0,06185	0,00383	-0,04419	0,00195		
400	$\phi$	-0,01319	0,00017	0,00006	0,00000	0,01504	0,00023		
	$\rho$	-0,00969	0,00009	-0,00823	0,00007	-0,00679	0,00005		
	$q$	-0,01115	0,00012	-0,00777	0,00006	-0,00501	0,00003		
	$\beta$	-0,07392	0,00546	-0,05645	0,00319	-0,03960	0,00157		
800	$\phi$	-0,01079	0,00012	0,00294	0,00001	0,01828	0,00033		
	$\rho$	-0,00802	0,00006	-0,00734	0,00005	-0,00635	0,00004		
	$q$	-0,00933	0,00009	-0,00672	0,00005	-0,00422	0,00002		
	$\beta$	-0,06310	0,00398	-0,04971	0,00247	-0,03476	0,00121		

Nota: Vício e EQM calculados a partir de  $m=1000$  replicações de Monte Carlo.

\* Percentual de censuras nas observações não curadas.

O Cenário 4 considera uma situação em que há um significativo percentual de curados e excessos de zeros. Os estimadores apresentam vício maior do que os percebidos em situações com baixo percentual de excessos de zeros (veja Tabela 5). Como nos Cenários 2 e 3, esse vício é maior quando a amostra é pequena e diminui à medida que aumentamos o tamanho da amostra.

## 5. Estudo de caso

O infarto agudo do miocárdio (IAM), conhecido popularmente como infarto do coração, enfarte ou ataque cardíaco, refere-se à morte de parte do músculo cardíaco (miocárdio), que ocorre de forma rápida (ou aguda) devido à obstrução do fluxo sanguíneo das artérias coronárias. O IAM está associado às altas taxas de mortalidade, chegando a aproximadamente 30% no ambiente pré-hospitalar em decorrência principalmente de arritmias ventriculares. No Brasil, estima-se a ocorrência de 300 mil a 400 mil casos anuais de infarto, e que a cada 5 a 7 casos, ocorra um óbito. (GARCIA, 2013)

Realizada sob anestesia local, a técnica de Intervenção Coronária Percutânea (ICP), angioplastia transluminal coronária, é o tratamento não cirúrgico, porém invasivo, das obstruções das artérias coronárias por meio de cateter balão, com o objetivo de aumentar o fluxo de sangue para o coração. (GARCIA, 2013)

A ICP é realizada com o auxílio de um delicado e sofisticado cateter-balão, disponíveis em diversos diâmetros e comprimentos, que é posicionado no local da lesão obstrutiva. A seguir este balão é expandido (insuflado) "esmagando" assim a placa contra as paredes do vaso, desobstruindo a artéria e permitindo que o fluxo sanguíneo retorne ao normal. Trata-se de um tratamento extremamente seguro e eficiente, onde as complicações graves ocorrem na ordem de 2 a 3%. (GARCIA, 2013)

Para esse estudo de caso foi definido como evento de interesse o tempo até o óbito. O tempo foi medido em dias, mas a análise foi realizada apenas para semanas completas.

Para esse estudo foram observados os casos de todos os pacientes submetidos à angioplastia em um hospital de referência em Brasília no período de janeiro a dezembro de 2011. Foram incluídos no estudo 97 pacientes com idade média de 63 anos, sendo a maioria do gênero masculino (69,4%). O interesse inicial do estudo era avaliar a evolução clínica e sobrevida dos pacientes por um período de 1 ano após a realização do procedimento. Após a observação de muitas censuras, esse tempo foi estendido até junho de 2013 e outros dois hospitais de Brasília foram considerados (GARCIA, 2013). A variável tempo de sobrevida dos pacientes foi medido em semanas (completas). Assim, pacientes que morreram antes de completar uma semana após a ICP tiveram seus tempos definidos como zero. A suposição do modelo WDFZ é que parte desses pacientes vieram a óbito antes de uma semana devido a complicações na ICP.

Os dados foram coletados pelo hospital conforme as ICP foram realizadas e foram obtidos acessando os prontuários após a intervenção. O acompanhamento dos pacientes no estudo foi realizado acessando o prontuário dos mesmos no hospital de intervenção e em dois outros hospitais de Brasília (GARCIA, 2013).

Como limitações do estudo destaca-se o pequeno tamanho amostral (97 pacientes), devido ao pouco tempo de implantação do serviço de hemodinâmica no Hospital; e a grande quantidade de censuras (aproximadamente 90%).

Para a análise de dados foram utilizados o software R, versão 3.2.0 (R CORE TEAM, 2015).

Abaixo serão apresentados gráficos com a função de sobrevivência estimada por Kaplan-Meier. Através desses gráficos é possível observar um grande número de censuras. Como a curva de sobrevivência estimada não tende a zero e se mantém constante por um longo período, é razoável supor a existência de indivíduos curados. Além disso, é possível supor um excesso de zeros (visto que 7 das 10 observações não censuradas são iguais a zero), fato que motivou a construção do principal modelo proposto nesse trabalho.

A Tabela 15 apresenta as estimativas para os parâmetros dos modelos analisados considerando o tempo em semanas completas. É interessante notar que a estimativa para o parâmetro  $\phi$  no modelo WDF está próxima de zero, o que implica que as estimativas dos modelos WDF e WD sejam muito parecidas. Nas Figura 20 e Figura 21 é possível visualizar essa semelhança.

Tabela 15 - Estimativas dos parâmetros dos modelos para os dados do tempo de sobrevivências de pacientes submetidos à ICP.

Parâmetro	Modelo				
	WDFZ	WDZ	WDF	WD	GEO
$\phi$	0,87576	-	0,00002	-	-
$\rho$	0,07208	0,05452	-	-	-
$q$	0,9984	0,9814	0,92863	0,92882	0,99727
$\beta$	2,32193	0,32811	0,13442	0,13489	-

A Figura 20 apresenta o estimador de Kaplan-Meier para os dados analisados considerando o tempo em semanas e o ajuste dos modelos Weibull discreta com fração de cura e excesso de zeros (WDFZ), Weibull discreta com Excesso de zeros (WDZ), Weibull discreta com fração de cura (WDF), Weibull discreta (WD) e Geométrico (GEO).

