Universidade Federal de Santa Catarina
Programa de Pós-Graduação em Engenharia de Produção

APLICAÇÃO DE MODELOS DE MISTURA DE LONGA DURAÇÃO EM DADOS DE REINCIDÊNCIA AO CRIME

Dissertação de Mestrado

Angela Maria Marcone de Araujo

Florianópolis
2004

APLICAÇÃO DE MODELOS DE MISTURA DE LONGA DURAÇÃO EM DADOS DE REINCIDÊNCIA AO CRIME

Dissertação apresentada no
Programa de Pós-Graduação em
Engenharia de Produção a
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do grau de Mestre em
Engenharia de Produção.

Orientador: Pedro Alberto Barbetta, Dr.

Florianópolis
2004

Angela Maria Marcone de Araujo

Aplicação de Modelos de Mistura de Longa Duração em dados de Reincidência ao Crime

Esta dissertação foi julgada adequada e aprovada para a obtenção do título de **Mestre em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina**

Florianópolis, 08 de janeiro de 2004.

Edson Pacheco Paladini, Dr.
Coordenador do Programa

BANCA EXAMINADORA

Prof. Pedro Alberto Barbetta, Dr.
Orientador

Prof.Josmar Mazucheli, Dr.

Prof. Paulo José Ogliari, Dr.

Ao meu esposo, José Luiz
pelo apoio constante.
A meus filhos Maria Gabriela,
Ana Roberta e Luiz Francisco.

AGRADECIMENTOS

À Deus por ter concedido a graça de mais uma realização;

À minha família pela compreensão nas ausências;

Ao Departamento de Estatística da UEM, pela colaboração e apoio;

À todos os alunos do curso pelo apoio e amizade e em especial a Clédina, Valentina, Clara, Zeza e Gazola;

Ao Programa de Pós-graduação em Engenharia de Produção da UFSC pelo empenho;

Ao Coronel Antonio Tadeu Rodrigues pela viabilização da coleta de dados;

Ao Prof Josmar Mazucheli pela atenção, colaboração e atuação de um co-orientador ;

Ao meu orientador, Prof. Pedro Alberto Barbetta, que apoiou e contribuiu para esta concretização;

a todos que direta ou indiretamente
contribuíram para a realização
desta pesquisa.

Resumo

ARAUJO, Angela Maria Marcone. **Aplicação de Modelos de Mistura de Longa Duração em dados de Reincidência ao Crime**. 2003. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

Segundo dados do Governo Federal, a média nacional do índice de reincidência de ex-detentos no Brasil é considerado elevado, 82%. No Paraná, esta taxa é de 30%, um dos menores índices do país. Entretanto, estes números não fornecem informações a respeito do tempo até que ocorra a primeira reincidência, como também não informam se existe alguma relação entre o tipo de benefício e a reincidência. Para cada indivíduo colocado em liberdade, pode-se definir uma variável aleatória que indica o tempo até a reincidência e então tentar modelar esta variável em função de alguma covariável, como por exemplo o tipo de benefício adquirido ao sair em liberdade. Na modelagem de dados relacionados ao tempo até a ocorrência de algum evento de interesse, utiliza-se métodos estatísticos de análise de sobrevivência. Nesta pesquisa buscou-se modelar o tempo de reincidência de 1172 ex-detentos da PEM - Penitenciária Estadual de Maringá no período de abril de 1996 a dezembro de 2002. A existência de um grande número de observações censuradas, induziu uma modelagem através de modelos de mistura de longa duração. O modelo forneceu a proporção, de ex-detentos imunes à reincidência, isto é, que não voltarão a reincidir ao crime, de acordo com os benefícios adquiridos para sair da PEM.

Palavras-chave: Análise de Sobrevivência, Modelos de Mistura de Longa Duração, Reincidência ao Crime.

Abstract

ARAUJO, Angela Maria Marcone. **Application of Long-term Mixture Models on Recidivism to Crime Data**. 2003. Dissertation (Master Course in Production Engineering) – Post-Graduation Program in Production Engineering, UFSC, Florianópolis.

According to the Federal Government data, the national average of the ex-prisoners recidivism index in Brazil is considered high, 82%. In Paraná state, this rate is of 30%, one of the lowest indexes in the country. However, these numbers neither provide information about the time until the first recidivism occurs, nor give information if there is any relation between the kind of benefit and the recidivism. For every subject to whom freedom is given, a random variable that indicates the time until the relapse can be defined, and, then, one tries to model this variable in function of some co-variables, such as, the kind of benefit acquired when going out in freedom. Concerning the data modeling related to time until the occurrence of some event of interest, survival analysis statistic methods are used. In this research we tried to model the relapse time of 1172 ex-prisoners from PEM – *Penitenciária Estadual de Maringá* (State Prison of Maringá city) from April, 1996 to December, 2002. The existence of a great number of censored observations induced a modelling by long-term mixture models. The model has given the proportion of ex-prisoners immune to relapse, that is, those who have not committed a crime anymore, according to the benefits acquired to be free from PEM.

Key-words: Survival Analysis. Long Term Mixture Models. Recidivism to crime.

Sumário

Capítulo 1

Introdução

1.1 Contextualização

Um dos estudos da área estatística que tem apresentado um crescimento freqüente é o da análise de sobrevivência. A análise de sobrevivência consiste de procedimentos estatísticos para análise de dados relacionados ao tempo até a ocorrência de algum evento de interesse, a partir de um tempo inicial até um tempo final pré-definido.

É importante observar que o uso da palavra "sobrevivência" vem do fato de que muitas vezes o evento de interesse é a morte, mas o raciocínio é o mesmo para qualquer outro evento de interesse em estudo.

Ao contrário que muitas pessoas possam imaginar, cartões de crédito, máquinas industriais, garantias de eletrodomésticos, ex-detentos, pacientes com doenças grave e recorrência de um tumor, possuem mais coisas em comum do que se pode supor a primeira vista. Ou seja, eles correm riscos que podem ser: inadimplência, falha, quebra, reincidência e morte. E os métodos de análise desses eventos, chama-se análise de sobrevivência.

Na área de engenharia, a análise de sobrevivência é conhecida como confiabilidade industrial e possui modelos estatísticos que permitem prever quando um equipamento pode falhar e, portanto, deve parar para a manutenção ou substituição de peças.

A análise do tempo que decorre até a ocorrência de um evento, envolve a

estimação da probabilidade de que um evento ocorrerá em pontos diferentes do tempo. A análise de sobrevivência estima a probabilidade de sobrevivência como uma função do tempo, a contar da data de partida.

Para dados que envolvam tempo de estudo, outros métodos estatísticos podem não ser eficientes, pois pode acontecer do evento de interesse não ter ocorrido em todos os sujeitos que estão fazendo parte do estudo, necessitando assim a introdução da variável censura, que indica que alguns dos participantes do estudo não experimentaram o evento e a análise de sobrevivência acomoda dados com esta característica peculiar.

Muitas vezes, pode também ocorrer de que algumas informações a respeito dos sujeitos que participam do evento de interesse podem esclarecer melhor os resultados obtidos. Essas informações adicionais são chamadas de covariáveis, e podem ser desde sexo, raça, uso de um medicamento, ou até mesmo o tipo de tratamento dos sujeitos envolvidos no estudo.

Também pode ocorrer que o conjunto de dados que está sendo analisado tenha um número muito grande de indivíduos com seus tempos de estudo censurados, sugerindo, então, a existência de uma proporção de indivíduos não suscetíveis (imunes) ao evento de interesse. Nessas condições, o conjunto de dados deve ser analisado como um modelo de mistura de longa duração.

Os modelos de mistura de longa duração pertencem a uma classe plausível de modelos para a distribuição de tempos até a ocorrência de algum evento de interesse, pois os mesmos consideram a existência na população de uma fração de participantes considerados "imunes", nos quais seus tempos de vida são observados somente como observações censuradas. Talvez o modelo de mistura de longa duração mais popular seja o discutido por Berkson e Gage (1952).

Quando ocorre de um evento de interesse ter um número grande de observações censuradas na amostra, em tempos altos, pode-se inferir a presença de participantes "imunes". Como não é possível identificar quais participantes são realmentes "imunes", devido ao fato de que seus tempos de vida são indistinguíveis dos tempos realmente censurados, existe a necessidade de se fazer um teste estatístico para concluir se estes participantes considerados "imunes"

ao evento de interesse o são realmente.

O presente trabalho tem como evento de interesse o tempo até a reincidência de ex-detentos a algum delito. Ou seja, verificar se os ex-detentos que tiveram seus tempos de liberdade censurados em 31 de dezembro de 2002 podem fazer parte da fração de indivíduos considerados "imunes", isto é, que não voltarão a reincidir no crime. Também, deseja-se analisar se as covariáveis (profissão do ex- detento, enquadramento do delito cometido, benefício utilizado para sair da PEM, os cursos profissionalizantes realizados na PEM, o estudo formal fornecido pela PEM) alteram a probabilidade de imunes.

Devido ao fato de que o conjunto de dados dos ex-detentos da Penitenciária Estadual de Maringá - PEM apresentarem um número muito grande de observações censuradas, será adotado o modelo de mistura de longa duração.

Além da análise gráfica será realizado o teste da razão de verossimilhanças, ou diferença das deviances, para testar hipóteses sobre os parâmetros dos modelos, ou seja, verificar se há evidência da existência de uma proporção de indivíduos imunes a reincidir ao crime.

1.2 Problema

É possível utilizar os modelos de mistura de longa duração para estimar a porcentagem de "imunes"no conjunto dos ex-detentos da PEM?

1.3 Objetivo Geral

Aplicar modelos de mistura de longa duração, em dados reais (dados do sistema penitenciário de Maringá), o que permitirá estimar a proporção de ex-detentos imunes à reincidir ao crime.

1.4 Objetivos Específicos

a) Apresentar os principais modelos de análise de sobrevivência com suas respectivas funções densidade de probabilidade, de sobrevivência e de risco;

b) Descrever os modelos de mistura de longa duração exponencial e Weibull e métodos para a estimação de seus parâmetros;

c) Aplicar o modelo de mistura de longa duração Weibull nos dados do sistema penitenciário;

d) Verificar se existe uma proporção de ex-detentos que não reincidirão no crime, ou seja, verificar se existem ex-detentos "imunes" a reincidência. Havendo evidência de ex-detentos imunes, estimar a proporção de imunes por tipo de benefício adquirido ao sair da PEM.

1.5 Métodos

Com a finalidade de alcançar o objetivo geral, utilizar-se-á modelos de mistura de longa duração para verificar se existe uma proporção de indivíduos que não são suscetíveis ao evento de interesse, que nesta pesquisa trata-se da reincidência de ex-detentos.

Porém, para realizar o estudo de modelos de longa duração, faz-se necessário uma abordagem inicial em análise de sobrevivência, desde a presença de censura, enfocando as distribuições de probabilidades mais utilizadas, bem como a técnica de estimação via teoria de verossimilhança.

1.6 Estrutura

Este trabalho está organizado em 6 capítulos de forma a facilitar o entendimento e compreensão do leitor.

No primeiro capítulo tem-se uma contextualização do assunto, problema, objetivos, métodos de desenvolvimento e sua estrutura.

No capítulo dois, apresenta-se uma revisão de literatura sobre análise de sobrevivência, desde os conceitos de censura, passando pelas distribuições de probabilidades mais utilizadas, bem como técnicas para estimar seus parâmetros.

O terceiro capítulo continuará com a revisão de literatura que enfocará modelos de longa duração que serão utilizados para que os objetivos deste

trabalho sejam atingidos. Apresentando ainda, exemplos para esclarecer melhor a teoria mencionada.

No capítulo quatro será relatado uma breve apresentação do sistema penitenciário com particular enfoque à PEM, como o banco de dados foi constituído e será realizada uma análise com os dados dos ex-detentos que passaram pela PEM no período de 1996 a 2002.

Finalizando, o capítulo cinco apresenta algumas conclusões e e propostas para pesquisas futuras.

Capítulo 2

Revisão da Literatura

2.1 Conceitos Básicos de Análise de Sobrevivência

O termo análise de sobrevivência refere-se ao estudo de dados relacionados ao tempo até a ocorrência de um determinado evento de interesse, a partir de um tempo inicial até um tempo final pré-definido. A diferença entre o tempo final e o tempo inicial é definido como tempo de estudo.

A análise de sobrevivência está associada a muitas áreas, em especial nas ciências médicas e na engenharia. Nas ciências médicas, o evento de interesse é o tempo transcorrido entre o tempo de entrada do indivíduo no estudo e o tempo quando o indivíduo apresenta características para o término do estudo. Esse evento pode ser o tempo transcorrido até a morte do indivíduo, o tempo de duração de uma determinada doença ou a complicação da mesma, o tempo do alívio da dor, o tempo de cura, entre outros. Já na área da engenharia, o evento de interesse está relacionado ao tempo até a falha (degradação, incapacidade total ou parcial do funcionamento) de um determinado artigo manufaturado. Esse evento pode ser: o tempo de ocorrência de uma falha fatal ou não; o tempo até a ocorrência de um reparo ou o tempo até a utilização da garantia de um produto.

Os métodos empregados na análise de dados provenientes de tempo de sobrevivência não estão restritos apenas para tempo de sobrevivência na forma literal, mas aplica-se igualmente para dados de outras áreas, tais como o tempo de sobrevivência de animais num estudo experimental, o tempo que

um indivíduo leva para completar uma tarefa num experimento de psicologia, o tempo de armazenamento de sementes até sua venda, o tempo para a reincidência ao crime, a vida industrial de um componente eletrônico, etc.

Para um melhor entendimento do evento de interesse a ser estudado, considere os seguintes exemplos comentados por Lawless (1982):

1. Artigos fabricados como componentes eletrônicos ou mecânicos são freqüentemente testados com o objetivo de revelar informações sobre suas resistências. Esses testes de vida geralmente são feitos em laboratórios, onde o evento de interesse é o tempo que transcorre até a ocorrência da falha.

2. Alguns tipos de artigos fabricados podem ser reparados, caso venham a falhar. Assim, o evento de interesse seria o tempo entre sucessivos fracassos ou falhas dos artigos que foram reparados.

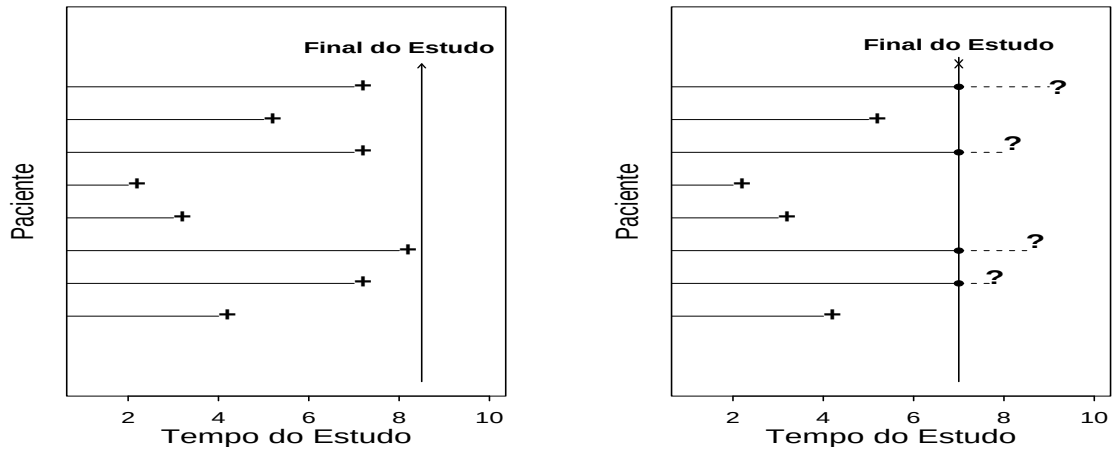
3. Em estudos médicos com pacientes sofrendo de uma doença fatal, um evento de interesse é o tempo de sobrevivência do indivíduo com a doença, a partir da data do diagnóstico ou algum outro ponto de partida. Sendo comum a comparação de tratamentos, pelo menos em parte, para pacientes que estão recebendo tratamentos diferentes.

4. Um experimento padrão na investigação de substâncias cancerígenas é aquele em que animais de um laboratório estão sujeitos a certas quantidades de doses e são observados para verificar se desenvolvem tumores. A variável principal de interesse nestes experimentos é o tempo até o aparecimento do tumor ou, talvez, até a morte do animal, medido desde que a dose é administrada.

Sem especificação da área de interesse, o tempo inicial de contagem, em geral, coincide com o final do tempo de recrutamento dos indivíduos ou itens para um determinado estudo experimental, enquanto que o tempo final dependerá do interesse específico que se deseja no estudo e da disponibilidade dos recursos necessários para a execução da pesquisa, sendo que as populações em estudo devem estar sempre sob as mesmas condições.

A variável de interesse, o tempo de sobrevivência ou até a falha, é uma variável aleatória estritamente positiva e, geralmente, medida em escala contínua.

O tempo de estudo do decorrer de um evento de interesse é ilustrado na figura 2.1.



(a) (b)

Figura 2.1: (a) Todos os pacientes experimentam o evento antes do final do estudo. (b) No final do estudo alguns pacientes ainda não haviam experimentado o evento de interesse.

2.2 A Presença de Censuras

Uma característica presente em análise de sobrevivência é o fato da variável de interesse (variável resposta), tempo de sobrevivência, muitas vezes não poder ser medida instantaneamente e independentemente do tamanho da resposta, o que pode comprometer a observação do valor da variável, dado que para alguns indivíduos o evento de interesse pode não ocorrer até o término do estudo. Também pode ocorrer do indivíduo abandonar o estudo antes da observação do evento de interesse ocorrer ou falecer devido a outras causas de morte, mas não a que está sendo estudada. Dessa maneira, para verificar se o valor do tempo de sobrevivência de um determinado indivíduo foi ou não observado no evento de interesse, existe a necessidade da introdução de uma

variável extra, conhecida como censura.