Considerando as Figura 20 e Figura 21 também é possível verificar que o modelo WDFZ possui o melhor ajuste dentre os modelos analisados. É importante destacar um comportamento em formato de “S” no início da curva estimada pelo modelo WDFZ que acompanha a estimativa de Kaplan-Meier. Isso sugere um bom ajuste do modelo. A função de sobrevivência estimada para os modelos WDF e WD aparentemente coincidem. Isso porque o valor estimado para  $\phi$ , parâmetro que representa a fração de cura no modelo WDF que o difere do modelo WD, é muito próximo de zero (0,00002). Ambos os modelos não se ajustaram tão bem ao conjunto de dados. No entanto, é possível destacar que os modelos WDF, WDZ e WD acompanharam bem o excesso de zeros no começo do estudo. É possível dizer o mesmo para o modelo WDZ. Além disso, claramente o modelo GEO não se ajusta bem aos dados.

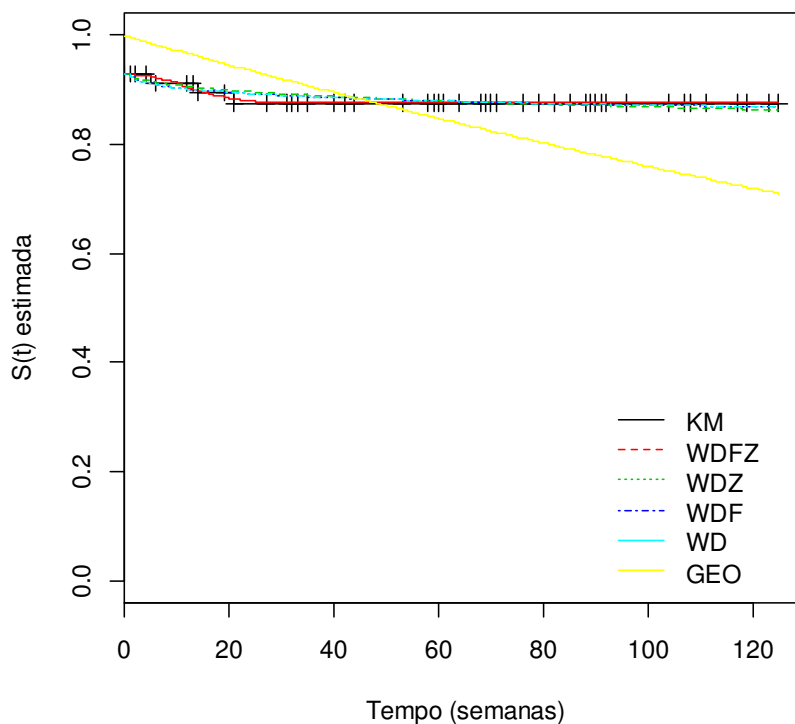


Figura 20- Curvas de sobrevivência estimadas do tempo de sobrevivências (em semanas) de pacientes submetidos à ICP.

A Figura 21 apresenta os mesmos dados apresentadas na Figura 20, porém com a escala do eixo y reduzido para proporcionar uma melhor visualização.

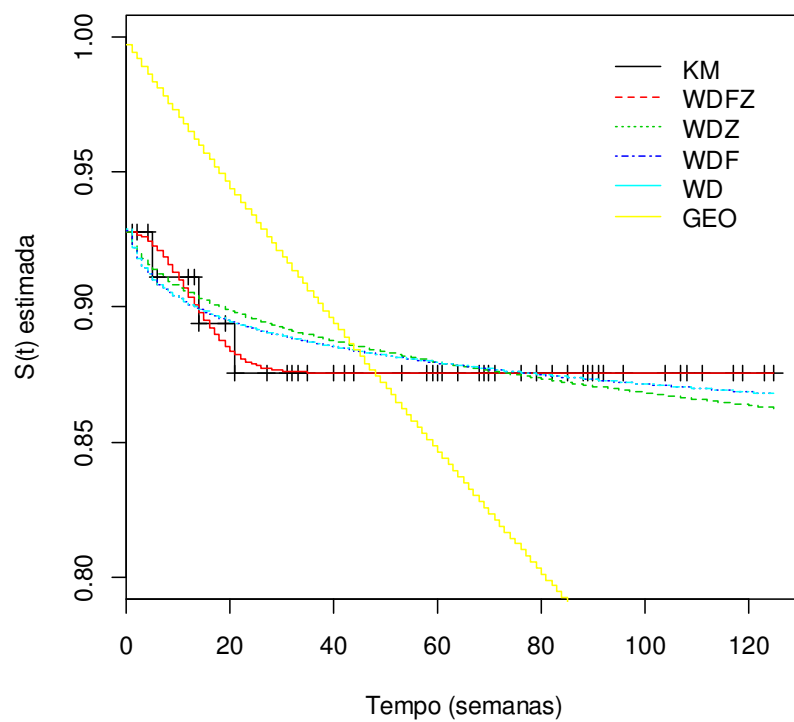


Figura 21 – Curvas de sobrevivência estimadas do tempo de sobrevivências (em semanas) de pacientes submetidos à ICP. (eixo y reduzido)

A Tabela 16 - Tempos de falha e suas respectivas estimativas da função de sobrevivência apresenta os tempos os quais ocorreram falhas e os respectivos valores estimados para cada modelo proposto. É interessante ressaltar a pequena quantidade de falhas (10) observadas no banco de dados. Considerando o tempo em semanas, foram observadas 7 falhas no tempo 0. Isso sugere que um modelo com excesso de zeros pode se ajustar melhor aos dados.

Tabela 16 - Tempos de falha e suas respectivas estimativas da função de sobrevivência.

Tempo (semanas)	Quantidade de falhas	Função de sobrevivência estimada					
		KM	WDFZ	WDZ	WDF	WD	GEO
0	7	0,928	0,928	0,928	0,929	0,929	0,997
5	1	0,911	0,923	0,914	0,910	0,910	0,984
14	1	0,894	0,898	0,903	0,899	0,899	0,959
21	1	0,876	0,882	0,898	0,893	0,894	0,941

Ao calcular a diferença entre o valor estimado de cada modelo e a estimativa de Kaplan-Meier para cada tempo de falha, é possível afirmar que ao comparar as maiores diferenças de cada modelo, o WDFZ possui o melhor ajuste (WDFZ = 0,012; WDZ=0,022; WDF=0,017; WD=0,018 e GEO=0,073).