No decorrer destas situações, indivíduos podem entrar no estudo no tempo t_0 e morrer no tempo $t_0 + t$. Embora t seja desconhecido, porque um ou outro indivíduo pode estar vivo ou ter abandonado o estudo durante a observação. Se o indivíduo for o último conhecido vivo do evento de interesse no tempo $t_0 + c$, o tempo c é chamado de censura do tempo de sobrevivência, conforme Collett (1994).

Quando ocorre de todos os indivíduos observados experimentarem o evento de interesse antes do término do estudo não existe censura. Esta é a situação mostrada na Figura 2.1-a.

De acordo com Lawless (1982) e Collett (1994), existem várias formas de censuras: censuras à direita ou censuras do tipo I; censuras do tipo II; censuras aleatórias; censuras à esquerda; censuras intervalares etc.

Os tipos de censuras mais usuais são censura à direita ou tipo I, censura tipo II e a censura aleatória, pois os demais tipos ocorrem com menos frequência em situações reais.

As censuras à direita e as censuras aleatórias são mais freqüentes nos estudos da área de ciências médicas, devido ao fato do pesquisador fixar um tempo máximo de observação para que o evento de interesse possa ocorrer. Enquanto que na área de engenharia, as censuras do tipo II são predominantes, pois o pesquisador prolonga o período de observação do evento de interesse até que ocorra r falhas de n possibilidades $(r < n)$.

Censura do tipo I ou censura à direita é aquela na qual o teste será terminado após um período pré-estabelecido de tempo e alguns dos tempos de sobrevivência não foram observados, situação esta mostrada na Figura $2.1 - b$. Como ilustração veja o exemplo citado por Borges, Colosimo e Freitas (1996). Considere o mecanismo de acionamento do vidro das portas de um veículo como parte essencial para o conforto do mesmo. Afim de obter informações importantes sobre a confiabilidade do produto, os fabricantes submetem os mecanismos a testes de funcionamento acelerado (colocando em uso muito mais intenso que o uso em condições normais). Um tipo de teste comum é

aquele em que o mecanismo é colocado sob o uso contínuo com uma carga correspondente ao peso do vidro, até ocorrer a falha, ou seja, parar de operar. Um lote de 30 mecanismos novos foram colocados em teste, que consistia em deixá-los funcionando por um período pré-estabelecido até completar os 50.000 ciclos (um ciclo corresponde ao ato de descer e subir o vidro) e registrou-se, para cada mecanismo, o número de ciclos que ele completou até falhar (tempo até a falha). Após os 50.000 ciclos (tempo de censura), foram registrados que 18 mecanismos haviam falhado e que o restante continuava funcionando.

Censura do tipo II é aquela em que o estudo será terminado após ter ocorrido a falha de um número pré-estabelecido de itens no decorrer do estudo. Por exemplo, para detectar o tempo de vida de lâmpadas, podemos realizar um experimento, no qual serão colocadas em situação de estresse 160 lâmpadas. O experimento será encerrado após a 45^a (número pré-estabelecido pelo experimentador) lâmpada queimar. O tempo que a 45^a lâmpada gastou para queimar será estipulado como tempo médio de vida para lâmpadas deste tipo.

Censura aleatória ocorre quando o paciente deixa o estudo sem ter experimentado o evento de interesse. Este tipo de censura pode ocorrer porque o indivíduo abandona o estudo, morre de uma outra causa que não seja a que está sendo estudada ou permanece vivo até o fim do estudo. Por exemplo, suponha que alguns indivíduos são recrutados para um estudo e um desses indivíduos muda-se de cidade e até mesmo de país e não é possível continuar fazendo sua observação. Como outro exemplo, considere que um indivíduo está sendo acompanhado em um estudo sobre o câncer de próstata e morre em um acidente automobilístico.

Censura à esquerda ocorre quando o evento de interesse já aconteceu em um momento anterior ao início do estudo. Essa censura é encontrada quando o tempo de sobrevivência atual de um indivíduo é menor que o observado. Como exemplo, uma pesquisa realizada na California, em escolas secundárias, tinha por objetivo verificar a distribuição do tempo até o uso da maconha pela primeira vez em meninas e meninos. A pesquisa obteve algumas respostas do tipo "já usou, porém não recorda quando foi a primeira vez". Os respondentes que forneceram esta resposta estão indicando que o evento já tinha acontecido

antes da entrevista, porém a idade exata que ele começou usar maconha é desconhecida e inferior a data que a pesquisa foi realizada.

Censura intervalar ocorre quando pacientes em uma tentativa clínica ou estudo longitudinal têm seguimento periódico e o tempo de estudo do paciente só é conhecido dentro de um intervalo. Tal tipo de censura também pode acontecer em experiências industriais em que existe inspeção periódica no funcionamento de artigos de equipamentos. Como exemplo, considere um estudo em que o interesse central é o tempo necessário para a recorrência de um câncer após a cirurgia de remoção de um tumor primário, no qual o tempo de estudo será o intervalo entre o terceiro e o sexto mês após a cirurgia.

Dependendo das condições em que o estudo será realizado e das informações históricas sobre o produto ou indivíduo em estudo, existem vantagens na utilização de algum tipo de censura com relação a outros tipos, porém, na prática, o tratamento estatístico para dados censurados geralmente é o mesmo.

2.3 A Presença de Covariáveis

Além do tempo de sobrevivência e da presença de censura, outra característica peculiar nos dados de sobrevivência é a possível observação de variáveis que podem representar a heterogeneidade existente na população ou os tratamentos a que os indivíduos são submetidos. Essas variáveis são conhecidas como variáveis explicativas ou covariáveis.

Nos casos em que as covariáveis estão presentes, é possível verificar se as mesmas estão influenciando o evento de interesse ou até mesmo se a interação entre tratamentos e as covariáveis são significativas.

As covariáveis podem ou não depender do tempo, que surgem quando o estado do indivíduo muda durante um evento de interesse. Assim, uma covariável X não pode ser representada por um simples valor x_i para o paciente i, mas aceita valores que podem mudar com o tempo.

2.4 Descrição do Comportamento do Tempo de Sobrevivência

Para expressar o comportamento da variável aleatória contínua que representa o tempo de sobrevivência, $T \geqslant 0$, pode-se recorrer a várias funções matemáticas.

As funções que podem ser utilizadas para descrever os diferentes aspectos apresentados no conjunto de dados a ser observado são a função de densidade de probabilidade $f(t)$; a função de sobrevivência $S(t)$, a função de risco $h(t)$ e a média residual $mrl(t)$.

Dado que estas funções caracterizam o comportamento do tempo de sobrevivência, caso uma delas for especificada as demais podem ser encontradas devido às relações existentes entre elas.

2.4.1 A Função Densidade de Probabilidade

A função densidade de probabilidade é definida, para descrever o tempo de sobrevivência, como o limite da probabilidade de um indivíduo morrer no intervalo $[t, t + \Delta t)$ por unidade de tempo, e é expressa por Lee (1992, p.11)

$$f\left(t\right) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t\right)}{\Delta t}, \tag{2.1}$$

onde $f(t) \geq 0$ para todo $t > 0$.

A função de distribuição acumulada de T é definida por:

$$F(t) = P\left(T \leq t\right) = \int_0^t f\left(u\right) du, \tag{2.2}$$

que representa a probabilidade do indivíduo sobreviver até algum tempo t qualquer.

2.4.2 A Função de Sobrevivência

A função de sobrevivência é a probabilidade de um indivíduo ter sobrevivido além de um determinado tempo t, ou seja, apresenta a probabilidade do tempo de sobrevivência ser maior ou igual a t, e é representada por

$$S\left(t\right) = P\left(T \geq t\right), \tag{2.3}$$

em que $S\left(t\right)$ é uma função monótona contínua decrescente com $S\left(0\right) = 1$ e com $\lim\limits_{\Delta t \to \infty} S(t) = 0$.

Uma vez que T é uma variável aleatória contínua, $S(t)$ também será uma função contínua estritamente decrescente. Assim, a função de sobrevivência é o complemento da função de distribuição acumulada, ou seja

$$S(t) = 1 - F(t). \tag{2.4}$$

Segundo Collett (1994), a função de sobrevivência é usada para representar a probabilidade de um indivíduo sobreviver desde o tempo inicial do estudo para algum tempo além de t.

No contexto de equipamentos ou características de itens manufaturados, $S(t)$ é referida como sendo a função de confiabilidade.

2.4.3 A Função de Risco

A função de risco é definida como o limite da probabilidade de um indivíduo morrer no intervalo de tempo $[t, t + \Delta t]$, dado que o mesmo tenha sobrevivido até o tempo t. Matematicamente, a função de risco é dada por:

$$h\left(t\right) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T \leq t + \Delta t \mid T \geq t\right)}{\Delta t}, \tag{2.5}$$

que representa a probabilidade do indivíduo sobrevivente morrer no intervalo $[t,\ t+\Delta t]$.

De acordo com Klein e Moeschberger (1997), a função de risco também é conhecida como taxa de falha incondicional em confiabilidade, a força da mortalidade em demografia, a função de intensidade em processos estocásticos, a taxa de falha de época-específica em epidemiologia, a inversa da razão de *Mill's* em economia ou simplesmente como a função de risco.

A função de risco descreve como a probabilidade instantânea de falha (taxa de falha) modifica-se com o passar do tempo, por isso tem sido preferida para descrever o comportamento do tempo de sobrevivência, podendo ser constante, crescente, decrescente ou mesmo não monótona.

Para representar o risco de morte de pessoas na faixa etária de 20 a 30 anos, a função de risco constante é a mais apropriada, pois a principal causa de morte, nesta faixa etária são os acidentes. Já para representar pessoas com alguma doença crônica que não respondem a determinado tratamento a função de risco empregada é crescente. Enquanto que a função de risco decrescente representaria as pessoas que respondem ao tratamento.

Em termos das equações (2.1) e (2.4), a função de risco pode também ser definida por meio da expressão:

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.6}$$

descrevendo assim o relacionamento entre as três funções que são utilizadas para representar o comportamento do tempo de sobrevivência.

2.4.4 Vida Média Residual

Uma outra função que pode ser usada para caracterizar a variável aleatória tempo de sobrevivência é a vida média residual. Ela é definida condicional a um certo tempo de vida t, dada por

$$mrl\,(t) = E\,(T - t \mid T > t)\,. \tag{2.7}$$

Para indivíduos com idade t, esta função mede o tempo médio restante de vida.

A vida média residual é a área sob a curva de sobrevivência à direita do tempo t dividida por $S\,(t)$, ou seja

$$mrl\,(t) = \frac{1}{S(t)} \int_t^{\infty} (u - t)\, f(u) du = \frac{1}{S\,(t)} \int_t^{\infty} S\,(u)\, du. \tag{2.8}$$

Como $S(0) = 1$ a vida média residual $mrl(0)$ será a área total sob a curva de sobrevivência para $t = 0$.

2.5 Relações entre a Função de Sobrevivência e a Função de Risco

As funções de risco e de sobrevivência são funções de interesse central, pois fornecem, respectivamente, as probabilidades de sobrevivência (garantia do produto) e de morte (falha do produto) e, por isso, podem resumir os dados de análise de sobrevivência.

A probabilidade da variável aleatória associada com indivíduos de tempo de sobrevivência, T, está compreendida no intervalo t e Δt, condicional a T ser maior ou igual a um valor de t, escrita como:

$$P\,(t \leq T \leq t + \Delta t \mid T \geq t)\,. \tag{2.9}$$

Pelo fato das funções $f\,(t)$, $S(t)$ e $h(t)$ serem matematicamente equivalente, é possível descrever algumas relações básicas que podem ser utilizadas na obtenção destas, caso uma delas for especificada.

A função densidade de probabilidade é a derivada da função densidade de distribuição em relação a t, ou seja

$$f\left(t\right) = \frac{d}{dt}F\left(t\right). \tag{2.10}$$

Uma vez que $F(t) = 1 - S(t)$, temos

$$f(t) = \frac{d}{dt}\left[1 - S(t)\right] = -S^{'}(t). \tag{2.11}$$

Substituindo (2.11) em (2.6), obtemos

$$h(t) = -\frac{S^{'}(t)}{S(t)} = -\frac{d}{dt}\left[\log S(t)\right]. \tag{2.12}$$

Dessa forma,

$$\log S(t) = -\int_0^t h(u)du, \tag{2.13}$$

e uma vez que $S(0) = 1$, segue que

$$S(t) = \exp\left(-\int_0^t h(u)du\right). \tag{2.14}$$

A função de risco dada pela equação(2.6), também pode ser utilizada na obtenção da função de risco acumulada, ou seja

$$H(t) = \int_0^t h(u)du. \tag{2.15}$$

Substituindo (2.15) em (2.14), temos

$$S(t) = \exp\left[-H(t)\right]. \tag{2.16}$$

Como, de (2.4), $\lim_{\Delta t \to \infty} S(t) = 0$, então

$$\lim_{t \to \infty} H\left(t\right) = \infty. \tag{2.17}$$

Também, como de (2.6) segue que

$$f(t) = h(t)S(t), \tag{2.18}$$

e substituindo (2.16) em (2.18) e usando (2.16), temos

$$f(t) = h(t) \exp\left(-\int_0^t h(u)du\right). \tag{2.19}$$

A expressão (2.19) é muito importante quando se deseja desenvolver procedimentos de estimação por meio da função de risco.

2.6 Estimador Não Paramétrico

Para estimar as funções de sobrevivência e de risco a partir de um conjunto de dados amostrais, na presença de observações censuradas, pode-se utilizar o estimador não paramétrico de Kaplan-Meier, também conhecido na literatura como estimador produto-limite, que foi proposto por Kaplan e Meier (1958).

O estimador de Kaplan-Meier é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como

$$\hat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo } t}{\text{nº total de observações no estudo}}. \tag{2.20}$$

$\hat{S}(t)$ é a uma função com degraus nos tempos observados de falha de tamanho $1/n$, onde n é o tamaho da amostra. E caso existam empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates.

Considerando um estudo envolvendo n indivíduos, em que os tempos de sobrevivência, incluindo as censuras, são ordenados, então o estimador Kaplan-Meier da função de sobrevivência empírica é dado por

$$\begin{aligned} \hat{S}_{KM}(t) &= \frac{n_1 - d_1}{n_1}\frac{n_2 - d_2}{n_2}\cdots\frac{n_r - d_r}{n_r} \\ &= \prod_{r;t_r<t}\frac{n_i - d_i}{n_i}, \end{aligned} \tag{2.21}$$

onde:

t_r é o maior tempo de sobrevivência menor ou igual a t;

n_i é o número de observações sob risco até o tempo t_i (inclusive);

d_i representa o número de mortes no tempo t_i ($d_i = 0$ para tempos de sobrevivência censurados) e $1 < i < r$.

Segundo Louzada, Mazucheli e Achar (2000), o estimador de Kaplan-Meier da função de risco no intervalo de tempo $[t, t+u)$ é dado por

$$\hat{h}_{KM}(t) = \frac{1}{u}\left(1 - \frac{n_i - d_i}{n_i}\right).$$

2.7 Modelos Probabilísticos mais Usados

Devido à caracterização da variável aleatória T, representando o comportamento do tempo de sobrevivência, serão apresentadas várias distribuições de probabilidade que são usadas para modelar, de forma acessível, dados relacionados ao tempo de sobrevivência.

2.7.1 Distribuição Exponencial

A distribuição exponencial é utilizada extensivamente como um modelo para o tempo de vida de certos produtos e materiais. Em parte por ser um dos modelos mais simplificados, mas principalmente por ser caracterizada por uma função de risco constante.

Muitas vezes pode ocorrer a redução da aplicação do modelo de probabilidade exponencial em situações realistas, devido o mesmo ter a característica de falta de memória, isto é, a função de risco não depende do tempo, é constante ao longo do mesmo.

Uma variável aleatória não negativa e contínua T tem distribuição exponencial com parâmetro $\mu > 0$ se sua função de densidade de probabilidade é dada por

$$f(t) = \frac{1}{\mu} \exp\left(-\frac{t}{\mu}\right), \qquad t > 0, \tag{2.22}$$

em que μ representa a média (com a mesma unidade de tempo de T). A variância de T é dada por μ^2.

A função de sobrevivência e o percentil $t_p = 100(1-\alpha)\%$ (quanto se quer obter informações a respeito de falhas ou mortes prematuras) são dados respectivamente, por

$$S(t) = \exp\left(-\frac{t}{\mu}\right), \tag{2.23}$$

$$t_p = -\mu \log(1-p). \tag{2.24}$$

A partir das equações (2.22) e (2.23), tem se a função de risco, dada por

$$h(t) = \frac{1}{\mu}. \tag{2.25}$$

Corroborando com a afirmação de que a função de risco do modelo exponencial é constante para todo t.

Em muitas situações práticas em que a morte de um indivíduo ou a falha de um equipamento eletrônico, que são assumidas como aleatórias no tempo e independentemente da idade do mesmo, a distribuição mais adequada é a exponencial.

A distribuição do tempo de sobrevivência adicional não é afetada pela informação de que o mesmo sobreviveu anteriormente a algum tempo, pois sua função de risco não depende do tempo, sendo constante ao longo do mesmo. Sendo assim, se supor o tempo de sobrevivência com distribuição de probabilidade exponencial, um equipamento novo (recém fabricado) e um equipamento com muitas horas de uso têm a mesma probabilidade de falhar em qualquer instante t. Porém, na prática, o que podemos verificar é que envelhecimento pode influenciar na forma da função de risco, ou seja, um equipamento novo pode apresentar um risco de falha que cresce com o tempo de uso, por exemplo, um televisor, ou um computador, ou um eletrodoméstico novo pode ter um risco de falha que cresce com o tempo de uso.

2.7.2 Distribuição Weibull

Outra distribuição de probabilidade muito utilizada para descrever o tempo de vida de produtos industriais e de dados biomédicos é a distribuição Weibull, que foi proposta por Weibull em 1951, que estudou a relação do tempo de falha devido a fadiga de metais.

A distribuição de probabilidade Weibull apresenta uma grande variedade de formas, porém todas com a mesma propriedade básica: função de risco monótona, isto é, ou ela é crescente, ou decrescente, ou constante, sendo a razão de ser popularmente utilizada em situações práticas.

A função densidade de probabilidade da distribuição Weibull, especificada

a partir de dois parâmetros, é dada por

$$f(t) = \frac{\beta}{\mu}\left(\frac{t}{\mu}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right], \tag{2.26}$$

com $t > 0$, e onde $\beta > 0$ e $\mu > 0$ são os parâmetros de forma e escala, respectivamente. Em (2.26) quando $\beta = 1$ tem-se a distribuição exponencial. Assim, a distribuição exponencial é um caso particular da distribuição Weibull.

A função de risco, de sobrevivência e os percentis da distribuição Weibull são dados, respectivamente, por

$$h(t) = \frac{\beta}{\mu}\left(\frac{t}{\mu}\right)^{\beta-1}, \tag{2.27}$$

$$S(t) = \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right], \tag{2.28}$$

$$t_p = \mu\left[-\log(1-p)\right]^{\frac{1}{\beta}}. \tag{2.29}$$

O parâmetro β indica a forma da curva. Quando $\beta > 1$, a função de risco $h(t)$ é estritamente crescente; quando $\beta < 1$, a função de risco é estritamente decrescente; e para $\beta = 1$, a função de risco é constante.

Uma distribuição bastante relacionada com a distribuição Weibull é a chamada distribuição de valor extremo que surge quando se toma o logaritmo de uma variável com distribuição Weibull. Isto é, para uma variável aleatória T com distribuição Weibull dada por (2.26), a variável aleatória $Y = \log(T)$ tem distribuição de valor extremo, cuja função densidade de probabilidade é dada por

$$f(y) = \frac{1}{\sigma}\exp\left[\frac{y-\alpha}{\sigma} - \exp\left(\frac{y-\alpha}{\sigma}\right)\right], \tag{2.30}$$

em que $\alpha = \log(\mu)$ e $\sigma = \frac{1}{\beta}$.

A função de sobrevivência e os percentis da variável Y são dadas, respectivamente, por

$$S(y) = \exp\left[-\exp\left(\frac{y-\alpha}{\sigma}\right)\right], \tag{2.31}$$

$$t_p = \alpha + \sigma \log\left[-(1-p)\right], \tag{2.32}$$

em que α representa o parâmetro de locação e σ o parâmetro de escala.

2.7.3 Distribuição Log-Normal

A distribuição log-normal, assim como a distribuição Weibull, é extensamente utilizada na modelagem de dados de sobrevivência, afim de caracterizar o tempo de vida de produtos e materiais, como isolantes elétricos, fadiga de materiais, etc. Nelson e Haln, 1972 (apud Lawless, 1982), ilustram a utilização da distribuição log-normal nas situações de mergulhadores, como a análise de falha de isolantes elétricos e o estudo de tempos até o aparecimento de câncer pulmonar em fumantes.

A função de densidade de probabilidade de uma variável aleatória com distribuição log-normal é dada por (Borges, Colosimo e Freitas, 1996)

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left\{-\frac{\left[\log(t) - \mu\right]^2}{2\sigma^2}\right\}, \tag{2.33}$$

em que μ é a média do logaritmo do tempo de falha e σ é o desvio padrão.

Assim como a relação existente entre a distribuição Weibull e de valor extremo, existe uma relação direta entre a distribuição log-normal e a normal, que facilita a apresentação e análise de dados provenientes da distribuição log-normal. O logaritmo de uma variável aleatória com distribuição log-normal com parâmetros μ e σ tem distribuição normal com média μ e desvio padrão

σ.

Dessa forma, pode-se concluir que dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, utilizando o logaritmo dos dados, ao invés de utilizar os dados em sua escala original.

A função de sobrevivência de uma variável aleatória log-normal é escrita em termos da distribuição normal padrão e é definida por

$$S(t) = 1 - \int_{-\infty}^{\frac{\log - \mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx \tag{2.34}$$

$$= 1 - \phi\left(\frac{\log t - \mu}{\sigma}\right), \tag{2.35}$$

em que $\phi(\cdot)$ é a função de distribuição acumulada normal padrão com média 0 e variância 1.

A distribuição log-normal não apresenta funções de sobrevivência e de risco monótonas como a distribuição Weibull. As funções de risco de uma distribuição log-normal crescem, atingem um valor máximo e depois decrescem, apresentando assim função de risco unimodal.

Os percentis da distribuição log-normal podem ser obtidos a partir da tabela da distribuição normal padrão por meio da expressão

$$t_p = \exp(Z_p\sigma + \mu). \tag{2.36}$$

onde Z_p é o $100(1-\alpha)\%$ percentil da distribuição normal padrão.

2.7.4 Distribuição Log-Logística

Outra distribuição de sobrevivência muito utilizada é a distribuição log-logística. Uma característica da distribuição log-logística é que sua função de risco apresenta forma unimodal, isto é, sua função de risco começa crescente

e depois muda para decrescente, ou vice versa.

Uma variável aleatória $T > 0$ tem distribuição log-logística de parâmetros $\beta > 0$ e $\mu > 0$, se seu logaritmo, $Y = \log(T)$, tem distribuição logística com função densidade dada por

$$f(y) = \frac{\exp\left(\frac{y-\alpha}{\sigma}\right)}{\sigma\left[1+\exp\left(\frac{y-\alpha}{\sigma}\right)\right]^2}, \tag{2.37}$$

em que -$\infty < \alpha < \infty$ e $\sigma > 0$ são os parâmetros de locação e de escala, respectivamente.

As funções de sobrevivência, de risco e os percentis de uma distribuição log-logística são dados, respectivamente, por

$$S(t) = \frac{1}{\left[1+\left(\frac{t}{\mu}\right)^{\beta}\right]}, \tag{2.38}$$

$$h(t) = \frac{\beta\left(\frac{t}{\mu}\right)^{\beta-1}}{\mu\left[1+\left(\frac{t}{\mu}\right)^{\beta}\right]}, \tag{2.39}$$

$$t_p = \mu\left[\frac{p}{(1-p)}\right]^{\frac{1}{\beta}}. \tag{2.40}$$

Apesar da função de risco da distribuição log-logística apresentar comportamento unimodal, a mesma tem vantagem em relação a distribuição log-normal, pois apresenta expressões explícitas para as funções de sobrevivência e de risco.

2.7.5 Distribuição Gama

Uma variável aleatória não negativa T com distribuição gama tem função de densidade dada por (Lawless, 1982)

$$f(t) = \frac{\mu(\mu t)^{\beta-1} e^{-\mu t}}{\Gamma(\beta)}, \tag{2.41}$$

em que os parâmetros $\mu > 0$ e $\beta > 0$, μ é o parâmetro de escala, β é o parâmetro de forma e $\Gamma(\cdot)$ é a função gama incompleta.

Assim como a distribuição Weibull, a distribuição gama também inclui a distribuição exponencial como caso particular quando $\beta = 1$.

As funções de sobrevivência e de risco de uma distribuição gama envolvem a integral gama incompleta

$$I(\beta, t) = \frac{1}{\Gamma(\beta)} \int_0^t u^{\beta-1} e^{-u} du. \tag{2.42}$$

Assim, integrando a função de densidade, tem-se que a função de sobrevivência, dada por

$$S(t) = 1 - I(\beta, \mu t). \tag{2.43}$$

Quando $\beta > 1$ a função de risco é monótona crescente com $h(0) = 0$ e $\lim_{t\to\infty} h(t) = \mu$. Se $0 \leq \beta \leq 1$, então a função de risco é monótona decrescente com $\lim_{t\to\infty} h(t) = \infty$ e $\lim_{\Delta t\to\infty} h(t) = \mu$.

Quando $\mu = 1$ a distribuição é conhecida como a distribuição gama de um parâmetro e sua função de densidade é dada por

$$f(t) = \frac{t^{\beta-1} e^{-t}}{\Gamma(\beta)}. \tag{2.44}$$

A distribuição gama de um parâmetro está relacionada com a distribuição qui-quadrado com graus de liberdade igual 2 vezes o parâmetro de escala.

A distribuição gama, apesar de ajustar uma larga variedade de dados de sobrevivência, não é tão utilizada para a modelagem de dados de sobrevivência como a distribuição Weibull, pois suas funções de sobrevivência e de risco não possuem expressões de forma fechada e simples.

Da mesma maneira que existe uma relação entre a distribuição Weibull e a distribuição valor extremo, pode-se observar algumas relações sobre a distribuição do logaritmo do tempo de vida, segundo uma distribuição gama.

Seja a função densidade de probabilidade de uma distribuição gama dada por (2.41), então fazendo $\varpi = \log(\mu t) = \log \mu + \log t$, temos a chamada distribuição log-gama, que possui função densidade de probabilidade dada por

$$f(t) = \frac{1}{\Gamma(\beta)} \exp\left(\beta\varpi - e^{\varpi}\right), \tag{2.45}$$

em que -$\infty < \varpi < \infty$.

As distribuições de probabilidade da distribuição log-gama são inviesadas negativamente, decrescendo a medida que o valor de β cresce. Quando $\beta = 1$ a distribuição se reduz a uma distribuição valor extremo.

2.7.6 Distribuição Gama Generalizada

A distribuição gama generalizada é uma distribuição com três parâmetros e sua função de densidade de probabilidade é dada por

$$f(t) = \frac{\mu\lambda}{\Gamma(\beta)} (\mu t)^{\beta\lambda - 1} \exp\left[-(\mu t)^{\lambda}\right], \tag{2.46}$$

com $t > 0$, e os parâmetros λ, β, e μ são todos positivos.

A distribuição gama generalizada inclui casos especiais, tais como se $\lambda = \beta = 1$, tem-se uma distribuição exponencial; para $\beta = 1$, tem-se uma distribuição Weibull; para $\lambda = 1$, tem-se a distribuição gama e para $\beta \to \infty$

torna-se a distribuição log-normal, como mostra Lawless (1982).

As funções de sobrevivência e de risco de uma distribuição gama generalizada também envolvem a função gama incompleta. Dessa maneira, tem-se que a função de sobrevivência é dada por

$$S(t) = 1 - I\left[\beta, (\mu t)^{\lambda}\right], \tag{2.47}$$

e a função de risco é dada pela relação $h(t) = f(t)/S(t)$.

A distribuição gama generalizada é uma família de distribuições flexíveis com três parâmetros, sendo muito útil na substituição de modelos alternativos como Weibull e log-normal.

2.8 Estimação Via Máxima Verossimilhança

Apesar de existirem técnicas não paramétricas para tratar de dados relativos ao tempo de vida, nesta pesquisa o foco é ajustar modelos paramétricos devido estes serem mais consistentes no cálculo das funções de sobrevivência e de risco.

Parâmetros são quantidades (valores) desconhecidas dos modelos probabilísticos. Porém, em cada estudo, os parâmetros devem ser estimados a partir de observações amostrais.

Dos modelos de probabilidades descritos até agora neste trabalho, tem-se que o modelo gama generalizado é caracterizado por três parâmetros; os modelos gama, log-logística, log-normal e Weibull são caracterizados por dois parâmetros e o modelo exponencial por um parâmetro.

Existem vários métodos de estimação, mas para os modelos de análise de sobrevivência o método mais utilizado é o de máxima verossimilhança, devido ao fato do mesmo ser capaz de incorporar dados censurados, além de possuir propriedades ótimas para amostras grandes.

O método de máxima verossimilhança consiste em adotar para o parâmetro o valor que maximiza a função de verossimilhança correspondente ao resultado

obtido através da amostra. Como afirma Cordeiro:

> "O método de máxima verossimilhança (MV) não contradiz os fatos (representados pelos dados) que nós realmente observamos e objetiva escolher o valor do parâmetro (ou a hipótese no sentido mais amplo) que dá a chance mais provável para os fatos que ocorreram novamente ocorram". (Cordeiro, 1992: p. 4 e 5).

Suponha uma amostra aleatória de observações, $t_1, t_2, ..., t_n$, da variável aleatória T, de uma população com função densidade $f(t;\theta)$, com $\theta \in \Theta$, onde Θ é o espaço paramétrico. A função de verossimilhança de θ, correspondente a amostra aleatória observada, é dada por

$$L(t_1, t_2, ..., t_n; \theta) = L(\theta) = \prod_{i=1}^{n} f(t_i; \theta). \tag{2.48}$$

O estimador de máxima verossimilhança de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta)$.

Considerando, primeiramente, todas as observações não censuradas da amostra observada, a função de verossimilhança $L(\theta)$ mostra que a contribuição de cada observação não censurada, ou seja observação completa, é a função de densidade $f(t)$ do modelo probabilístico.

Porém, o mesmo não ocorre quando as observações são censuradas. As observações censuradas informam que o tempo de não sobrevivência ou falha é maior que o tempo de censura observado, e portanto, que a sua contribuição para $L(\theta)$ é a função de sobrevivência $S(t)$ do modelo probabilístico.

Assim, as observações da amostra aleatória podem ser divididas em dois conjuntos, as r primeiras são as não censuradas $(1, 2, ..., r)$, e as $(n - r)$ seguintes são as censuradas $(r + 1, r + 2, ..., n)$.

Dessa forma, a função de verossimilhança, segundo Lawless (1982), assume

a seguinte forma

$$L(\theta) = \prod_{i=1}^{r} f(t_i;\theta) \prod_{i=r+1}^{n} S(t_i;\theta). \tag{2.49}$$

A expressão (2.49) é válida para as censuras do tipo I, II e aleatória e também sob a suposição que o mecanismo de censura é não informativo.

Dependendo do modelo probabilístico, θ pode estar representando um único parâmetro ou um conjunto de parâmetros.

Para que se tenha uma notação mais prática, pode-se introduzir uma variável indicadora de censura δ_i, onde $\delta_i = 1$ se t_i é observado diretamente, ou seja, não tem censura ou é censura do tipo II; e $\delta_i = 0$ se t_i é censurado à direita, dessa forma, a função de verossimilhança é escrita como

$$\begin{aligned} L(\theta) &= \prod_{\delta_i=1} f(t_i;\theta) \prod_{\delta_i=0} S(t_i;\theta) \\ &= \prod_{\delta_i=1} h(t_i;\theta) S(t_i;\theta) \prod_{\delta_i=0} S(t_i;\theta) \\ &= \prod_{i=1}^{n} h(t_i;\theta)^{\delta_i} S(t_i;\theta). \end{aligned} \tag{2.50}$$

É sempre conveniente trabalhar com o logaritmo da função de verossimilhança, em que os estimadores de θ que maximizam $L(\theta)$ são equivalentes aos que maximizam $\log[L(\theta)]$.

Os estimadores de máxima verossimilhança são obtidos resolvendo o sistema de equações

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = 0. \tag{2.51}$$

É importante lembrar que tem que ser definido primeiramente o modelo probabilístico adequado para os dados para depois utilizar-se o método de máxima verossimilhança.

2.8.1 Estimação do Parâmetro para o Modelo Exponencial

Para um melhor entendimento da estimação do parâmetro para o modelo exponencial, a mesma será dividida em três partes: observações sem censura, observações com censura do tipo II e observações com censura do tipo I.

Inicialmente, considere que $t_1, t_2, ..., t_n$ é uma amostra aleatória dos tempos de vida, ou seja, não existe censura, para os quais se supõe a distribuição exponencial, cuja função de densidade de probabilidade é dada por

$$f(t_i;\mu) = \frac{1}{\mu}\exp\left(-\frac{t_i}{\mu}\right),$$

com média μ e função de risco $h(t) = \frac{1}{\mu}$.

Assim, a função de verossimilhança é dada por

$$L(\mu) = \prod_{i=1}^{n} f(t_i;\mu) = \frac{1}{\mu^n}\exp\left(-\sum_{i=1}^{n}\frac{t_i}{\mu}\right). \tag{2.52}$$

O estimador de máxima verossimilhança de μ, neste caso, é obtido maximizando (2.52). Este processo leva ao estimador $\hat{\mu} = \frac{T^*}{n}$, e sendo $T^* = \sum_{i=1}^{n} t_i$ então $\hat{\mu}= \bar{t}$ (tempo médio de vida).

A segunda parte da estimação refere-se a amostras com censura do tipo II, onde somente as r primeiras observações $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)}$ são analisadas numa amostra de tamanho n, sendo a função densidade de probabilidade conjunta das r primeiras estatísticas de ordem dada por

$$f(t_i;\mu) = \frac{n!}{(n-r)!} f(t_1)\cdots f(t_r)\left[S(t_r)\right],$$

então a função de verossimilhança é dada por

$$L(\mu) = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r}\frac{1}{\mu}\exp\left(-\frac{t_i}{\mu}\right)\right]\left[\exp\left(-\frac{t_r}{\mu}\right)^{n-r}\right] \tag{2.53}$$

$$= \frac{n!}{(n-r)!}\frac{1}{\mu}\exp\left[\frac{1}{\mu} - \left(\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}\right)\right]. \tag{2.54}$$

Seja agora T^*

$$T^* = \sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}. \tag{2.55}$$

Como o objetivo é realizar uma maximização, o termo constante pode ser excluído, pois isto não altera o resultado do mesmo, assim de acordo com a expressão (2.54), a função de verossimilhança é dada por

$$L(\mu) = \frac{1}{\mu^r}\exp\left(-\frac{T^*}{\mu}\right), \tag{2.56}$$

em que o estimador de máxima verossimilhança de μ é $\hat{\mu} = \frac{T^*}{r}$

Na primeira parte T^* é referido como "o tempo total observado," pois é uma amostra sem censura. Já na segunda parte, onde a amostra é com censura do tipo II, T^* é "tempo total em teste", porém, tanto na primeira parte como na segunda, T^* é "o tempo total observado" para todos os indivíduos da amostra.