Para comparar o ajuste dos modelos foram utilizadas as medidas AIC (Akaike information criterion) e BIC (Bayesian information criterion). A Tabela 17 apresenta as medidas encontradas para os dados analisados considerando o tempo em semanas.

Tabela 17 - Comparação das medidas AIC e BIC para cada modelo, considerando as estimativas da Tabela 16.

Medidas	Modelos				
	WDFZ	WDZ	WDF	WD	GEO
AIC	101,0055	102,8217	103,0064	101,0064	140,0144
BIC	111,3044	110,5458	110,7306	106,1558	142,5891

Considerando apenas esses critérios de seleção de modelos, o modelo que apresenta menor AIC é a distribuição Weibull Discreta com fração de cura e excesso de zeros (WDFZ). No entanto, o modelo que apresentou o menor BIC foi a distribuição Weibull Discreta (WD). Porém, observando os gráficos apresentados anteriormente é possível afirmar que tal modelo não obteve um bom ajuste aos dados analisados. Além disso, sabe-se que o BIC penaliza mais os modelos com a maior quantidade de parâmetros, com objetivo de auxiliar na escolha de um modelo mais parcimonioso. Provavelmente a maior penalização considerada para o modelo WDFZ foi o motivo para o modelo WDZ ter apresentado menor BIC.

No teste de Kolmogorov-Smirnov, o ajuste do modelo só será rejeitado, sob um nível de significância  $\alpha$ , se a estatística KS for maior que o quantil  $\alpha$  do teste de Kolmogorov-Smirnov. Considerando  $\alpha = 0,95$ , o ajuste do modelo WDFZ não foi rejeitado pois a estatística KS encontrada foi 0,012, menor que o valor crítico 0,409 (Tabela 1).

A Tabela 18 apresenta as estatísticas do teste obtidas no Teste da Razão de Verossimilhança (TRV) e o valor p para cada teste. Assim como os critérios de informação apresentados anteriormente, o TRV também penaliza modelos com muitos parâmetros. Sendo assim, considerando  $\alpha = 0,05$ , o modelo WDFZ é melhor apenas na comparação com o modelo WDF. No entanto, a Figura 21 ilustra claramente que o modelo WDFZ é o que se ajusta melhor aos dados analisados.

*Tabela 18 - Teste da Razão de Verossimilhança considerando como modelo completo o modelo WDFZ.*

Modelo Completo	Modelo Reduzido	Estatística do teste	gl	valor p
WDFZ	WDZ	3,816135	1	0,0508
WDFZ	WDF	4,000916	1	0,0455
WDFZ	WD	4,000842	2	0,1353

## 6. Considerações finais

A motivação deste trabalho surgiu, particularmente, em um problema real, cuja necessidade foi modelar o tempo de sobrevivência de pacientes submetidos à intervenção coronária percutânea. Esse tipo de dado, geralmente medido em escala discreta (dias, semanas, meses, ...) apresenta, em geral, duas características peculiares: uma parcela desses pacientes vem a óbito logo após a intervenção e outra parcela apresenta sucesso no tratamento. Essa peculiaridade sugere que, após a intervenção, existam as três populações distintas: 1. Pacientes que morreram por complicações que surgiram devido à intervenção (apresentam tempo de sobrevivência igual a zero); 2. Pacientes cuja intervenção pode ter alterado seu tempo de sobrevivência padrão, mas não necessariamente foram curados e; 3. Pacientes cuja intervenção foi bem-sucedida (curados).

Neste contexto, o objetivo deste trabalho foi formular um modelo que considere tempos de sobrevivência discretos com fração de cura e excessos de zeros. O modelo utilizado foi o Weibull discreto (WD) proposto por Nakagawa e Osaki (1975), que corresponde ao modelo Weibull contínuo, muito utilizado na análise do tempo até a ocorrência do evento de interesse. O modelo Weibull discreto com fração de cura e excessos de zeros (WDFZ) proposto se baseou na ideia inicial desse trabalho, em que um indivíduo poder fazer parte de três populações distintas: uma população representada por uma distribuição com o tempo de sobrevivência degenerada no zero, uma população com tempos de sobrevivência segundo uma distribuição Weibull discreta e uma população de curados. Essa formulação permitiu caracterizar o modelo através de uma mistura dessas três distribuições. O modelo WDFZ tem como casos particulares o modelo Weibull discreto (WD) de Nakagawa e Osaki, o modelo Weibull discreto com excesso de zeros (WDZ), o modelo Weibull discreto com fração de cura (WDF) apresentado por (Fernandes, 2014) e o modelo Geométrico (GEO).

Além da formulação do modelo WDFZ, esse trabalho apresentou o desenvolvimento dos seguintes temas, descritos a seguir:

- Elaboração de um algoritmo computacional que permita a realização da Inferência dos parâmetros do modelo WDFZ. Obtenção das estimativas de máxima verossimilhança dos parâmetros;
- Realização de testes de significância dos parâmetros do modelo WDFZ para verificar em que casos um modelo mais simples pode ser considerado para o ajuste dos dados.
- Desenvolvimento de métodos de simulação estocástica para permitir a geração de dados de sobrevivência discretos com fração de curados e excessos de zeros. Os dados foram gerados considerando diversos percentuais de curados, zeros e censuras. Foi considerando um mecanismo aleatório de censura à direita. Obtenção de estimativas de máxima verossimilhança dos parâmetros, assim como seu vício e erro quadrático médio.

Ao término deste trabalho foi possível concluir que o modelo WDFZ se ajusta bem a dados com característica de fração de cura e excesso de zeros. O modelo proposto

também mostrou um bom ajuste para os dados da aplicação sobre o tempo de sobrevivência de pacientes submetidos à intervenção coronária percutânea. Apesar do banco de dados apresentar poucas observações não censuradas, o mesmo apresentou indícios de presença de fração de cura e excessos de zeros. Nesta aplicação, o modelo proposto apresentou um melhor ajuste, quando comparado com os modelos sem fração de cura e/ou excessos de zeros.