Finalizando, considere uma amostra aleatória de n indivíduos com tempo de vida $T_1, ..., T_n$, em que a cada indivíduo está associado também um tempo limite de censura $L_i > 0$. Dessa forma, para valores de $1 \leq i \leq n$, cada T_i consiste de um par (t_i, δ_i), onde $t_i = \min(T_i, L_i)$ e o indicador de censura será $\delta_i = 1$ se $t_i = T_i$, e $\delta_i = 0$ se $t_i = L_i$.

Assim, para uma amostra de dados com censura tipo I, sua função de

verossimilhança para a distribuição exponencial será

$$\begin{aligned} L(\mu) &= \left(\prod_{i=1}^{n} \frac{1}{\mu} \exp -\frac{T_i L_i}{\delta_i}\right) \exp\left(-\frac{L_i(1-\delta_i)}{\mu}\right) \\ &= \frac{1}{\mu^r} \exp\left(-\sum_{i=1}^{n} \frac{t_i}{\mu}\right), \end{aligned} \tag{2.57}$$

onde $r = \sum_{i=1}^{n} \delta_i$ é a quantidade de observações completas. Fazendo com que T^* agora seja

$$T^* = \sum_{i=1}^{n} t_i = \sum_{i \in D} T_i + \sum_{i \in C} L_i, \tag{2.58}$$

em que T é o total de tempo observado para n indivíduos; D é o conjunto de índices para os indivíduos com tempo observado (isto é, para amostra sem censura ou com censura do tipo II); e C é o conjunto de índices para os tempos censurados (amostra com censura do tipo I).

Da equação (2.57) é possível obter o estimador de máxima verossimilhança para μ, assumindo que $r > 0$, o qual dado por $\hat{\mu} = \frac{T^*}{r}$.

Caso $r = 0$, a função de verossimilhança é monótona decrescente, aproximando-se de 1 quando $\mu \to \infty$, e assim não possui um máximo finito.

Devido ao fato de r ser aleatório, o procedimento de estimação torna-se uma tarefa não muito fácil. Assim, faz-se necessário considerar as propriedades dos estimadores de máxima verossimilhança baseado em amostras grandes, ou seja, aplicar a primeira e a segunda derivada na função log-verossimilhança

De uma maneira geral, considerando uma variável aleatória T, sendo com censura do tipo I, do tipo II ou sem censura, tem-se que a função de verossimilhança para o modelo exponencial fica expressa na forma

$$
\begin{aligned}
L(\mu) &= \prod_{\delta_i=1} \frac{1}{\mu} \exp\left(-\frac{t_i}{\mu}\right) \prod_{\delta_i=0} \exp\left(-\frac{t_i}{\mu}\right) \\
&= \prod_{i=1}^{n} \left(\frac{1}{\mu}\right)^{\delta_i} \exp\left(-\frac{t_i}{\mu}\right) \\
&= \mu^{-\delta_i} \exp\left(-\frac{t_i}{\mu}\right)
\end{aligned} \tag{2.59}
$$

Aplicando o logaritmo em(2.59), tem-se

$$
\log\left[L(\mu)\right] = -\sum_{i=1}^{n} \delta_i \log(\mu) - \mu \sum_{i=1}^{n} t_i. \tag{2.60}
$$

Derivando (2.60) com relação a μ e igualando a zero, tem-se como solução o estimador de máxima verossimilhança de μ, dado por

$$
\hat{\mu} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} \delta_i}. \tag{2.61}
$$

O método de máxima verossimilhança também permite a construção de intervalos de confiança para os parâmetros, o que é feito a partir das propriedades para amostras grandes Cox e Hinkley (1974) e Cordeiro (1992) .

A partir de $Var(\hat{\mu}) \cong \left(\frac{partial^2 \log L(\mu)}{partial \mu}\right)^{-1}$, tem-se que $Var(\hat{\mu}) \cong \frac{\mu^2}{\sum_{i=1}^{n} \delta_i}$. Utilizando o fato de que a quantidade $\frac{2r\hat{\mu}}{\mu}$ tem distribuição qui-quadrado com $2r$ graus de liberdade, como afirma Lawless (1982), um intervalo de confiança de $100\ (1-\alpha)\%$ para μ é obtido a partir de

$$
P\left(\chi^2_{(2n;\ \alpha/2)} \leq \frac{2r\hat{\mu}}{\mu} \leq \chi^2_{(2n;\ 1-\alpha/2)}\right) = 1 - \alpha \tag{2.62}
$$

onde $\chi^2_{(2n;\,p)}$ é o p-ésimo percentil da distribuição qui-quadrado com $2r$ graus de liberdade. Assim, o intervalo de confiança de $100\ (1-\alpha)\%$ para μ é dado por

$$\frac{2r\hat{\mu}}{\chi^2_{(2n;\,\alpha/2)}} \leq \mu \leq \frac{2r\hat{\mu}}{\chi^2_{(2n;\,1-\alpha/2)}}. \tag{2.63}$$

Na ausência de observações censuradas $(r=n)$, $\frac{2n\hat{\mu}}{\mu} \sim \chi^2_{(2n)}$.

A partir de grandes amostras é possível utilizar a normalidade assintótica dos estimadores de máxima verossimilhança para a obtenção do intervalo de confiança para o parâmetro desejado.

Em Lawless (1982, p108) são apresentadas outras aproximações assintóticas que podem ser consideradas no processo de estimação e obtenção de intervalo de confiança para μ.

2.8.2 Estimação dos Parâmetros para o Modelo Weibull

Suponha que $t_1 \leq \cdots \leq t_n$ são r observações de uma amostra aleatória sem censura ou com censura do tipo II de tamanho n, proveniente de uma distribuição Weibull com parâmetros μ e β , cuja a função densidade de probabilidade é dada por

$$f\left(t_i\right) = \frac{\beta}{\mu}\left(\frac{t_i}{\mu}\right)^{\beta-1} \exp\left[-\left(\frac{t_i}{\mu}\right)^{\beta}\right], \tag{2.64}$$

ou equivalentemente, $x_1 \leq \cdots \leq x_r$, para $x_i = \log t_{(i)}$ são as r observações de uma amostra de tamanho n de uma distribuição de valor extremo dada por

$$f\left(x\right) = \frac{1}{b} e^{\frac{(x-v)}{b}} \exp\left[-e^{\frac{(x-v)}{b}}\right], \tag{2.65}$$

em que $v = \log\left(\mu\right)$, é o parâmetro de locação, $-\infty < v < \infty$ e $b = \beta^{-1}$ é o parâmetro de escala, $b > 0$.

A função densidade conjunta de probabilidade de $x_1, ..., x_r$ é dada por

$$\frac{n!}{(n-r)!} f(x_1) \cdots f(x_r) \left[S(x_r)\right]^{n-r},$$

e a função de verossimilhança para v e b pode ser escrita como

$$L(v,b) = \frac{n!}{(n-r)!} \left[\prod_{i=1}^{r} \frac{1}{b} e^{\frac{(x_i-v)}{b}} \exp\left(-e^{\frac{(x_i-v)}{b}}\right)\right] \left[\exp\left(-e^{\frac{(x_r-v)}{b}}\right)\right]^{n-r} \tag{2.66}$$

$$= \frac{1}{b^r} \exp\left(\sum_{i=1}^{r} \frac{x_i - v}{b} - \sum_{i=1}^{r} \exp \frac{x_i - v}{b}\right). \tag{2.67}$$

O logaritmo da função verossimilhança é

$$\log L(v,b) = -r \log b + \sum_{i=1}^{r} \frac{x_i - v}{b} - \sum_{i=1}^{r} \exp\left(\frac{x_i - v}{b}\right). \tag{2.68}$$

Derivando a equação (2.68) em relação ao parâmetro v e em relação ao parâmetro b, e resolvendo simultaneamente $\frac{\partial \log L}{\partial v} = 0$ e $\frac{\partial \log L}{\partial b} = 0$, é possível encontrar os estimadores de máxima verossimilhança para v e b. Assim, temos as equações

$$e^{\hat{v}} = \left[\frac{1}{r} \sum_{i=1}^{r} \exp\left(\frac{x_i}{\hat{b}}\right)\right]^{\hat{b}}, \tag{2.69}$$

e

$$\frac{\sum_{i=1}^{r} x_i \exp\left(\frac{x_i}{\hat{b}}\right)}{\sum_{i=1}^{r} \exp\left(\frac{x_i}{\hat{b}}\right) - \hat{b} - \frac{1}{r}\sum_{i=1}^{r} x_i} = 0. \tag{2.70}$$

Para encontrar os estimadores $\hat{v}$ e $\hat{b}$, primeiramente, deve-se determinar $\hat{b}$ resolvendo a equação (2.70) , utilizando-se de um método numérico. O método mais apropriado é o de Newton-Raphson, que utiliza a matriz de derivadas segundas da função log verossimilhança e é baseado na expansão em série de Taylor, que parte de um valor inicial, usualmente zero, e atualiza-o a cada iteração. Geralmente a convergência se dá em poucas iterações e, encontrado o estimador $\hat{b}$, obtêm-se o estimador $\hat{v}$ a partir da equação (2.69).

Então, encontrados $\hat{v}$ e $\hat{b}$, os estimadores de máxima verossimilhança dos parâmetros de uma distribuição Weibull são:

$$\hat{\mu} = \exp(\hat{v}) \quad \text{e} \quad \hat{\beta} = \hat{b}^{-1} \tag{2.71}$$

Usando as equações (2.69) e(2.70) , tem-se que

$$\hat{\mu} = \left(\frac{1}{r} \sum_{i=1}^{r} t_i^{\hat{\beta}} \right)^{1/\hat{\beta}}, \tag{2.72}$$

$$\frac{\sum_{i=1}^{r} t_i^{\hat{\beta}} \log t_i}{\sum_{i=1}^{r} t_i \hat{\beta} - \frac{1}{\hat{\beta}} - \frac{1}{r} \sum_{i=1}^{r} \log t_i} = 0. \tag{2.73}$$

Considere uma amostra com censura do tipo I (censura a direita), onde T_i representa o tempo de vida e L_i o tempo de censura fixado numa amostra de n indivíduos. Sendo $t_i = \min(T_i, L_i)$ se a observação é um tempo de vida ou um tempo censurado.

Considerando os $T_i^{'}s$ provenientes da distribuição Weibull ou, equivalentemente, $X_i = \log T_i$, seja $\eta_i = \log L_i$, $x_i = \log t_{(i)}$ e o indicador de censura δ_i, onde $\delta_i = 1$ se $t_i = T_i$ ou $\delta_i = 0$ caso $t_i = L_i$. Então, a função de

verossimilhança da distribuição de valor extremo é dada por

$$L(v,b) = \prod_{i=1}^{n} \left[\frac{1}{b}\exp\left(\frac{x_i - v}{b} - e^{\frac{(x_i - v)}{b}}\right)\right]^{\delta_i} \left[\exp\left(-e^{\frac{(\eta_i - v)}{b}}\right)\right]^{1-\delta_i}. \quad (2.74)$$

Tomando $r = \sum_{i=1}^{n} \delta_i$ como sendo o de tempo de vida observado e D o conjunto formado pelos indivíduos com tempo de vida sem censura, ou seja, $\delta_i = 1$, tem-se

$$\log L(v,b) = -r\log b + \sum_{i\in D}\left(\frac{x_i - v}{b}\right) - \sum_{i=1}^{n}\exp\left(\frac{x_i - v}{b}\right). \quad (2.75)$$

Observe que o logaritmo da função de verossimilhança da distribuição de valor extremo tem a mesma forma apresentada no caso de censura do tipo II, desde que X_n tenha o tempo censurado para os $(n-r)$ indivíduos que não são observados.

Assim, os estimadores de máxima verossimilhança para a distribuição de valor extremo são conseqüentemente da mesma forma como apresentados em (2.69) e (2.70) e, para $r > 0$, podem ser escritos respectivamente como (2.72) e (2.73).

A estimação dos parâmetros para os modelos de distribuição log-normal, log-logística, gama e gama generalizada é feita pelo método de máxima verossimilhança de maneira análoga, como mostra Lawless (1982).

2.9 Estimadores de Máxima Verossimilhança para Amostras Grandes

Suponha que n tempos de sobrevivência observados, $t_1, t_2, \cdots, t_n$, foram ajustado por um modelo de probabilidade para estimar os valores de p parâmetros desconhecidos $\beta_1, \beta_2, \ldots \beta_p$, tal que sua função de verossimilhança é dada por $L(\beta)$. Os estimadores de máxima verossimilhança do vetor de parâmetros p são os valores $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p$ que maximizam $L(\beta)$ e são encontrados resolvendo

a derivada de primeira ordem do logaritmo da função de verossimilhança, ou seja

$$\frac{d \log L(\beta)}{d\beta_j} \tag{2.76}$$

em que $j = 1, 2, \ldots, p$.

Calculando a segunda derivada parcial da função de log verossimilhança $\log L\left(\hat{\beta}\right)$, tem-se a matriz $p \times p$, denotada por $H(\beta)$, chamada de matriz hessiana, onde o elemento (j,k)-ésimo de $H(\beta)$ é

$$\frac{\partial^2 \log L\left(\hat{\beta}\right)}{\partial\beta_j \partial\beta_k}, \tag{2.77}$$

para $j = 1, 2, \ldots, p$.

A matriz $I(\beta) = -H(\beta)$ é conhecida como a matriz de informação observada, onde o (j,k)-ésimo elemento correspondente à matriz de informação esperada é

$$-E\left(\frac{\partial^2 \log L(\beta)}{\partial\beta_j \partial\beta_k}\right). \tag{2.78}$$

A matriz de covariância p, dos estimadores de máxima verossimilhança $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$, escrita como $Var\left(\hat{\beta}\right)$, pode ser aproximada pela inversa da matriz de informação observada, calculada por $\hat{\beta}$, isto é

$$Var\left(\hat{\beta}\right) \approx I^{-1}\left(\hat{\beta}\right). \tag{2.79}$$

2.9.1 Estimadores de Máxima Verossimilhança para Amostras Grandes, Modelo Weibull

Sendo a contribuição da log verossimilhança de uma única observação de uma distribuição de valor extremo

$$\log L = \delta_i \left[Z_i - \log b - \exp Z_i \right] + (1 - \delta_i) \left[- \exp \left(c_i \right) \right], \tag{2.80}$$

em que $Z_i = \left(\frac{x_i - v}{b} \right)$ e $c_i = \left(\frac{\eta_i - v}{b} \right)$, então a matriz de informação de Fisher, denominada $I\left(v, b\right)$, é composta pelo menor valor esperado das derivadas de segunda ordem dos parâmetros, cujas derivadas podem ser encontradas assumindo $\delta_i = 1$.

Considerando ainda, que Z_i tem uma distribuição de valor extremo padrão truncada em c_i, possuindo função densidade de probabilidade dada por

$$f\left(t\right) = \frac{1}{b} e^{-Z_i} \exp \left(-e^{Z_i} \right), \tag{2.81}$$

e observando também que

$$\Pr(\delta_i = 1) = 1 - \Pr(\delta_i = 0)$$

é possível obter

$$\begin{aligned} I_{vv,i} &= E \left(- \frac{\partial^2 \log L_i}{\partial v^2} \right), \\ I_{bb,i} &= E \left(- \frac{\partial^2 \log L_i}{\partial b^2} \right), \\ I_{vb,i} &= E \left(- \frac{\partial^2 \log L}{\partial v \partial b} \right), \end{aligned} \tag{2.82}$$

em que: $I_{vv,i}$ é o menor valor esperado da segunda derivada em relação ao

parâmetro v, $I_{bb,i}$ é o menor valor esperado da segunda derivada em relação ao parâmetro b e $I_{vb,i}$ é a derivada parcial de segunda ordem em relação aos parâmetros v e b.

Na prática, a estimativa de $I\left(v,b\right)^{-1}$ é dada por $I\left(\hat{v},\hat{b}\right)^{-1}$. Para encontrar a matriz de Fisher $I\left(v,b\right)$, os cálculos envolvidos nas integrais são de grau elevado, e o logaritmo do tempo censurado pode não ser conhecido para todos os indivíduos da amostra. Um procedimento simples é a aproximação de $\left(\hat{v},\hat{b}\right) \sim N_2\left[\left(v,b\right), I_0^{-1}\right]$, onde I_0 é a matriz de informação observada, dada por:

$$I_0 = \left[\begin{array}{cc} -\frac{\partial^2 \log L}{\partial v^2} & -\frac{\partial^2 \log L}{\partial v \partial b} \\ -\frac{\partial^2 \log L}{\partial v \partial b} & -\frac{\partial^2 \log L}{\partial b^2} \end{array}\right]_{\left(\hat{v},\hat{b}\right)} . \tag{2.83}$$

Considerando $\hat{Z}_i = \frac{(x_i - \hat{v})}{\hat{b}}$ no cálculo das segundas derivadas da função log verossimilhança, igualando as derivadas de primeira ordem em relação aos parâmetros v e b a zero e tomando também $\sum_{i=1}^{n} \exp \hat{Z}_i = r$ e $\sum_{i=1}^{n} \hat{Z}_i + \sum_{i=1}^{n} \hat{Z}_i^2 \exp \hat{Z}_i = r$, tem-se que

$$I_0 = \frac{1}{\hat{b}^2}\left[\begin{array}{cc} r & \sum_{i=1}^{n} \hat{Z}_i \exp \hat{Z}_i \\ \sum_{i=1}^{n} \hat{Z}_i \exp \hat{Z}_i & r + \sum_{i=1}^{n} \hat{Z}_i^2 \exp \hat{Z}_i \end{array}\right] . \tag{2.84}$$

Embora estes procedimentos são totalmente adequados para amostras grandes, Billmann (1972) e Lawless (1975), alertam que se o tamanho da amostra for pequeno ou moderado, a aproximação normal pode não ser válida. Por isso, é sempre sensato verificar se a distribuição dos estimadores de máxima verossimilhança realmente se aproxima da distribuição normal.