Como propostas para trabalhos futuros, pode se indicar a inclusão de covariáveis no modelo, a comparação entre os modelos WDFZ, WDF(Z) e WDZ(F), considerar outras distribuições discretas ao invés da Weibull e realizar a inferência do modelo dentro de um contexto bayesiano.



## BIBLIOGRAFIA

- AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v.19, n.6, p.716-723.
- BERKSON, J.; GAGE, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501-515.
- BRESLOW, N.; CROWLEY, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *Ann. Statist.* 2, no. 3, 437--453. doi:10.1214/aos/1176342705. <http://projecteuclid.org/euclid.aos/1176342705>.
- BRAGA, J. R. (2011). Registros de síndromes coronarianas agudas: O que são, quais suas lições e como interpretá-los corretamente?, Reseach fellow, Division of Cardiology, Toronto General Hospital, Canada.
- BRUNELLO, G. H. V; NAKANO, E. Y. (2015). Inferência bayesiana no modelo Weibull discreto em dado com presença de censura. *TEMA - Tend. Mat. Apl. Comput.*, v.16, n.2, p. 97-110..
- CARRASCO, C. G.; TUTIA, M. H; NAKANO, E. Y. (2012). Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. *TEMA - Tend. Mat. Apl. Comput.*, v.13, n.3, p. 247-255.
- CARVALHO, M. S.; ANDREOZZI, V.L.; CODEÇO, C. T.; BARBOSA, M. T. S. e SHIMAKURA, S.E.; (2011). Análise de sobrevivência: Teoria e aplicações em saúde. FIOCRUZ, RIO DE JANEIRO.
- CASELLA, G.; BERGER, R. L. (2002). *Statistical Inference*. 2. ed. Eua: Cengage Learning Int.
- CHEN, M. H.; IBRAHIM, J. G.; SINHA, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94, 909-919.
- COLOSIMO, E. A.; GIOLO, S. R.. (2006). *Análise de Sobrevivência*. São Paulo: Abe - Projeto Fisher.
- COONER, F.; BANERJEE, S.; CARLIN, B.; SINHA, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of The American Statistical Association*, 102, 560-572.
- COX, D. R.; HINKLEY, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London
- FACHINI, J. B. (2006). Análise de influência local nos modelos de riscos múltiplos. Dissertação (Mestrado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.
- FACHINI, J. B. (2011). Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência. Tese (Doutorado em Ciências) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.
- FERNANDES, L. M. (2014). Inferência bayesiana em modelos discretos com fração de

cura, Dissertação. Mestrado em Estatística, Universidade de Brasília.

GARCIA, P. N. A. (2013). Aplicação de Técnicas de Análise de Sobrevida em pacientes submetidos à intervenção coronária percutânea, Monografia. Bacharelado em Estatística, Universidade de Brasília.

JONHSON, N. X.; KOTZ, S.; KEMP, A. W. (1992). Univariate Discrete Distributions, second edition, John Wiley and Sons, New York.

KALBFLEISCH, J. D.; PRENTICE, R.L. (2002). The statistical analysis of failure time data. 2nd ed. New York: John Wiley. 439p.

LAWLESS, J. F. (1982). Statistical Methods and Models for Lifetime Data. John Wiley & Sons, New York.

MALLER, R. A., ZHOU, X. (1996). Survival Analysis with Long-Term Survivors. Wiley, New York.

MONTEIRO, G. P.; COSTA, R. A.; ABIZAID, A. (2011). Intervenção coronária percutânea primária. Sca revista. São Paulo, Brasil, ano 1, v. 1.

NAKAGAWA, T.; OSAKI, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, v. R-24, n. 5, p.300-301.

NAKANO, E. Y. ; CARRASCO, C. G. (2006). Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *TEMA - Tend. Mat. Apl. Comput.*, v.7, n.1, p. 91-100.

OSPINA, R.; FERRARI, S.L.P. (2010). Inflated beta distributions. *Statistical Papers*, New York, v.51, p.111-126.

R CORE TEAM (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

SCHWARZ, G. (1978). Estimating the dimensional of a model. *Annals of Statistics*, v.6, n.2, p.461-464.

SGRILLO, R. B., (1982). A distribuição de Weibull como modelo de sobrevivência de insetos, Centro de Energia Nuclear na Agricultura – CENA/USP, Piracicaba.

SIEGEL, S; CASTELLAN, N.J. (2006). Estatística não-paramétrica para as ciências do comportamento, Ed. Artmed, 488p.

WEIBULL, W. (1951). A Statistical Distribution Function Of Wide Applicability, *ASME Journal of Applied Mechanics*, Transactions of the American Society Of Mechanical Engineers, September, p. 293-297.

## APÊNDICE: SCRIPT DESENVOLVIDO

```

rm(list=ls(all=TRUE)) # remover todos os objetos anteriores
library(survival)
set.seed(7)

##### DISTRIBUIÇÃO DE PROBABILIDADES DA WEIBULL DISCRETA
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
PWD<-function(t,q,b){ ((q)^(t^b)) - (q)^((t+1)^b) }

##### FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
SWD<-function(t,q,b){ (q)^((t+1)^b) }

##### FUNÇÃO DE RISCO DA WEIBULL DISCRETA
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
HWD<-function(t,q,b){ 1-(q)^((t+1)^b)-(t)^b) }

##### FUNÇÃO DE VEROSSIMILHANÇA DA WEIBULL DISCRETA
LWD <-function(par,t,censura){
  q<-par[1]
  b<-par[2]

  if ( (b>0)&&(q>0)&&(q<1) )
    return (-1*sum( censura*log(PWD(t,q,b)) + (1-
censura)*log(SWD(t,q,b)) ) )
    else return (-Inf)
}

##### FUNÇÃO DE VEROSSIMILHANÇA DA GEOMETRICA
LGEO <-function(q,t,censura){

  if ( (q>0)&&(q<1) )
    return (-1*sum( censura*log(PWD(t,q,1)) + (1-
censura)*log(SWD(t,q,1)) ) )
    else return (-Inf)
}

##### DISTRIBUIÇÃO DE PROBABILIDADES DA WEIBULL DISCRETA COM
FRAÇÃO DE CURADOS
###   fi --- parâmetro que modela a fração de curados
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
PWDF<-function(t,fi,q,b){ (1-fi)*(((q)^(t^b)) - (q)^((t+1)^b)) }

##### FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA COM FRAÇÃO
DE CURADOS
###   fi --- parâmetro que modela a fração de curados
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
SWDF<-function(t,fi,q,b){ fi + (1-fi)*((q)^((t+1)^b)) }

```

```
##### FUNÇÃO DE RISCO DA WEIBULL DISCRETA COM FRAÇÃO DE CURA
###   fi --- parâmetro que modela a fração de curados
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
HWDF<-function(t, fi, q, b) {                                     ((1-fi) ((q)^(t)^b-
(q)^(t+1)^b)) / (fi+(1-fi)*(q)^(t)^b) }

##### FUNÇÃO DE VEROSSIMILHANÇA DA WEIBULL DISCRETA COM
FRAÇÃO DE CURA
LWDF <-function(par, t, censura) {
  fi<-par[1]
  q<-par[2]
  b<-par[3]

  if ( (q>0) && (q<1) && (b>0) && (fi>0) && (fi<1) )
    return ( -1*sum( censura*log(PWDF(t, fi, q, b)) + (1-
censura)*log(SWDF(t, fi, q, b)) ) )
    else return (-Inf)
}

##### DISTRIBUIÇÃO DE PROBABILIDADES DA WEIBULL DISCRETA COM
EXCESSOS DE ZEROS
###   rho --- parâmetro que modela o excesso de zeros
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
PWDZ<-function(t, rho, q, b) {
  zero<-(t==0)
  (1-rho)*(q^(t^b)-q^((t+1)^b)) + rho*zero
}

##### FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA COM
EXCESSOS DE ZEROS
###   rho --- parâmetro que modela o excesso de zeros
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
SWDZ<-function(t, rho, q, b) {
  (1-rho)*q^((t+1)^b)
}

##### FUNÇÃO DE RISCO DA WEIBULL DISCRETA COM EXCESSOS DE
ZEROS
###   rho --- parâmetro que modela o excesso de zeros
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
HWZ<-function(t, rho, q, b) {                                     ((1-rho) ((q)^(t)^b-
(q)^(t+1)^b)+rho*zero) / (1-rho)*(q)^(t)^b+rho*zero }

##### FUNÇÃO DE VEROSSIMILHANÇA DA WEIBULL DISCRETA COM
EXCESSOS DE ZEROS
LWDZ<-function(par, t, censura) {
  rho<-par[1]
  q<-par[2]
  b<-par[3]

  if ( (b>0) && (q>0) && (q<1) && (rho>0) && (rho < 1) )
```

```

    return (-1*sum( censura*log(PWDZ(t,rho,q,b)) + (1-
censura)*log(SWDZ(t,rho,q,b)) ) )
    else return (-Inf)
}

##### DISTRIBUIÇÃO DE PROBABILIDADES DA WEIBULL DISCRETA COM
FRAÇÃO DE CURADOS E EXCESSOS DE ZEROS
### fi --- parâmetro que modela a fração de curados
### rho --- parâmetro que modela o excesso de zeros
### q --- parâmetro do modelo
### b --- parâmetro do modelo
PWDFZ<-function(t,fi,rho,q,b){
  zero<-(t==0)
  (1-fi-rho)*(q^(t^b)-q^((t+1)^b)) + rho*zero
}

##### FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA COM FRAÇÃO
DE CURADOS E EXCESSOS DE ZEROS
### fi --- parâmetro que modela a fração de curados
### rho --- parâmetro que modela o excesso de zeros
### q --- parâmetro do modelo
### b --- parâmetro do modelo
SWDFZ<-function(t,fi,rho,q,b){
  fi + (1-fi-rho)*q^((t+1)^b)
}

##### FUNÇÃO DE RISCO DA WEIBULL DISCRETA COM FRAÇÃO DE CURA
E EXCESSOS DE ZEROS
### fi --- parâmetro que modela a fração de curados
### rho --- parâmetro que modela o excesso de zeros
### q --- parâmetro do modelo
### b --- parâmetro do modelo
HWDFZ<-function(t,fi,rho,q,b){
  ((1-fi-rho)((q)^(t)^b-
(q)^(t+1)^b)+rho*zero)/(fi+(1-fi-rho)*(q)^(t)^b+rho*zero) }

##### FUNÇÃO DE VEROSSIMILHANÇA DA WEIBULL DISCRETA COM
FRAÇÃO DE CURADOS E EXCESSOS DE ZEROS
LWDFZ<-function(par,t,censura){
  fi<-par[1]
  rho<-par[2]
  q<-par[3]
  b<-par[4]

  if ( (b>0)&&(q>0)&&(q<1)&&(fi>0)&&(rho>0)&&(fi+rho < 1) )
  return (-1*sum( censura*log(PWDFZ(t,fi,rho,q,b)) + (1-
censura)*log(SWDFZ(t,fi,rho,q,b)) ) )
  else return (-Inf)
}

#####
##### UTILIZANDO O BANCO DE DADOS #####
#####

dados<-read.table(file.choose(), header = T, sep=";")
names(dados)
attach(dados)