Uma maneira de contornar este problema é utilizar uma transformação no

estimador de máxima verossimilhança em amostras pequenas, isto é, em vez de tratar $\hat{b}$ como aproximadamente normal é preferível tratar $\log \hat{b}$ como aproximadamente normal.

2.10 Método da Razão de Verossimilhanças

Qualquer teste de hipótese divide o espaço amostral em duas regiões mutuamente exclusivas, isto é, a região de rejeição e de aceitação. A decisão de um teste consiste em verificar em qual região o vetor de parâmetros pertence.

Para estimar o vetor de parâmetros verdadeiro θ_n escolhe-se o vetor de parâmetros que maximiza a função de verossimilhança no domínio de θ.

A função de verossimilhança maximizada sob a hipótese H_0 produz outra estimativa $\tilde{\theta}_{H_0}$. Dessa forma, o teste da razão de verossimilhança consiste em comparar o valor da verossimilhança do valor do parâmetro da hipótese com o valor da máxima verossimilhança, ou seja

$$\frac{L\left(\tilde{\theta}_{H_0}\right)}{L\left(\tilde{\theta}_n\right)}, \tag{2.85}$$

em que a razão desta comparação não deve exceder 1. H_0 é plausível se a razão em (2.85) não for muito distante de 1. Reciprocamente, quando (2.85) assume valores muito inferiores que 1, H_0 não é plausível.

Denotando a deviance do modelo ajustado sob H_0 por $-2\log\tilde{\theta}_{H_0}$, então

$$\Lambda = 2\left[\log\left(\tilde{\theta}_n\right) - \log\left(\tilde{\theta}_{H_0}\right)\right] \tag{2.86}$$

é a diferença estatística das deviances. Valores de Λ próximos de zero sugerem que a hipótese H_0 é plausível, em contrapartida H_0 não será aceita para valores grandes de Λ.

Por exemplo, a partir do logaritmo da função de verossimilhança do modelo

Weibull, que é dado por

$$\log L(\beta, \mu) = r \log \beta - r\beta \log \mu + (\beta - 1) \sum_{i \in D} \log t_i - \sum_{i=1}^{n} \left(\frac{t_i}{\mu} \right)^{\beta}. \quad (2.87)$$

é possível obter intervalos de confiança ou realizar testes de hipóteses para os parâmetros β, μ e para os percentis t_p. Sendo que os t_p são obtidos através de $t_p = \mu \left[-\log (1-p) \right]^{\frac{1}{\beta}}$.

2.10.1 Testes e Intervalos de Confiança para β

Suponha que tem-se interesse em testar a hipótese $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. O teste da razão de verossimilhanças pode ser utilizado. Primeiramente, é necessário obter o estimador de máxima verossimilhança de μ e β, sob H_0.

Considerando o estimador de máxima verossimilhança de β, sendo $\tilde{\beta} = \beta_0$, e o estimador de máxima verossimilhança μ que é encontrado maximizando $\log L(\mu, \beta_0)$ com respeito a μ, resulta em

$$\tilde{\mu} = \left(\sum_{i=1}^{n} \frac{t_i^{\beta_0}}{r} \right)^{\frac{1}{\beta_0}}. \quad (2.88)$$

Para o modelo Weibull, a estatística da razão de verossimilhanças para testar H_0 contra H_1 é dada por

$$\Lambda = -2 \left[\log L(\hat{\beta}, \hat{\mu}) - \log L(\beta_0, \tilde{\mu}) \right], \quad (2.89)$$

onde $\hat{\mu}$ e $\hat{\beta}$ são os estimadores de máxima verossimilhança do modelo Weibull sem restrição, dado como solução para as equações (2.72) e(2.73). Assim, para valores grandes de Λ rejeita-se H_0.

Para amostras de tamanho grande, Λ tem aproximadamente uma distribuição qui-quadrado com um grau de liberdade, sob H_0. Então, com esta distribuição

de referência, pode-se operacionalizar o teste estatístico.

Para obter um intervalo de confiança para β, deve-se encontrar um conjunto de valores β_0 para que H_0 não seja rejeitada ao nível de significância α, isto é, um conjunto de valores β_0, tal que $\Lambda \leq \chi^2_{1,\alpha}$.

2.10.2 Testes e Intervalos de Confiança para μ

Para testar a hipótese $H_0 : \mu = \mu_0$ versus a hipótese $H_1 : \mu \neq \mu_0$ deve-se maximizar $\log L(\beta, \mu)$, sob a restrição $\mu = \mu_0$. Encontrando $\frac{\partial \log L(\mu_0, \beta)}{\partial \beta} = 0$, tem-se a equação

$$\frac{r}{\beta} - r \log \mu_0 + \sum_{i \in D} \log t_i - \sum_{i=1}^{n} \left(\frac{t_i}{\mu_0}\right)^{\beta} \log\left(\frac{t_i}{\mu_0}\right) = 0. \tag{2.90}$$

A resolução da equação (2.90) é encontrada iterativamente para um dado $\tilde{\beta}$. E a estatística da razão da verossimilhança para testar a hipótese H_0 é dada por

$$\Lambda = 2\left[\log L(\hat{\beta}, \hat{\mu}) - \log\left(\tilde{\beta}, \mu_0\right)\right]. \tag{2.91}$$

Valores grandes de Λ faz com que H_0 seja rejeitada.

Um intervalo de confiança para μ, com nível de confiança de $(1 - \alpha)$ é obtido através do conjunto de todos os valores μ_0, tal que $\Lambda \leq \chi^2_{1,\alpha}$.

2.10.3 Testes e Intervalos de Confiança para t_p

Considere as hipóteses $H_0 : t_p = Q_0$ versus $H_1 : t_p \neq Q_0$ ou seja, H_0 especifica que $\mu\left[-\log(1-p)\right]^{1/\beta} = Q_0$ ou que $\mu = \frac{Q_0}{\left[-\log(1-p)\right]^{\frac{1}{\beta}}}$. Para achar os estimadores de máxima verossimilhanças $\tilde{\mu}$ e $\tilde{\beta}$ de μ e β, respectivamente, sob H_0 pode-se considerar $\mu = Q_0 \diagup \left[-\log(1-p)\right]^{1/\beta}$ em $\log L(\mu, \beta)$. Difer-

enciando com respeito a β e igualando a zero tem-se

$$\frac{r}{\beta} - r \log Q_0 + \sum_{i \in D} \log t_i + \log (1-p) \sum_{i=1}^{n} \left(\frac{t_i}{Q_0} \right)^{\beta} \log \left(\frac{t_i}{Q_o} \right) = 0. \quad (2.92)$$

Para encontrar o estimador $\tilde{\beta}$, resolve-se iterativamente a equação (2.92) e logo após encontra-se o valor de $\tilde{\mu} = Q_0 / \left[-\log (1-p) \right]^{1/\tilde{\beta}}$. A estatística para a realização do teste da razão de verossimilhanças é

$$\Lambda = 2 \left[\log L \left(\hat{\beta}, \hat{\mu} \right) - \log L \left(\tilde{\beta}, \tilde{\mu} \right) \right]. \quad (2.93)$$

Para valores de Λ grande, rejeita-se H_0. Determinando o conjunto de valores Q_0 que satisfaz $\Lambda \leq \chi^2_{1,\alpha}$, é possível encontrar um intervalo de confiança para t_p.

Também pode ser realizado testes e intervalos de confiança para a função de sobrevivência utilizando o mesmo procedimento do teste da razão de verossimilhanças para t_p.

Suponha-se que se queira realizar um teste ou encontrar um intervalo de confiança para $S(t_0) = \exp \left[- \left(\frac{t_0}{\mu} \right)^{\beta} \right]$, em que t_0 é fornecido. Considerando as hipóteses $H_0 : S(t_0) = S_0$ e $H_1 : S(t_0) \neq S_0$, desde que $S(t_0) = S_0$,tem-se

$$\mu = \frac{t_0}{\left[-\log (S_0) \right]^{\frac{1}{\beta}}}. \quad (2.94)$$

Para maximizar $\log L(\beta, \mu)$ sob a hipótese nula, utiliza-se a equação (2.92) a fim de obter $\tilde{\beta}$, substituindo Q_0 por t_0 e $(1-p)$ por S_0.

Um intervalo de confiança com nível de confiança $(1-\alpha)$ para $S(t_0)$ consiste no conjunto de valores de S_0, tal que $\Lambda = -2 \log L \left(\tilde{\beta}, \tilde{\mu} \right) + 2 \log L \left(\hat{\beta}, \hat{\mu} \right) \leq \chi^2_{1,\alpha}$, onde $\tilde{\mu} = t_0 / \left(-\log S_0 \right)^{1/\tilde{\beta}}$.

Capítulo 3

Modelos de Mistura de Longa Duração

3.1 Caracterização

Ao estudar análise de sobrevivência, na maioria das vezes, o evento de interesse está centrado em investigar a morte de um paciente, a ocorrência ou a recorrência de uma doença, o retorno de um prisioneiro liberado da prisão e assim por diante.

Quando acontece de uma população ter um grande número de observações censuradas, há um indício de que nesta população existe uma fração de indivíduos que não estão sujeitos em experimentar o evento de interesse. Esses indivíduos são considerados "imunes", "curados" ou "imortais".

Um indivíduo é considerado "imune" quando não experimenta o evento de interesse no tempo de observação definido, dessa maneira um indivíduo imune sempre ter sua observação censurada. Mazucheli (2002) define

> "Os modelos de mistura de longa duração são os modelos que consideram que existe na população uma fração de itens ou indivíduos que não estão sujeitos a experimentar o evento de interesse. Esses itens ou indivíduos que pertencem a esta fração são considerados "curados" ou "imortais" e seus tempos de vida são observados como observações censuradas."

Uma análise de dados de sobrevivência com indivíduos imunes, consiste em ajustar um modelo paramétrico, que é uma mistura de duas distribuições: uma

que representa o tempo de sobrevivência para os suscetíveis (ou seja, dos itens que falharam, dos indivíduos que morreram, dos ex-detentos que reincidiram no crime) e a outra distribuição degenerada, que permite cronometrar o tempo de sobrevivência dos imunes. Este modelo assume que uma certa fração p da população são suscetíveis e o complemento $(1-p)=q$ que são imunes, onde $(0 \leq p \leq 1)$.

O modelo de mistura de longa duração apresenta função de sobrevivência dado por

$$S(t) = (1-p) + pS^*(t), \tag{3.1}$$

onde $S^*(t)$ denota a função de sobrevivência para a proporção dos indivíduos considerados não imunes.

A função de distribuição acumulada do tempo de sobrevivência de um modelo de mistura tem representação paramétrica dada por

$$F(t) = pF^*(t), \tag{3.2}$$

em que $F^*(t)$ é a função distribuição acumulada do tempo de sobrevivência dos indivíduos suscetíveis, ou seja, os não imunes.

3.1.1 Modelo de Mistura de Longa Duração Exponencial

Na situação onde a variável aleatória T, não negativa, representa o tempo de sobrevivência de um indivíduo, a função distribuição acumulada do tempo de vida do modelo de mistura de longa duração exponencial fica expressa na forma

$$F(t) = p\left[1 - \exp\left(-\frac{t}{\mu}\right)\right], \tag{3.3}$$

em que $\mu > 0$.

Derivando a equação (3.3), é possível encontrar a função densidade de

probabilidade para o modelo de mistura exponencial, dada por

$$f(t) = (p)\,\frac{1}{\mu}\exp\left(-\frac{t}{\mu}\right). \tag{3.4}$$

A função de sobrevivência para o modelo de mistura de longa duração exponencial é dada na forma

$$S(t) = (1-p) + p\exp\left(-\frac{t}{\mu}\right). \tag{3.5}$$

Com função de risco dada por

$$h(t) = \frac{f(t)}{1-F(t)}, \tag{3.6}$$

ou seja

$$h(t) = \frac{\frac{1}{\mu}p\exp\left(-\frac{t}{\mu}\right)}{1-p\left[1-\exp\left(-\frac{t}{\mu}\right)\right]}. \tag{3.7}$$

3.1.2 Modelos de Mistura de Longa Duração Weibull

Suponha que a vida dos elementos não imunes (suscetíveis) a um evento de interesse tenha uma função de distribuição acumulada dada por uma distribuição Weibull. Assim, a função de distribuição acumulada de toda a população é representada por

$$F\left(t\right) = p\left\{1-\left[\exp\left(-\frac{t}{\mu}\right)^{\beta}\right]\right\}, \tag{3.8}$$

em que p é a proporção de suscetíveis, $\beta > 0$ parâmetro de forma e $\mu > 0$

parâmetro de escala.

Derivando (3.8) encontra-se a função densidade de probabilidade do modelo de mistura Weibull, dada por

$$f(t) = \frac{1}{\mu^{\beta}} (1-p)\, \beta t^{\beta-1} \exp\left(-\frac{t}{\mu}\right)^{\beta}. \tag{3.9}$$

A função de sobrevivência do modelo de mistura Weibull é expressa na forma,

$$S(t) = (1-p) + p \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right]. \tag{3.10}$$

Apresentando função de risco

$$h(t) = \frac{\frac{1}{\mu^{\beta}} p\beta t^{\beta-1} \exp - \left(\frac{t}{\mu}\right)^{\beta}}{1 - p\left\{1 - \left[\exp - \left(\frac{t}{\mu}\right)^{\beta}\right]\right\}}, \tag{3.11}$$

onde $h(t) \geq 0$ para todos os $t \geq 0$.

3.2 Áreas de Aplicações

Muitos autores contribuíram para a teoria dos modelos de mistura de longa duração, o pioneiro no trabalho foi Boag (1949) que publicou um artigo no Journal of Royal Statistical Society. Ele usou o método de máxima verossimilhança para estimar a proporção de pacientes curados, numa população de 121 mulheres com câncer de mama, experimento esse que teve a duração de 14 anos.

Na área médica, os seguidores de Boag foram Berkson e Gage (1952), Haybittle (1959), Mould e Boag (1975), Bitthel e Upton (1977), Langlands et al (1979), Goldman (1984), Kimbre e Crowder (1984) Farawell (1982, 1986), recentemente têm-se Gamel et al (1990), Ghitany e Maller (1992), Ng e McLach-

lan (1998), Peng e Dear (2000) Sy e Taylor (2000).

Outra área de aplicação de modelos de mistura de longa duração é a de criminologia, que analisa a proporção da reincidência de indivíduos que foram libertados após a primeira apreensão. Também pode ser de interesse o complemento do reincidismo, ou seja a probabilidade de que um indivíduo não retorne à prisão após ser libertado. O primeiro a modelar dados em criminologia foi Partanen (1969), seguido por Maltz e McCleary (1977), Bloom(1979), Schimidt e Witte (1988). Recentemente, Broadhurst e Maller (1991) verificaram que uma das características desta área é a possibilidade de fazer a análise em arquivos gigantescos.

A avaliação da confiabilidade, na área da engenharia, teve como precursor Nelson (1982), Meeker e Lu Valle (1995). Nesta área, os modelos de longa duração são utilizados para verificar a proporção da vida dos componentes que são colocados em teste no tempo zero e expostos a vários regimes de tensão ou uso.

Na área de mercado, os imunes são considerados os indivíduos que nunca comprarão um certo produto (Anscombe, 1961). Para os investigadores da comercialização é necessário a estimação da distribuição do tempo gasto para a aquisição do novo produto e a proporção dos imunes ao produto.

Gulland (1955) sugeriu um modelo de mistura para estudo da recaptura de peixes que são marcados e jogados no lago, sendo que a proporção dos imunes representa o peixe que nunca é recapturado.

Em educação, os modelos de mistura foram estudados por Regal e Larntz (1978) . Estes apresentaram a um grupo de estudantes um quebra-cabeça para ser montado, e o tempo que cada um levou para resolver foi registrado. Os estudantes que não resolveram o quebra cabeça foram considerados indivíduos imunes para a resolução.

Dunsmuir (1989) analisou que no mercado de trabalho existe uma proporção de indivíduos que não conseguem emprego. Yamaguchi (1994) trabalhou com uma aplicação do modelo de mistura de longa duração numa análise dos empregos permanentes no Japão.

3.3 Estimação de Máxima Verossimilhança para Modelos de Mistura de Longa Duração.

Especificado o modelo paramétrico para representar a função de distribuição acumulada $F^*(t)$ das vidas dos indivíduos suscetíveis a um certo evento, função essa que pode ser obtida através de uma observação visual ou pelo ajustamento de um modelo padrão (por exemplo, o Weibull, exponencial, log-logística, etc.), os parâmetros envolvidos na função podem ser estimados através do método de máxima verossimilhança.

Considere o caso de censura do tipo II, onde n itens são colocados em teste no tempo zero e a observação cessará após um número pré-determinado $(r < n)$ de fracassos. Sejam $T_1, T_2, ..., T_n$ variáveis aleatórias independentes, representando as possíveis observações censuradas.

Sejam estas variáveis aleatórias em ordem crescente

$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}. \tag{3.12}$$

Supondo-se, ainda, a função distribuição acumulada de sobrevivência, como uma função de distribuição contínua com densidade $f(t) = F'(t)$, pode-se considerar os tempos de falha $t_1, t_2, ...t_n$ ordenados como

$$t_{(1)} < t_{(2)} < \cdots < t_{(r)} = t_{(r+1)} = \cdots = t_n.$$

Usando os argumentos da teoria de estatísticas de ordem, Lawless (1982), tem-se que a função densidade de probabilidade das variáveis aleatórias independentes $T_1, T_2, ..., T_r$ é dada por

$$f(t) = \left(\prod_{i=1}^{n} f(t_i) \right) [1 - F(t_r)]^{n-r}, \tag{3.13}$$

onde $t_1, t_2, ...t_r$ são números reais que satisfazem $0 < t_{(1)} < t_{(2)} < \cdots < t_{(r)}$.

Dessa maneira pode-se escrever a equação (3.13) em termos originais, ou seja, sem ordem e com indicadores de censura $\delta_1, \delta_2, ..., \delta_n$, onde tem-se que $\delta_i = 1$ se t_i é sem censura; e $\delta_i = 0$ se t_i for censurado.

Considerando censura do tipo II, tem-se que $\delta_i = 1$, se $t_1 \leq t_r$; e $\delta_i = 0$, em caso contrário. Assim, a equação (3.13) pode ser escrita como a função de verossimilhança

$$L\left(\mathbf{t_i}\right) = k \prod_{i=1}^{n} \left[f\left(t_i\right)^{\delta_i} \left(1 - F\left(t_i\right)\right)^{1-\delta_i} \right] \tag{3.14}$$

onde k é uma constante que não depende de t_i, δ_i ou de qualquer parâmetro desconhecido.

Similarmente, quando tem-se censura do tipo I, onde n itens são observados desde o tempo zero até algum tempo pré-determinado $(\Delta > 0)$, fixado um número aleatório R de falhas, pode-se observar o tempo de falhas sem censura $t_{(1)} < t_{(2)} < \cdots < t_{(R)}$, com o restante $(n - R)$ tempos censurados até Δ. Assim, a função densidade conjunta de $t_{(1)}, \ldots, t_{(R)}$, e R é a mesma dada por (3.13) mas com o fator $\left(1 - F\left(t_r\right)\right)^{n-r}$ substituído por $\left(1 - F\left(\Delta\right)\right)^{n-r}$ e r substituído por R, quando for definido para observações censuradas em que $t_i = \Delta$ para $\delta_i = 0$.

A equação (3.14) sugere que o princípio geral para escrever a probabilidade de uma amostra com algumas observações censuradas, particularmente censura à direita, é multiplicar um fator $f\left(t_i\right)$ para qualquer observação t_i não censurada, onde $f\left(\cdot\right)$ é a função densidade do tempo de sobrevivência, e multiplicar um fator $\left(1 - F\left(t_i\right)\right)$ para qualquer observação censurada em t_i.

Considerando o modelo de mistura de longa duração Weibull, cuja função de distribuição acumulada é dada pela equação (3.8) e a função densidade de

probabilidade dada pela equação (3.9), então sua função de verossimilhança é

$$L(p,\beta,\mu) = k\prod_{i=1}^{n}\left[\frac{1}{\mu^{\beta}}p\beta t_i^{\beta-1}\exp-\left(\frac{t_i}{\mu}\right)^{\beta}\right]^{\delta_i}\left\{1-\left[p+p\exp-\left(\frac{t_i}{\mu}\right)\right]^{\beta}\right\}^{1-\delta_i}. \tag{3.15}$$

Observe que, quando $(p=1)$ implica na ausência de imunes na população e, dessa forma, o modelo de mistura se reduz ao modelo Weibull padrão.

Fixando $\beta = 1$ em (3.15) tem-se a função de verossimilhança do modelo de mistura de longa duração exponencial, dada por

$$L(p,1,\mu) = k\prod_{i=1}^{n}\left[\frac{1}{\mu}p\exp\left(-\frac{1}{\mu}\right)\right]^{\delta_i}\left\{1-\left[p+p\exp\left(-\frac{1}{\mu}\right)\right]\right\}^{1-\delta_i}. \tag{3.16}$$

Quando $p=1$, a equação (3.16) torna-se a função verossimilhança para o modelo exponencial padrão.

A função de verossimilhança L pode ser usada para estimar os parâmetros (p,β,μ), e também testar hipóteses sobre eles.

Como já foi dito, é sempre conveniente trabalhar com o logaritmo da função de verossimilhança, desde de que o produto final seja uma soma de verossimilhanças formada por observações independentes.

3.4 Teste da Razão de Verossimilhanças e Deviance para Modelos de Mistura de Longa Duração

Ao desejar testar uma hipótese H_0 em relação a algum ou todos os componentes que o vetor de parâmetros θ assume, utilizar-se-à o teste da razão de máxima verossimilhanças.

Quando ocorrer dos dados se ajustarem, tanto ao modelo de mistura de longa duração Weibull e ao modelo de mistura exponencial, deve-se testar a hipótese $H_0 : \beta = 1$. Caso o teste da razão de máxima verossimilhanças

aceitar a hipótese $H_0 : \beta = 1$, então pode reduzir o modelo de mistura de longa duração Weibull para o modelo de mistura exponencial.

Caso deseja-se verificar se a proporção de imunes é significativa no modelo ajustado, deve-se testar a hipótese $H_0 : p = 1$. Se não for possível rejeitar H_0, o modelo se reduz ao modelo Weibull padrão, o que implica na não existência de indivíduos imunes, pois ocorre a restrição do vetor paramétrico para um subconjunto do mesmo.

A estatística para a realização do teste da razão de máxima verossimilhança ou a diferença de deviances para modelos de mistura de longa duração são as mesmas apresentadass pelas equações (2.85) e (2.86) do capítulo 2.

Para ilustração dos conceitos abordados até o presente momento, considere os dados da Tabela 3.1, apresentados em Maller e Zhou (1996 p. 82), onde tem-se o tempo de recorrência da leucemia ou de censura, em anos, de 46 pacientes que foram submetidos ao transplante alogênico.

A partir dos dados apresentados na Tabela 3.1, será conduzida uma comparação com relação ao ajuste dos modelos Weibull padrão e o de mistura de longa duração Weibull, isto quer dizer que será analisada a existência ou não de uma proporção de indivíduos imunes.

Com o uso dos pacotes estatísticos SAS e R e utilizando o método de máxima verossimilhança, primeiramente o ajuste será feito para o modelo Weibull padrão e depois para o modelo de mistura de longa duração Weibull.

As estimativas de máxima verossimilhança e seus erros padrões para o modelo

Weibull padrão são $\hat{\mu}_0 = 2,0488\,(\pm 0,0926)$ e $\hat{\beta}_0 = 0,6208(\pm 0,5881)$.

As estimativas de máxima verossimilhança e seus erros padrões para o modelo de mistura de longa duração Weibull são $\hat{p}_1 = 0,7311\,(\pm 0,0759)$, $\hat{\beta}_1 = 0,9460\,(\pm 0,1487)$ e $\hat{\mu}_1 = 0,6890(\pm 0,1410)$.

Colocando a curva do modelo Weibull padrão ajustado (com seus parâmetros estimados), e a curva do modelo de mistura de longa duração Weibull, (com seus parâmetros estimados), juntamente com a curva das estimativas de Kaplan-Meier (KMEs) na 3.1, pode-se observar que o modelo de mistura de longa

Tabela 3.1: Tempos de Vida de Pacientes com Leucemia Submetidos ao Transplante Alogênico.

t_i	c_i	t_i	c_i	t_i	c_i	t_i	c_i
0,0301	1	0,2384	1	0,9096	1	3,0384	0
0,0384	1	0,2712	1	0,9644	1	3,1726	0
0,0630	1	0,2740	1	1,0082	1	3,4411	1
0,0849	1	0,3863	1	1,2822	1	4,4219	0
0,0877	1	0,4384	1	1,3452	1	4,4356	0
0,0959	1	0,4548	1	1,4000	1	4,5863	0
0,1397	1	0,5918	1	1,5260	1	4,6904	0
0,1616	1	0,6000	1	1,7205	0	4,7808	0
0,1699	1	0,6438	1	1,9890	0	4,9863	0
0,2137	1	0,6849	1	2,2438	1	5,0000	0
0,2137	1	0,7397	1	2,5068	0		
0,2164	1	0,8575	1	2,6466	0		

duração Weibull está melhor ajustado do que o modelo de Weibull padrão.

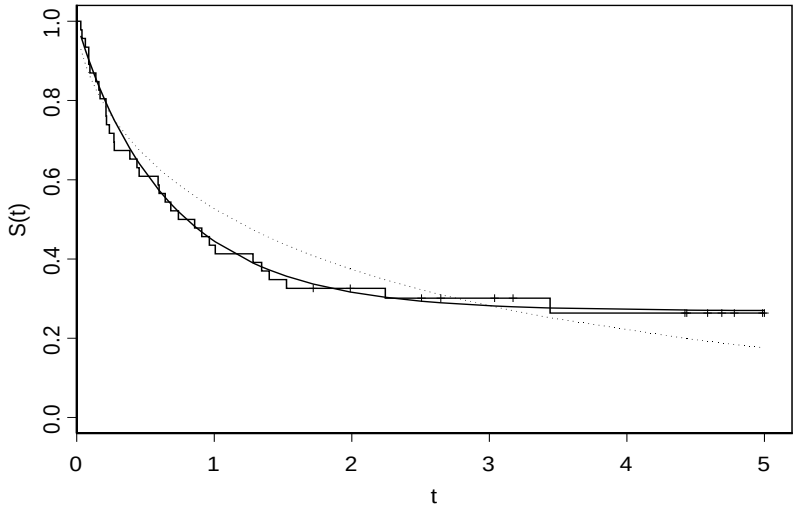


Figura 3.1: (——): Kaplan-Meier, $(\cdots\cdots)$: Modelo Weibull Padrão e $(\cdot - \cdot - \cdot)$: Modelo de mistura de longa duração Weibull.

Para comprovar tal afirmação, também será usado o valor das deviances dos dois modelos que estão sendo analisados. O valor da deviance para o modelo Weibull padrão ajustado é

$$\log \hat{L}_0(\hat{\beta}, \hat{\mu}) = -51,0653 \tag{3.17}$$

e o valor da deviance para o modelo de mistura de longa duração Weibull ajustado com seus parâmetros estimados é

$$\log \hat{L}_1 \left(\hat{p}, \hat{\beta}, \hat{\mu}\right) = -46,1550. \tag{3.18}$$

Para a confirmação da visualização gráfica a hipótese a ser testada será: $H_0 : p = 1$ versus $H_1 : p \neq 1$ (a aceitação de H_0 implica na não existência de indivíduos imunes e, então, conclui-se que o melhor modelo ajustado é o

modelo Weibull padrão). Calculando a diferença das deviances

$$\Lambda = 2\left[\left(\log \hat{L}_1\right) - \left(\log \hat{L}_0\right)\right] = 2\left[(-46,1550) - (-51,0653)\right] = 9,8206$$

tem-se que o valor encontrado para a diferença das deviances dos modelos ajustados foi de $9,8206$, valor muito grande, comparado ao valor $3,84$ (valor tabelado da distribuição de qui quadrado com um grau de liberdade). Então H_0 não é plausível, implicando, dessa forma, que o modelo de mistura de longa duração Weibull, ajustado ao conjunto dos dados da recorrência de leucemia nos 46 pacientes submetidos ao transplante alogênico, é melhor, confirmando assim a existência de uma proporção de imunes.

A mesma análise feita para os pacientes submetidos ao transplante alogênico (grupo-1), será agora considerada para 44 pacientes de leucemia que foram submetidos ao transplante autólogo (grupo-2).

Tabela 3.2: Tempo de Vida de Pacientes com Leucemia Submetidos ao Transplante Autólogo.

tempo (t_i)	c_i	tempo (t_i)	c_i	tempo (t_i)	c_i	tempo (t_i)	c_i
0,0575	1	0,2164	1	0,3589	1	0,7589	1
0,1096	1	0,2219	1	0,4027	1	1,9836	0
0,1370	1	0,2411	1	0,4685	1	1,9973	0
0,1452	1	0,2603	1	0,4712	1	2,0110	1
0,1479	1	0,2685	1	0,4904	1	2,8849	0
0,1534	1	0,2685	1	0,5178	1	2,9973	0
0,1671	1	0,2712	1	0,5342	1	3,2658	0
0,1753	1	0,2849	1	0,5452	1	4,0411	0
0,1836	1	0,2877	1	0,5836	1	4,2055	0
0,2000	1	0,2904	1	0,6110	1	4,2055	0
0,2082	1	0,3068	1	0,6137	1	5,0548	0

Os parâmetros estimados para o conjunto de dados do grupo 2 para o modelo Weibull padrão, juntamente com seus erros padrões, foram $\hat{\mu}_0 = 1,0988\,(\pm 0,2839)$ e $\hat{\beta}_0 = 0,6721\,(\pm 0,0906)$.

O modelo de mistura de longa duração Weibull obteve os seguintes estimadores de máxima verossimilhança, com seus erros padrões: $\hat{p} = 0,7975\,(\pm 0,0630)$, $\hat{\beta}_1$=$1,3711(\pm 0,1612)$ e $\hat{\mu}_1 = 0,4104\,(\pm 0,0557)$.

A curva do ajuste dos modelos Weibull padrão e o de mistura de longa duração Weibull foram colocados juntamente com as estimativas de Kaplan-Meier.

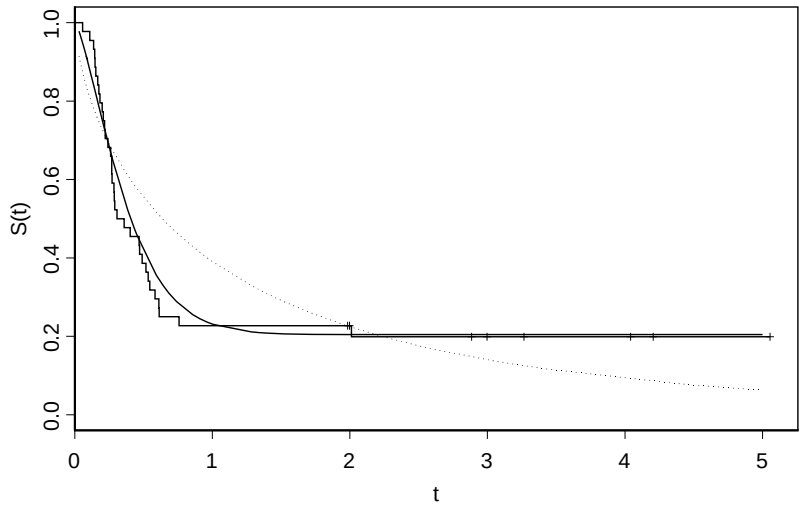


Figura 3.2: (——): Kaplan-Meier, $(\cdots\cdots)$: Modelo Weibull Padrão e $(\cdot - \cdot - \cdot)$: Modelo de mistura de longa duração Weibull.

Novamente, tem-se a impressão de que o modelo de mistura de longa duração Weibull está melhor ajustado. Porém deve-se testar a hipótese H_0 : $p = 1$, para confirmar.

Os valores das deviance dos modelos Weibull padrão e do modelo de mistura de longa duração Weibull foram: $\log \hat{L}_0(\hat{\beta}, \hat{\mu}) = 37,0310$ e $\log \hat{L}_1\left(\hat{p}, \hat{\beta}, \hat{\mu}\right) = -19,4406$, respectivamente. Assim $\Lambda = 2\,[(37,0310) - (-19,4406)] = 112,9432$, um valor grande, comparado com o valor crítico $3,84$, concluindo assim que a hipótese H_0 é rejeitada, ou seja, há evidência da existência de indivíduos imunes na população dos submetidos ao transplante autólogo.

Pela análise descrita pelos dois tipos de transplantes que os pacientes foram

submetidos, pôde-se confirmar que existe uma proporção de indivíduos imunes tanto para os pacientes que foram submetidos ao transplante alogênico, como para os pacientes que foram submetidos ao transplante autólogo.

3.5 Presença de Covariáveis

Usualmente, os dados de sobrevivência ou tempo de falha vêem com informações associadas ou concomitantes, nas quais o tempo de sobrevivência é dependente destas informações. Na área médica, estas informações poderão ser: o efeito de um ou mais tratamentos comparados com um grupo controle, a classificação do paciente pelo sexo masculino ou feminino, ser jovem ou velho, etc.

Já em criminologia, a reincidência de um prisioneiro após sua liberação pode depender do sexo, raça, tipo de liberação e tipo de transgressão cometida. Também, pode ocorrer de os prisioneiros serem participantes de grupos de aconselhamento terapêutico ou não.

As informações das covariáveis (também chamadas de variáveis explicativas ou regressoras) sobre um indivíduo i podem ser expressas por um vetor da forma

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}),$$

de maneira que os dados consistem de n observações na forma (t_i, δ_i, x_i). As covariáveis afetam o tempo de sobrevivência e podem ser formalizados parametricamente pela distribuição de cada t_i , que dependa do vetor de covariáveis $\mathbf{X}_i$.

Análises realizadas com dados que possuem covariáveis muitas vezes chegam a ser mais importantes e completas do que para conjunto de dados sem a presença de informações concomitantes. Na maioria das vezes, o conjunto de dados em que as covariáveis estão presentes, pode ser representado como sendo o "grupo de tratamento" e o conjunto de dados sem a presença de covariávies, como o "grupo controle" .

Para verificar se o modelo ajustado está explicitamente melhor com o uso das covariáveis, pode-se analisar:

a) a proporção de falha, a qual é mensurada pela magnitude do parâmetro β;

b) a proporção de imunes, caso exista, cuja magnitude é mensurada por $(1-p)$;

c) através do parâmetro de escala μ.