```

```

t<- (dados$TEMPOOBITO)
t<-t-1
censura<-dados$CENSURAOBITO

#####
# Considerando o tempo em semanas completas#
#####
t7<-floor(t/7)
##### OBTENÇÃO DAS ESTIMATIVAS
## WEIBULL DISCRETA COM FRAÇÃO DE CURA E EXCESSO DE ZEROS
est.WDFZ7<-
optim(c(.7, .1, .8, .8), LWDFZ, t=t7, censura=censura, control=(maxit=2
0000))
est.WDFZ7

## WEIBULL DISCRETA COM EXCESSO DE ZEROS
est.WDZ7<-
optim(c(.4, .5, .5), LWDFZ, t=t7, censura=censura, control=(maxit=20000
0))
est.WDZ7

## WEIBULL DISCRETA COM FRAÇÃO DE CURA
est.WDF7<-
optim(c(.4, .5, .5), LWDF, t=t7, censura=censura, control=(maxit=20000
0))
est.WDF7

## WEIBULL DISCRETA
est.WD7<-
optim(c(.1, .5), LWD, t=t7, censura=censura, control=(maxit=200000))
est.WD7

## GEOMETRICA
est.GEO7<-nlm(LGEO, .9, t=t7, censura=censura)
#est.GEO7<- 1-(sum(censura)/(sum(t7)+sum(censura)))
est.GEO7

est7<-
matrix(0,5,4,dimnames=list(c("WDFZ", "WDZ", "WDF", "WD", "GEO"), c("p
hi", "rho", "q", "beta")))
est7[1,]<-c(est.WDFZ7$par)
est7[2,]<-c(0, est.WDZ7$par)
est7[3,]<-c(est.WDF7$par[1], 0, est.WDF7$par[2], est.WDF7$par[3])
est7[4,]<-c(0, 0, est.WD7$par)
est7[5,]<-c(0, 0, est.GEO7$estimate, 0)
round(est7,5)

##### ESTIMADOR EMPIRICO DE KM
KM7<-survfit(Surv(t7, censura)~1)
plot(KM7, conf.int=F, xlab="Tempo (semanas)", ylab="S(t) estimada")

tt7<-seq(0, max(t7))
S.est.WDFZ7<-
SWDFZ(tt7, est.WDFZ7$par[1], est.WDFZ7$par[2], est.WDFZ7$par[3], est
.WDFZ7$par[4])

```

```

points(tt7,S.est.WDFZ7,type="s",col=2,pch=16,lty=1)

S.est.WDZ7<-
SWDZ(tt7,est.WDZ7$par[1],est.WDZ7$par[2],est.WDZ7$par[3])
points(tt7,S.est.WDZ7,type="s",col=3,pch=16,lty=2)

S.est.WDF7<-
SWDF(tt7,est.WDF7$par[1],est.WDF7$par[2],est.WDF7$par[3])
points(tt7,S.est.WDF7,type="s",col=4,pch=16,lty=3)

S.est.WD7<-SWD(tt7,est.WD7$par[1],est.WD7$par[2])
points(tt7,S.est.WD7,type="s",col=5,pch=16,lty=4)

S.est.GEO7<-SWD(tt7,est.GEO7$estimate,1)
points(tt7,S.est.GEO7,type="s",col=7,pch=16,lty=1)

legend (90,.35,bty="n",lty=c(1,2,3,4,1),c("WDFZ","WDZ","WDF",
"WD","GEO"),col=c(2,3,4,5,7))

#Diminuindo o eixo Y para visualizar melhor o gráfico
plot(KM7,conf.int=F,ylim=c(.8,1),xlab="Tempo
(semanas)",ylab="S(t) estimada")
points(tt7,S.est.WDFZ7,type="s",col=2,pch=16,lty=1)
points(tt7,S.est.WDZ7,type="s",col=3,pch=16,lty=2)
points(tt7,S.est.WDF7,type="s",col=4,pch=16,lty=3)
points(tt7,S.est.WD7,type="s",col=5,pch=16,lty=4)
points(tt7,S.est.GEO7,type="s",col=7,pch=16,lty=1)
legend (90,1,bty="n",lty=c(1,2,3,4,1),c("WDFZ","WDZ","WDF",
"WD","GEO"),col=c(2,3,4,5,7))

#Comparando os modelos - AIC e BIC
AICWDFZ7<-(2*est.WDFZ7$value)+(2*4)
AICWDZ7<-(2*est.WDZ7$value)+(2*3)
AICWDF7<-(2*est.WDF7$value)+(2*3)
AICWD7<-(2*est.WD7$value)+(2*2)
AICGEO7<-(2*est.GEO7$minimum)+(2*1)

BICWDFZ7<-(2*est.WDFZ7$value)+4*log(length(t))
BICWDZ7<-(2*est.WDZ7$value)+3*log(length(t))
BICWDF7<-(2*est.WDF7$value)+3*log(length(t))
BICWD7<-(2*est.WD7$value)+2*log(length(t))
BICGEO7<-(2*est.GEO7$minimum)+1*log(length(t))

AIC7<-cbind(AICWDFZ7,AICWDZ7,AICWDF7,AICWD7,AICGEO7)
BIC7<-cbind(BICWDFZ7,BICWDZ7,BICWDF7,BICWD7,BICGEO7)
criterio7<-rbind(AIC7,BIC7)
colnames(criterio7)<-c("WDFZ","WDZ","WDF","WD","GEO")
rownames(criterio7)<-c("AIC","BIC")
criterio7

#TRV: WDFZ WDZ7
TRVWDFZWDZ7<-(-2)*(est.WDFZ7$value-est.WDZ7$value)
TRVWDFZWDZ7
pvalor<-1-pchisq(TRVWDFZWDZ7,1)
pvalor

```

```

#TRV: WDFZ WDF7
TRVWDFZWDF7<-(-2)*(est.WDFZ7$value-est.WDF7$value)
TRVWDFZWDF7
pvalor<-1-pchisq(TRVWDFZWDF7,1)
pvalor

#TRV: WDFZ WD7
TRVWDFZWD7<-(-2)*(est.WDFZ7$value-est.WD7$value)
TRVWDFZWD7
pvalor<-1-pchisq(TRVWDFZWD7,2)
pvalor

#Tempos de falha
TFWDFZ7<-
c(S.est.WDFZ7[1],S.est.WDFZ7[6],S.est.WDFZ7[15],S.est.WDFZ7[22])
TFWDZ7<-
c(S.est.WDZ7[1],S.est.WDZ7[6],S.est.WDZ7[15],S.est.WDZ7[22])
TFWDF7<-
c(S.est.WDF7[1],S.est.WDF7[6],S.est.WDF7[15],S.est.WDF7[22])
TFWD7<-c(S.est.WD7[1],S.est.WD7[6],S.est.WD7[15],S.est.WD7[22])
TFGEO7<-
c(S.est.GEO7[1],S.est.GEO7[6],S.est.GEO7[15],S.est.GEO7[22])
TF<-cbind(TFWDFZ7,TFWDZ7,TFWDF7,TFWD7,TFGEO7)
TF

## KS
summary(KM7)
ttt<-c(0,5,14,21) # tempos de falha
S.est.KSWDFZ7<-
SWDFZ(ttt,est.WDFZ7$par[1],est.WDFZ7$par[2],est.WDFZ7$par[3],est
.WDFZ7$par[4])
round(S.est.KSWDFZ7,3)

KS<-max(abs(S.est.WDFZ7<-
SWDFZ(KM7$time,est.WDFZ7$par[1],est.WDFZ7$par[2],est.WDFZ7$par[3
],est.WDFZ7$par[4])-KM7$urv))
KS

#####
##### SIMULAÇÃO #####
#####

#####
##CENÁRIO 1 ##   ### Para os demais cenários, basta alterar
#####   ### os valores dos parâmetros do modelo

##### GERAR VALORES DA WEIBULL DISCRETA COM FRAÇÃO DE CURADOS
E EXCESSO DE ZEROS
fi<-0.05   ## proporção de curados
rho<-0.05  ## proporção de zeros
q<-0.9     ## parâmetro do modelo
b<-0.9     ## parâmetro do modelo
n<-200     ## tamanho da amostra
p.cens<-0.1 ## proporção de censura (dos não curados)

set.seed(12345)

```



```
##### FUNÇÃO PARA GERAR VALORES DA WEIBULL DISCRETAR
###   n --- tamanho da amostra
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
rweidc<-function(n,q,b) {
mt<-matrix(ncol=n,nrow=1,0)
m<-PWD(0:20000,q,b)
for (i in 2:20001) {
m[i]<-sum(m[i],m[i-1])    }
for (j in 1:n) {
k<-runif(1,0,1)
mt[j]<-which(m>k)[1]-1    }
as.vector(mt)  }

ni<-5   ### quantidade de amostras (50, 100, 200, 400, 800)
k=matrix(0,5,1)

estimativas.WDFZ<-
matrix(0,m,4,dimnames=list((1:m),c("phi","rho","q","beta")))
estimativas.WDZ<-
matrix(0,m,3,dimnames=list((1:m),c("rho","q","beta")))
estimativas.WDF<-
matrix(0,m,3,dimnames=list((1:m),c("phi","q","beta")))
estimativas.WD<-matrix(0,m,2,dimnames=list((1:m),c("q","beta")))
estimativas.GEO<-matrix(0,m,1,dimnames=list((1:m),c("q")))

t<-numeric(n)
censura<-numeric(n)
nx<-rbinom(1,n,(1-fi-rho))      # gerando a quantidade de X
nz<-rbinom(1,n-nx,rho/(fi+rho)) # gerando a quantidade de Z
ny<-n-nx-nz                    # gerando a quantidade de Y

x<-rweidc(nx,q,b)
censura.x<-rbinom(nx,1,1-p.cens)

t<-c(x,rep(max(x),ny),rep(0,nz))
censura<-c(censura.x,rep(0,ny),rep(1,nz))

##### OBTENÇÃO DAS ESTIMATIVAS
est.WDFZ<-
optim(c(fi,rho,q,b),LWDFZ,t=t,censura=censura,control=(maxit=200
000))
est.WDZ<-
optim(c(rho,q,b),LWDZ,t=t,censura=censura,control=(maxit=200000)
)
est.WDF<-
optim(c(fi,q,b),LWDF,t=t,censura=censura,control=(maxit=200000))
est.WD<-
optim(c(q,b),LWD,t=t,censura=censura,control=(maxit=200000))
est.GEO<-nlm(LGEO,q,t=t,censura=censura)
estimativas.WDFZ[1,]<-est.WDFZ$par
estimativas.WDZ[1,]<-est.WDZ$par
estimativas.WDF[1,]<-est.WDF$par
estimativas.WD[1,]<-est.WD$par
estimativas.GEO[1,]<-est.GEO$estimate
```

```

matriz.WDFZ<-
c(n,p.cens,estimativas.WDFZ[,1],estimativas.WDFZ[,2],estimativas
.WDFZ[,3],estimativas.WDFZ[,4])
matriz.WDZ<-
c(n,p.cens,NA,estimativas.WDZ[,1],estimativas.WDZ[,2],estimativa
s.WDZ[,3])
matriz.WDF<-
c(n,p.cens,estimativas.WDF[,1],NA,estimativas.WDF[,2],estimativa
s.WDF[,3])
matriz.WD<-
c(n,p.cens,NA,NA,estimativas.WD[,1],estimativas.WD[,2])
matriz.GEO<-c(n,p.cens,NA,NA,estimativas.GEO[,1],NA)

matriz.est.WDFZ<-data.frame(0,0,0,0,0,0)
colnames(matriz.est.WDFZ)<-c("n","%censura","phi", "rho", "q",
"beta")
matriz.est<-
rbind(matriz.WDFZ,matriz.WDZ,matriz.WDF,matriz.WD,matriz.GEO)
colnames(matriz.est)<-c("n","%censura","phi", "rho", "q",
"beta")
round(matriz.est,5)

#Comparando os modelos - AIC e BIC
AICWDFZ<-(2*est.WDFZ$value)+(2*4)
AICWDZ<-(2*est.WDZ$value)+(2*3)
AICWDF<-(2*est.WDF$value)+(2*3)
AICWD<-(2*est.WD$value)+(2*2)
AICGEO<-(2*est.GEO$minimum)+(2*1)

BICWDFZ<-(2*est.WDFZ$value)+4*log(length(t))
BICWDZ<-(2*est.WDZ$value)+3*log(length(t))
BICWDF<-(2*est.WDF$value)+3*log(length(t))
BICWD<-(2*est.WD$value)+2*log(length(t))
BICGEO<-(2*est.GEO$minimum)+1*log(length(t))

AIC<-cbind(AICWDFZ,AICWDZ,AICWDF,AICWD,AICGEO)
BIC<-cbind(BICWDFZ,BICWDZ,BICWDF,BICWD,BICGEO)
criterio<-rbind(AIC,BIC)
colnames(criterio)<-c("WDFZ","WDZ","WDF", "WD", "GEO")
rownames(criterio)<-c("AIC","BIC")
criterio

##### Gráfico
KM<-survfit(Surv(t,censura)~1)
plot(KM,conf.int=F,xlab="Tempo",ylab="S(t) estimada",main=n)

tt<-seq(0,max(t))
S.est.WDFZ<-
SWDFZ(tt,mean(estimativas.WDFZ[,1]),mean(estimativas.WDFZ[,2]),m
ean(estimativas.WDFZ[,3]),mean(estimativas.WDFZ[,4]))
points(tt,S.est.WDFZ,type="s",col=2,pch=16,lty=1)

S.est.WDZ<-
SWDZ(tt,mean(estimativas.WDZ[,1]),mean(estimativas.WDZ[,2]),mean
(estimativas.WDZ[,3]))