Considerando que para um conjunto de dados qualquer, o modelo paramétrico ajustado foi o modelo de mistura Weibull, então a função de distribuição acumulada assumida para o tempo de sobrevivência do grupo tratamento e do grupo controle será, respectivamente:

$$F_{trat}(t) = p_1 \left[1 - \exp\left(-\frac{t}{\mu_1}\right)^{\beta_1}\right], \tag{3.19}$$

$$F_{con}(t) = p_2 \left[1 - \exp\left(-\frac{t}{\mu_2}\right)^{\beta_2}\right], \tag{3.20}$$

onde os parâmetros que descrevem os dois grupos podem ser diferentes, ou não; e tanto o parâmetro p_1 como o p_2 assumem valores entre 0 e 1.

Quando o vetor de covariáveis $\mathbf{X}_i$ reduz-se a uma covariável discreta (unidimensional), em que o indivíduo i recebe o valor 1 se está no grupo controle; e o valor 2 se está recebendo algum tipo de droga, é visto como um caso trivial de covariáveis.

Porém, quando não existe a redução do vetor de covariáveis, faz-se necessário o uso de uma codificação, a fim de simplificar as hipóteses a serem testadas.

Colocar-se-á índices nas hipóteses anunciando os modelos a serem testados. Sempre que o índice for 0 (zero) indica que os parâmetros são iguais nos dois grupos; e sempre que o índice for 1 indica que os parâmetros podem diferir. O primeiro índice refere-se ao parâmetro de proporção p; o segundo índice ao

parâmetro de forma β; e o terceiro índice ao parâmetro de escala μ.Veja, por exemplo, os modelos:

a) Modelo 011 - Os $p's$ dos dois grupos são o mesmo, porém os $\beta's$ e os $\mu's$ podem diferir. (hipótese H_{011});

b) Modelo 101 - Os $\beta's$ dos dois grupos são iguais, enquanto os $p's$ e os $\mu's$ podem diferir. (hipótese H_{101});

Como exposto, pode ocorrer dos parâmetros de forma (β_1 e β_2) no modelo Weibull nos dois grupos ser o mesmo, independentemente dos demais parâmetros.

c) Modelo 110 - Os $\mu's$ dos dois grupos são iguais, enquanto os $p's$ e os $\beta's$ podem diferir. (hipótese H_{110}).

Também existem modelos em que dois parâmetros são iguais:

a) Modelo 001 - Os valores de $p_1 = p_2$ e de $\beta_1 = \beta_2$, mas μ_1 pode diferir de μ_2.

b) Modelo 010 - Os valores de $p_1 = p_2$ e de $\mu_1 = \mu_2$, mas β_1 pode diferir de β_2.

c) Modelo 100 - Os valores de $\mu_1 = \mu_2$ e de $\beta_1 = \beta_2$, mas p_1 pode diferir de p_2.

Existem também os modelos que permitem ao grupo ter sua própria descrição em relação aos seus parâmetros, tal como o Modelo 111, no qual todos podem diferir.

Já o Modelo 000 indica que não existe diferença entre os parâmetros e, neste caso, não faz diferença entre utilizar a distribuição de sobrevivência para o modelo ajustado do grupo controle ou do grupo tratamento, ou seja, com ou sem a utilização de covariáveis.

Considerando, primeiramente, o ajustamento do Modelo 111, que permite que os parâmetros possam diferir nos dois grupos em questão, e que do total de observações as primeiras n_1 venham do grupo tratamento e o restante $(n_2 = n - n_1)$ do grupo controle, então a função de máxima verossimilhança para o Modelo 111 tem a forma

$$\begin{aligned} L_{111}(\theta) &= k\prod_{i=1}^{n_1}\left[\frac{1}{\mu_1^{\beta_1}}p_1\beta_1 t_i^{\beta_1-1}\exp\left(-\frac{t_i}{\mu_1}\right)^{\beta_1}\right]^{\delta_i}\left\{1-\left[p_1+p_1\exp\left(-\frac{t_i}{\mu_1}\right)^{\beta_1}\right]\right\}^{1-\delta_i} \\ &\quad \times k\prod_{i=n_{1+1}}^{n}\left[\frac{1}{\mu_2^{\beta_2}}p_2\beta_2 t_i^{\beta_2-1}\exp\left(-\frac{t_i}{\mu_2}\right)^{\beta_2}\right]^{\delta_i}\left\{1-\left[p_2+p_2\exp\left(-\frac{t_i}{\mu_2}\right)^{\beta_2}\right](\right. \end{aligned}$$

onde: $\boldsymbol{\theta} = (p_1, p_2, \beta_1, \beta_2, \mu_1, \mu_2)$; e k é uma constante que não depende dos parâmetros.

Observe que, para o Modelo 111, a função de verossimilhança entre dois componentes separados, sendo que um depende unicamente dos parâmetros $p_1, \beta_1 \; e \; \mu_1$, e o outro dos parâmetros $p_2, \beta_2 \; e \; \mu_2$, os componentes são exatamente a função de verossimilhança de cada um dos dois grupos considerados separados, ou seja, os parâmetros estimados: $\hat{p}_1, \hat{\beta}_1, \hat{\mu}_1$(do grupo de tratamento) e $\hat{p}_2, \hat{\beta}_2 \; e \; \hat{\mu}_2$ (do grupo controle), são os mesmos se obtidos pelo ajustamento do modelo de mistura de longa duração Weibull. E também o valor da log verossimilhança maximizada será a soma das log verossimilhanças maximizadas de cada grupo.

Assim:

$$\hat{L}_{111} = L_{111}\left(\hat{p}_1, \hat{\beta}_1, \hat{\mu}_1, \hat{p}_2, \hat{\beta}_2, \hat{\mu}_2\right). \tag{3.22}$$

Porém, quando se tem um modelo em que os grupos podem ter um ou mais parâmetros em comum, esta fatoração dentro dos componentes não pode ocorrer. Por exemplo, caso deseja-se testar a hipótese H_{011}, onde os valores dos parâmetros $p_1 = p_2$ e os $\beta' s$ e os $\mu' s$ podem diferir, a função de verossimilhança

tem a forma:

$$\begin{aligned} L_{011} &= k\prod_{i=1}^{n_1}\left[\frac{1}{\mu_1^{\beta_1}}p\beta_1 t_i^{\beta_1-1}\exp\left(-\frac{t_i}{\mu_1}\right)^{\beta_1}\right]^{\delta_i}\left\{1-\left[p+p\exp\left(-\frac{t_i}{\mu_1}\right)^{\beta_1}\right]\right\}^{1-\delta_1} \\ &\times k\prod_{i=n_{1+1}}^{n}\left[\frac{1}{\mu_2^{\beta_2}}p\beta_2 t_i^{\beta_2-1}\exp\left(-\frac{t_i}{\mu_2}\right)^{\beta_2}\right]^{\delta_i}\left\{1-\left[p+p\exp\left(-\frac{t_i}{\mu_2}\right)^{\beta_2}\right]\right\}^{1-\delta_1} \end{aligned} \quad (3.23)$$

em que p é o valor comum da estimação de p_1 e p_2, e k não depende dos parâmetros.

Assim, o ajustamento para o Modelo 011 é obtido unicamente pelos parâmetros estimados: $\hat{p}, \hat{\beta}_1, \hat{\mu}_1, \hat{\beta}_2\ e\ \hat{\mu}_2$ cuja a verossimilhança maximizada é

$$\hat{L}_{011} = L_{011}\left(\hat{p}, \hat{\beta}_1, \hat{\mu}_1, \hat{\beta}_2, \hat{\mu}_2\right). \quad (3.24)$$

Neste caso, o estimador de p é contribuído com informações de ambos os grupos, o que também faz modificar os estimadores dos demais parâmetros envolvidos.

Para comparar o ajustamento de dois modelos utilizar-se-á, como já abordado, o teste da razão de verossimilhança ou Análise de Deviance.

Assim, desejando-se testar a hipótese H_{111} (onde os parâmetros diferem nos dois grupos) versus a hipótese H_{011}(onde $p_1 = p_2$ e os $\beta's$ e os $\mu's$ podem diferir), toma-se os valores das verossimilhanças $\hat{L}_{111}$ e $\hat{L}_{011}$ e verifica-se através da diferença das deviances,

$$\Lambda = 2\left(\log\hat{L}_{111} - \log\hat{L}_{011}\right) \quad (3.25)$$

em que Λ tem, aproximadamente, uma distribuição de qui quadrado com graus de liberdade igual a diferença entre o número dos parâmetros estimados pelos modelos que estão sendo testados.

Mais uma vez, o teste de H_{011} versus H_{111} é realizado comparando o valor

de Λ com o maior percentil da distribuição χ^2_1.

Para realizar o teste a um nível de significância de 5%, basta verificar se Λ excede ou não o valor $3,84$ (valor tabelado da distribuição de qui quadrado com um grau de liberdade, pois o número de parâmetros estimados pelo Modelo 111 é 6 e o número de parâmetros estimados pelo Modelo 011 é 5). Se Λ exceder ao valor $3,84$, rejeita-se a hipótese H_{011} e conclui-se que a proporção de imunes nos dois grupos difere significativamente. Caso a hipótese H_{011} não seja rejeitada, conclui-se que a proporção de imunes nos dois grupos é a mesma, podendo assim o parâmetro p ser condensado em um único valor, sem perda de informações, ou seja $p_1 = p_2$.

Analogamente, pode-se testar qualquer hipótese e, então, analisar todos os modelos permitidos para escolher qual é o melhor modelo ajustado.

O que pode ser feito é começar com o modelo com os mesmos parâmetros (000) e ir somando parâmetros um por um, até que se obtenha uma melhoria adicional no ajuste (diferença significativa das deviances).

Lembrando que a análise descrita acima depende se o modelo de mistura (Weibull, exponencial ou outro qualquer) fornece uma boa descrição dos dados, ou seja, deve-se testar a qualidade do ajuste da distribuição para cada grupo separadamente ou testar a qualidade do ajuste do modelo.

Como um exemplo do ajustamento do modelo com uso de covariáveis trivial, novamente será utilizado o exemplo apresentado em Maller e Zhou (1996 p. 82), agora com os 90 indivíduos com leucemia, em que a covariável será o tipo de transplante recebido pelo paciente. Será comparado se o modelo de mistura de longa duração Weibull é melhor ajustado, levando em consideração o tipo de transplante que o paciente foi submetido. No grupo-1 estão os 46 pacientes que receberam o transplante alogênico e, no grupo-2, os 44 pacientes que receberam o transplante autólogo.

Para este exemplo, será comparado apenas os modelos 000 e 001, pois o tipo de tratamento está relacionado ao parâmetro de escala.

Comparando o Modelo 000, o qual afirma que os parâmetros com ou sem a presença da covariável tipo de transplante são iguais, contra o Modelo 001, em

que considera a presença da covariável tipo de transplante e tem por hipótese que os parâmetros p e β são iguais nos dois grupos, porém o parâmetro μ pode diferir.

Pelos valores das deviances dos modelos 000 e 001 tem-se $-2\log L_{000} = 143,752$ e $-2\log L_{001} = 136,006$. A diferença das deviances é: $7,746$. Comparando o valor da diferença das deviances com o valor da distribuição χ^2_1 (valor tabelado igual a $3,84$), conclui-se que as proporções de imunes p_1 e p_2 nos dois grupos são iguais e que as formas das curvas são iguais nos dois grupos $(\beta_1 = \beta_2)$, porém, o parâmetro de escala μ difere com o acréscimo da covariável tipo de transplante.

Um exemplo mais completo, com vários fatores, é o de câncer de ovário de 26 mulheres, apresentado por Maller e Zhou (1996, p.134), onde tem-se os tempos de vida ou de censuras e quatro covariáveis: o tipo de tratamento (x_{i1}), a idade que a paciente iniciou o tratamento (x_{i2}), a extensão da doença (x_{i3}) e o desempenho da paciente (x_{i4}), pois, além de modelar o parâmetro μ, pode-se também modelar os parâmetros p e β. Os dados encontram-se na 3.3.

A variável indicadora de censura é: 0 se a paciente morreu e 1 se a paciente teve seu tempo censurado.

A covariável tipo de tratamento é: 1 se a paciente recebe a quimioterapia padrão e 2 se recebe uma combinação de quimioterapia.

Quanto à covariável desempenho da paciente é: 1 se apresenta desempenho bom e 2 se apresenta desempenho ruim.

O interesse principal centra-se no efeito dos tratamentos de quimioterapia, depois leva-se em conta as diferentes idades, a extensão da doença e o desempenho das pacientes.

Observando os dados, pode-se perceber que os dois grupos, com diferentes tipos de tratamento, apresentam pacientes com observações censuradas, o que indica a possibilidade da existência de pacientes imunes a morte devido a doença.

Para a realização da análise, primeiramente será considerado o Modelo 000

Tabela 3.3: Tempos de Vida de Pacientes com Câncer de Ovário.

Tempo (t_i)	δ_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	Tempo (t_i)	δ_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
0,1616	1	1	72	2	1	1,3068	0	1	64	2	1
0,3151	1	1	74	2	1	1,5425	1	2	55	1	2
0,4274	1	1	66	2	2	1,7479	1	1	56	1	2
0,7342	1	1	74	2	2	2,0384	0	2	50	1	1
0,9014	1	1	43	2	1	2,1068	0	2	59	2	2
0,9671	1	2	63	1	2	2,1096	0	2	57	2	1
1,0000	1	2	64	1	1	2,2000	0	1	39	1	1
1,0329	0	2	58	1	1	2,3425	0	1	43	1	2
1,1534	0	2	53	2	1	2,8493	0	1	38	2	2
1,1808	1	1	50	2	1	3,0301	0	1	44	1	1
1,2274	0	1	56	1	2	3,0932	0	2	53	1	1
1,2712	1	2	56	2	2	3,3041	0	2	44	2	1
1,3014	1	2	59	2	2	3,3616	0	2	59	1	2

e assim, sucessivamente, modelando os parâmetros. A Tabela 3.4 indica os modelos, as estimativas dos parâmetros, juntamente com os valores de suas deviances, para cada tipo de tratamento.

Tabela 3.4: Estimativas de máxima verossimilhança dos modelos de mistura de longa duração.

Modelo	Número de Parâmetros	Tratamento	$\hat{p}$	$\hat{\beta}$	$\hat{\mu}$	Deviance
000	3		0,51	1,1530	2,11	49,31
100	4	1	0,59	1,1532	2,11	48,76
		2	0,43	1,1532	2,11	
110	5	1	0,57	0,9974	2,09	47,84
		2	0,44	1,3578	2,09	
111	6	1	0,58	0,9583	1,50	40,23
		2	0,44	1,3198	6,61	
101	5	1	0,65	1,3115	1,29	40,84
		2	0,44	1,3115	6,61	
001	4	1	0,53	1,3075	1,35	41,69
		2	0,53	1,3075	6,56	

O ajuste do modelo de mistura de longa duração Weibull 000 foi feito com todos os dados da Tabela 3.3, ignorando o tipo de tratamento de quimioterapia que cada um recebeu.

Inicialmente, sugere-se a possibilidade de uma proporção de imunes diferente para cada tipo de tratamento, ou seja, $p's$ diferentes. Dessa forma, foi ajustado o modelo 100. Porém, a diferença das deviances entre os dois modelos 000 e 100 foi de $0,55$, valor muito pequeno, comparado a $3,84$ (valor da distribuição Qui-quadrado com um grau de liberdade), refutando então a hipótese de que apenas a proporção de imunes nos grupos é diferente.

O próximo modelo ajustado foi 110, indicando que apenas os parâmetros

$\mu's$ seriam igual nos grupos e que os $p's$ e $\beta's$ podem ser diferentes. A diferença das deviances entre os modelos foi de $0,92$, uma mudança não significativa para um parâmetro extra, quando comparada ao valor crítico $3,84$, sugerindo assim o ajuste do modelo 111.

A diferença das deviances entre os modelos 110 e 111 foi de $7,61$, que comparado ao valor $3,84$ é um valor significativo, resultando então que os parâmetros $\mu's$ são diferentes para os grupos.

Para confirmar a diferença encontrada em relação aos parâmetros $\mu's$, foi realizado o ajuste 101, no qual tem-se $p's$ e $\mu's$ diferentes, mas restringe $\beta's$ iguais. A diferença das deviances foi $0,61$, novamente um valor não significativo comparado ao valor crítico $3,84$.

Para encerar o ajuste feito com o modelo 001, considerando $p's$ e $\beta's$ iguais e apenas $\mu's$ diferentes, a diferença das deviances foi de $0,85$ (outro valor não significativo em relação a $3,84$), demonstrando que até mesmo as proporções de imunes nos dois grupos podem ser consideradas iguais.

Analisando diretamente os modelos 000 e 001, a diferença das deviances foi $7,62$ (valor significativo, comparado com $3,84$), o que deixa claro que um modelo apenas com $\mu's$ diferentes é suficiente para descrever o efeito do tratamento. Nenhuma mudança ocorre no ajuste do modelo até que permita $\mu's$ diferentes, sugerindo que os tratamentos diferem na forma de sua distribuição, porém a proporção de imunes e a taxa de risco não apresentam diferença significativa entre os dois tipos de tratamentos.

Capítulo 4

Aplicação em Dados de Reincidência ao Crime

4.1 Sistema Penitenciário

Atualmente, o Brasil administra um dos dez maiores sistemas penais do mundo, com um Sistema Carcerário possuindo aproximadamente 170.000 detentos agrupados em 512 prisões, milhares de delegacias e vários outros estabelecimentos. No entanto, o sistema carcerário é considerado relativamente moderado, pois a taxa aproximada é de 108 presos por 100.000 habitantes e em relação aos países vizinhos, o Brasil encarcera menos pessoas *per capita* que muitos outros países.

Durante o período ditatorial, em virtude das perseguições políticas e da ideologia de manutenção da ordem social, os presídios brasileiros passaram a ter uma superpopulação e culminou, nos anos 80, com os abusos e desrespeito aos presidiários, interpretados como uma das formas mais sérias e crônicas de violação dos direitos humanos no país.