```

```

points(tt,S.est.WDZ,type="s",col=3,pch=16,lty=2)

S.est.WDF<-
SWDF(tt,mean(estimativas.WDF[,1]),mean(estimativas.WDF[,2]),mean
(estimativas.WDF[,3]))
points(tt,S.est.WDF,type="s",col=4,pch=16,lty=3)

S.est.WD<-
SWD(tt,mean(estimativas.WD[,1]),mean(estimativas.WD[,2]))
points(tt,S.est.WD,type="s",col=5,pch=16,lty=4)

S.est.GEO<-SWD(tt,mean(estimativas.GEO[,1]),1)
points(tt,S.est.GEO,type="s",col=7,pch=16,lty=1)

legend      (0.75*max(t),1,bty="n",lty=c(1,2,3,4,1),c("WDFZ","WDZ",
"WDF", "WD", "GEO"), col=c(2,3,4,5,7))

#Diminuindo o eixo Y para visualizar melhor o gráfico
plot(KM,conf.int=F,xlab="Tempo",ylab="S(t)
estimada",main=n,ylim=c(.03,1),xlim=c(0,40))
points(tt,S.est.WDFZ,type="s",col=2,pch=16,lty=1)
points(tt,S.est.WDZ,type="s",col=3,pch=16,lty=2)
points(tt,S.est.WDF,type="s",col=4,pch=16,lty=3)
points(tt,S.est.WD,type="s",col=5,pch=16,lty=4)
points(tt,S.est.GEO,type="s",col=7,pch=16,lty=1)
legend      (28,1,bty="n",lty=c(1,2,3,4,1),c("WDFZ","WDZ",      "WDF",
"WD", "GEO"), col=c(2,3,4,5,7))

#### VICIO E EQM
inicio<-format(Sys.time(), "%a %b %d %H:%M:%S %Y")
m=1000 ### número de réplicas de Monte Carlo
p.cens<-0

matriz.est.WDFZ<-data.frame(0,0,0,0,0,0)
colnames(matriz.est.WDFZ)<-c("n","%censura","phi", "rho", "q",
"beta")

#### MATRIZ DE VICIOS
vicios<-
matrix(0,4,1,dimnames=list(c("phi","rho","q","beta"),"WDFZ"))

#### MATRIZ DE EQM
EQM<-
matrix(0,4,1,dimnames=list(c("phi","rho","q","beta"),"WDFZ"))

matriz.VE<-data.frame(0,0,0)
matriz.VE.final<-matrix(0,4,4)

for (h in 1:3){
  n<-50      ## tamanho da amostra
  for (j in 1:5){
    for (i in 1:m){
      t<-numeric(n)
      censura<-numeric(n)

```

```

        nx<-rbinom(1,n,(1-fi-rho))           # gerando a
quantidade de X
        nz<-rbinom(1,n-nx,rho/(fi+rho))     # gerando a
quantidade de Z
        ny<-n-nx-nz                         # gerando a
quantidade de Y

        x<-rweidc(nx,q,b)
        censura.x<-rbinom(nx,1,1-p.cens)

        t<-c(x,rep(max(x),ny),rep(0,nz))
        censura<-c(censura.x,rep(0,ny),rep(1,nz))

        ##### OBTENÇÃO DAS ESTIMATIVAS
        est.WDFZ<-
optim(c(fi,rho,q,b),LWDFZ,t=t,censura=censura,control=(maxit=200
000))
        estimativa.WDFZ<-
matrix(c(n,p.cens,est.WDFZ$par),1,6,dimnames=list(c(1),c("n","%c
ensura","phi","rho","q","beta")))
        matriz.est.WDFZ<-
rbind(matriz.est.WDFZ,estimativa.WDFZ)
        cat(i,"\n")
}

### VICIO WDFZ
vicios[1,1]<-mean(matriz.est.WDFZ[,3]) - fi
vicios[2,1]<-mean(matriz.est.WDFZ[,4]) - rho
vicios[3,1]<-mean(matriz.est.WDFZ[,5]) - q
vicios[4,1]<-mean(matriz.est.WDFZ[,6]) - b

### EQM WDFZ
EQM[1,1]<- (mean(matriz.est.WDFZ[,3]) - fi)^2
EQM[2,1]<- (mean(matriz.est.WDFZ[,4]) - rho)^2
EQM[3,1]<- (mean(matriz.est.WDFZ[,5]) - q)^2
EQM[4,1]<- (mean(matriz.est.WDFZ[,6]) - b)^2

matriz.p.cens<-matrix(p.cens,4,1)
matriz.n<-matrix(n,4,1)
matriz.VE<-cbind(matriz.n,matriz.p.cens,vicios,EQM)
colnames(matriz.VE)<-c("n","%censura","Vicio","EQM")
matriz.VE.final<-rbind(matriz.VE.final,matriz.VE)

k[j,]<-n
print(n)
n<-n*2
}

print(p.cens)
p.cens<-p.cens+0.1

}

###RESULTADOS###
matriz.VE.final<-matriz.VE.final[-1,]
matriz.VE.final<-matriz.VE.final[-1,]

```

```
matriz.VE.final<-matriz.VE.final[-1,]  
matriz.VE.final<-matriz.VE.final[-1,]  
round(matriz.VE.final,5)  
  
round(matriz.est.WDFZ[-1,],5)  
  
fim<-format(Sys.time(), "%a %b %d %H:%M:%S %Y")  
inicio  
fim
```