Nos anos 90, especificamente em 1997, ocorreram nos grandes presídios brasileiros várias rebeliões dramáticas e episódios com reféns e mortes. Uma fiscalização internacional contínua apontou para a garantia dos direitos humanos aos presos e o cumprimento da Lei de Execução Penal.

A Lei de Execução Penal adotada em 1984 reconhece um respeito saudável aos direitos humanos dos presos e contém várias provisões ordenando o tratamento individualizado, protegendo os direitos substantivos e processuais dos presos e garantindo a assistência médica, jurídica educacional, social, religiosa

e material. É uma Lei que não visa a punição, mas a "ressocialização das pessoas condenadas" .

Estudos e o senso comum apontam que as prisões são espaços de exercício da violência e aprendizagem do crime o que inviabiliza a reintegração do preso à sociedade.

O país possui um índice elevado de reincidência ao crime, segundo dados do governo federal, a média nacional do índice de reincidência de ex-presidiários é de 82%. Estados como Minas Gerais e Pernambuco apresentam as médias de 48% e 45%, enquanto que no Paraná a taxa é de 30%, figurando como um dos menores índices de reincidência ao crime no Brasil.

Diante do quadro de dificuldades que assola o Sistema Penitenciário, o Estado do Paraná vem apontando soluções que perpassam pela infra-estrutura e pelo resgate de valores humanos indispensáveis à vida saudável. Isso se deve ao modelo de reclusão implantado no Estado nos últimos anos, pois, atualmente, 72% dos 5.319 internos do Sistema Penitenciário Paranaense têm alguma ocupação nos presídios, quer seja na prestação de serviços ou na realização de cursos profissionalizantes.

Para ressocializar os presos, o Sistema Penitenciário do Paraná vem adotando políticas que contemplam a construção e a melhoria da infra-estrutura dos espaços de detenção e o resgate de valores humanos. Aos presos do Paraná são oportunizadas as seguintes condições: a)trabalho; b)estudo; c)cumprimento da pena na região de domicílio.

Nesse sentido, o presidiário trabalha em regime de indústria, como empregado e sem os estigmas gerados pela prisão e a possibilidade de com a pena cumprida permanecer na mesma indústria. Estudando, o presidiário tem a possibilidade de profissionalizar-se para manter-se no mercado de trabalho ou preparar-se para o enfrentamento do mesmo após o cumprimento da pena. A proximidade com as famílias, durante o período de reclusão, não quebra os laços familiares, pois a atenção da família, a afetividade e a convivência contribuem para a recuperação mais humana dos presidiários.

O Paraná investe em torno de R$ 500 mil em programas de recuperação

que incentivam o processo de reintegração do preso à sociedade. O Estado oferece cursos de capacitação profissional nas áreas da construção civil, da informática, da panificação, do artesanato e da jardinagem, além disso, proporciona a oportunidade aos detentos ou ex-detentos de completar o ensino fundamental e médio, a trabalhar e profissionalizar-se.

O sistema possui o Patronato Penitenciário que acompanha a recuperação de presos para reintegrá-los à sociedade e proporciona aos presidiários trabalhos de assistentes de obra e nas horas vagas em serviço no próprio Patronato.

O Patronato em parceria com o Programa Pró-egresso encaminha, em média, 104 presos, oriundos das unidades penitenciárias do Paraná, para o mercado de trabalho.

Nesse modelo de reclusão é preciso destacar, também, o papel da sociedade civil organizada, pois as ações educativas e de profissionalização são realizadas em parceria com a Secretaria de Estado da Educação, SESC, SENAC, SENAI, SENAR e com empresários que montam as suas fábricas e as estruturas de cursos no interior das unidades penais, investindo na mão-de-obra e na capacidade dos presos.

As entidades religiosas fornecem contribuições espirituais e morais, ressaltando a vida e o lado bom que todos os seres humanos possuem.

Os servidores públicos e os funcionários do setor privado que prestam serviços no Sistema Penitenciário Paranaense recebem atenção especial no que diz respeito à capacitação profissional através de cursos, seminários, palestras e encontros realizados na Escola Penitenciária. Também, iniciou-se um curso de pós-graduação, no âmbito nacional, na Universidade Federal do Paraná, com o apoio do Ministério da Justiça.

> "Nossa visão é de que, diante das dificuldades nacionais e internacionais, temos de fazer a nossa lição de casa: buscar alternativas políticas, legais, econômicas, sociais e técnicas que elevem o ser humano preso" (Tavares, 2000).

4.2 Caracterização da PEM

A Penitenciária Estadual de Maringá é um estabelecimento de segurança máxima e destina-se ao atendimento de presos do sexo masculino que cumprem pena em regime fechado, com capacidade de atendimento para 360 presos. Localiza-se na área agrícola de Maringá -Paraná, no limite entre os municípios de Maringá e Paiçandu.

A unidade possui uma área construída de 5.800 metros quadrados num terreno de 24 mil metros quadrados, possui 60 celas com capacidade para 6 presos cada.

É dotada de infra-estrutura com guaritas, galerias, solários, refeitórios, salas de aula, salas de atendimento, cozinha, panificadora, lavanderia, consultório médico, odontológico e área íntima.

A população carcerária, atualmente, é de 345 presos com idade entre 18 e 90 anos. Desse total, 55% empregam a mão-de-obra em 24 canteiros de trabalho nas áreas de manutenção, artesanato, agricultura, fábrica de bolas de futebol, marcenaria e outros.

Compete a PEM:

1. **I** - a segurança e a custódia dos presos do sexo masculino que se encontram internados no estabelecimento, por decisão judicial, em cumprimento de pena em regime fechado;

   II - a promoção da reintegração social dos internos e o zelo pelo seu bem-estar, através da profissionalização, educação, prestação de assistência jurídica, psicológica, social, médica, odontológica, religiosa e material;

   III - a prestação de assistência social aos familiares dos internos.

São desenvolvidas as ações educativas e formativas em parceria com o SENAR, SENAC, SENAI, SESC, SEBRAE, SEED -Secretaria de Educação do Estado do Paraná-, empresários e religiosos. Dentre as atividades desenvolvidas destacam-se:

a) a produção de mudas de árvores nativas;

b) o curso de informática;

c) os cursos de artes;

d) os cursos de educação de jovens e adultos;

e) as palestras de prevenção da DST/AIDS;

f) o programa para a terceira idade;

g) a assistência religiosa.

Para constituir o banco de dados foram tomados os detentos que passaram pela PEM até 31 de dezembro de 2002 (data da nossa censura). Coletou-se a data de nascimento do indivíduo, o enquadramento do delito que levou a sua condenação, a data da entrada na PEM, a data da saída da PEM, o tipo de benefício que recebeu para sair e a data da reincidência no sistema penitenciário do Paraná, no caso de ser reincidente. Também, fizeram parte do banco de dados a participação dos detentos em cursos de formação profissionalizantes oferecidos pela PEM, em parceria com a comunidade e a educação formal, que vai desde primeira fase do ensino fundamental até o ensino médio.

Ressalta-se aqui que quando foi selecionada a covariável "enquadramento do delito cometido", considera-se o enquadramento dos artigos do código penal que o ex-detento cometeu para ser preso e condenado, pois a maioria infligiu mais de um artigo antes de serem condenados.

Quanto ao enquadramento do delito tem-se:

a) **Crimes Contra o Patrimônio**:

Furto:- Art. 155 - Subtrair, para si ou para outrem, coisa alheia móvel.

Roubo: Art. 157 - Subtrair coisa móvel alheia, para si ou para outrem, mediante grave ameaça ou violência a pessoa, ou depois de havê-la, por qualquer meio, reduzido à impossibilidade de resistência.

Extorsão: Art. 158 - Constranger alguém, mediante violência ou grave ameaça, e com o intuito de obter para si ou para outrem indevida vantagem econômica, a fazer, tolerar que se faça ou deixar fazer alguma coisa.

Extorsão mediante seqüestro: Art. 159 - Seqüestrar pessoa a fim de obter, para si ou para outrem, qualquer vantagem, com condição ou preço de resgate.

Extorção indireta: Art. 160 - Exigir ou receber, como garantia de dívida, abusando da situação de alguém, documento que pode dar causa a procedimento criminal contra a vítima ou contra terceiro.

Dano: Art. 163 - Destruir, inutilizar ou deteriorar coisa alheia.

Estelionato: Art. 171- Obter, para si ou para outrem, vantagem ilícita, em prejuízo alheio, induzindo ou mantendo alguém em erro, mediante artifício, ardil, ou qualquer outro meio fraudulento.

Receptação: Art. 180 - Adquirir, receber, transportar, conduzir ou ocultar, em proveito próprio ou alheio, coisa que sabe ser produto de crime, ou influi para que terceiro, de boa-fé, a adquira, receba ou oculte.

b) **Crimes Contra a Pessoa**:

Homicídio simples: Art. 121 - Matar alguém.

Aborto provocado pela gestante ou com se consentimento: Art. 124 - Provocar Aborto em si mesma ou consentir que outrem lho provoque.

Lesão corporal: Art. 129 - Ofender a integridade corporal ou a saúde outrem.

c) **Crimes Contra a Liberdade Sexual**:

Estupro: Art. 213 - Constranger mulher à conjunção carnal, mediante violência ou grave ameaça.

Atentado violento ao pudor: Art. 214 - Constranger alguém, mediante violência ou grave ameaça, a praticar ou permitir que com ele se pratique ato libidinoso diverso da conjunção carnal.

Corrupção de menores: Art. 218 - Corromper ou facilitar a corrupção de pessoas de maior de 14 (catorze) e menor de 18 (dezoito) anos, com ela praticando ato de libidinagem, ou induzindo-a a praticá-lo ou presenciá-lo.

Presunção de violência: Art. 224 - Presume-se a violência, se a vítima:

i) não é de maior de 14 (catorze) anos;

ii) é alienada ou débil mental, e o agente conhecia esta circunstância;

iii) não pode, por qualquer outra causa, oferecer resistência.

d) **Crimes Contra a fé Pública**:

Moeda falsa: Art. 289 - Falsificar, fabricando-a ou alterando-a, moeda metálica ou papel-moeda de curso legal no país ou no estrangeiro.

Falsificação de documento público: Art. 297 - Falsificar, no todo ou em parte, documento público, ou alterar documento público verdadeiro.

Falsificação de documento particular: Art. 298 - Falsificar, no todo ou em parte, documento particular ou alterar documento verdadeiro.

e) **Tráfico:**

Artigo 12 - Lei 6368: Tráfico de drogas.

Artigo 16 - Lei 6368: Uso de entorpecentes.

f) **Porte de arma:**

Artigo 10 - Lei 9437: Porte de arma ilegal.

Em relação ao benefício adquirido ao sair da PEM:

a) Liberdade Definitiva;

b) Liberdade Condicional;

c) Regime Aberto;

d) Regime Semi-Aberto (colônia penitencial agrícola - CPA);

e) Indulto.

Na educação formal ofertados pela SEED:

a) Conclusão da 1ª fase do ensino fundamental;

b) Conclusão da 1ª fase do ensino fundamental e em curso a 2ª fase do ensino fundamental;

c) Conclusão da 2ª fase do ensino fundamental;

d) Conclusão da 1ª e 2ª fases do ensino fundamental e em curso o ensino médio;

e) Conclusão da 2ª fase do ensino fundamental e em curso o ensino médio;

f) Conclusão do ensino médio;

g) Conclusão da 1ª e 2ª fases do ensino fundamental e em curso o ensino médio;

h) Conclusão da 1ª e 2ª fases do ensino fundamental e conclusão do ensino médio;

i) Não estudou na PEM.

Os cursos profissionalizantes oferecidos em parcerias com SENAR, SENAC, SENAI, UEM, SEBRAE e por voluntários no período de 1996 a 2002 foram:

a) Ofertados pelo **SENAC**: Serigrafia; Básico de Confeiteiro; Cozinha Executiva; Corte de Cabelos; Técnicas de Preparo de Pizzas; Técnicas de Serviços para Garçon; Técnicas de Serviço para restaurante; Técnicas de Preparo de lanches; Técnicas de Confeiteiro; Primeiros Socorros; Básico de Cozinha.

b) Ofertados pelo **SENAR**: Cultivo e Padronização de legumes; Corte de cana de açúcar; Plasticultura - Construção de Estufas; Produção de hortaliças em estufas; Casqueamento de bovinos; Conservas caseiras; Operações e Manutenção de Tratores; Aplicação de Defensivos; Produção de Mudas - Básico Citros e Café; Derivados de Leite; Produção de Mudas -

Viveiros; Olericultura Básica - Cultivo e Padronização de Hortaliças; Jardinagem; Olericultura - Raízes, Bulbos e Tubérculos; Olericultura - Frutos e Sementes; Fruticultura Tropical; Fruticultura temperada - Básica; Fruticiltura temperada - Morango; Fruticultura temperada - uva para mesa.

c) Ofertados pelo **SENAI**: Eletricista Instalados Predial; Eletricista Bobinador; Pintor de Obras; Eletricista de Automóveis; Pedreiro; Azulejista; Borbinador de Motores.

d) Ofertados pelo **SEBRAE**: Orientação para o Crédito - Brasil Empreendedor; Aprendendo a Empreender.

e) Ofertado pela **UEM**: Restauração de Material Bibliográfico.

f) Ofertados por **Voluntários**:Técnica de Pintura em Relevo; Decoração de Festas Infantis; Corte de Cabelos; Pintura Decorativa; Fabricação de Materiais de Limpeza; Bordados em Ponto Cruz; Técnica de Tingimento de Tecidos; Bordados em Pedrarias; Produção de derivados de Milho.

Algumas Características dos Ex-Detentos

Nos 1172 ex-detentos que fazem parte do banco de dados a ser analisado, constatou-se que:

a) Em relação à profissão que exerciam antes de serem condenados, as que mais se destacaram foram: 11% pedreiros; 19% eram trabalhadores rurais (vulgo bóia-fria); 4% mecânicos; 7% motoristas; 7% pintores; 2% ensacadores; 3% auxiliares de serviços gerais; 2% tratoristas; 3% comerciantes e 2% marceneiros. Também, foi detectado algumas profissões fora do comum como: policial, professor, decorador, atleta e segurança.

b) Dos benefícios adquiridos para sair da PEM: $49,23\%$ saíram em Regime Semi-aberto (CPA); $34,64\%$ saíram em Liberdade Condicional; $9,56\%$ saíram em Liberdade Definitiva; $5,63\%$ saíram por Indulto e $0,94\%$ em Regime Aberto.

c) Quanto ao enquadramento do delito cometido para a condenação: 43% foram condenados por infligirem artigos em relação a Crimes contra o Patrimônio; 8% em relação a Crimes contra a Pessoa; 5% em relação a Crimes contra a Liberdade Sexual; $0,1\%$ em relação a Crimes contra Fé Pública; 20% foram condenados como traficantes de drogas; 7% foram condenados como traficantes e infligiram artigos de crimes contra o Patrimônio; 3% crimes contra a Fé Pública e contra o Patrimônio; 8% crimes contra a Pessoa e Patrimônio; 2% crimes contra Liberdade Sexual e patrimônio; a porcentagem restante tiveram várias combinações entre o enquadramento dos crimes cuja uma proporção teve enquadramento em todos os tipos de crime citados.

d) Dos cursos profissionalizantes ofertados constatou-se que: 62% dos ex-detentos não participaram de curso oferecido; 14% fizeram apenas um curso; 9% fizeram dois cursos profissionalizantes; 5% fizeram três cursos; 4% fizeram quatro cursos ofertados; 2% fizeram cinco cursos; 1% aperfeiçoou com seis cursos; os demais 3% fizeram de sete a dezesseis cursos profissionalizantes.

e) Da educação formal: 26% dos ex- detentos concluíram a 1ª fase do ensino fundamental; 9% concluíram a 1ª fase do ensino fundamental e foram matriculados na 2ª fase do ensino fundamental; 19% concluíram a 2ª fase do ensino fundamental; 6% concluíram a 1ª e 2ª fases do ensino fundamental e foram matriculados no ensino médio; 5% concluíram a 2ª fase do ensino fundamental e foram matriculados no ensino médio; 4% concluíram a 1ª e 2ª fases do ensino fundamental e também o ensino médio; 31% não estudaram durante o período que permaneceram na PEM.

4.3 Aplicação de Modelos de Mistura de Longa Duração

Sendo o evento de interesse a reincidência de ex-detentos, a variável resposta é o tempo de liberdade dos ex-detentos. Para determinar o tempo de liberdade foi comparado o dia em que o ex-detento adquiriu o benefício até

a data de 31 de dezembro de 2002, data de censura. Para alguns ex-detentos não foi necessária a data de censura, pois eles já haviam retornado ao sistema penitenciário do estado do Paraná antes da mesma, constituindo assim observação completa.

A primeira análise foi ajustar os dados em um modelo de distribuição de probabilidade, trabalhando com a variável resposta tempo de liberdade e um indicador de censura. O ex-detento recebia indicador de censura 1 se fosse reincidente no sistema penitenciário até a data de censura, caso contrário recebia 0 (zero).

Como os dados não apresentavam uma função de risco constante, o modelo proposto foi o Weibull.

No primeiro ajuste, foi comprovado que dos 1172 ex-detentos, apenas $22,44\%$ que corresponde a 263 ex-detentos, tinham observações completas e $77,56\%$ que corresponde a 909 ex-detentos, eram observações censuradas, fornecendo forte indício de que o modelo de mistura de longa duração seria o mais apropriado para descrever o conjunto dos ex-detentos do sistema penitencário da PEM, pois o mesmo apresentava um número muito grande de observações censuradas.

As modelagens para o modelo padrão Weibull e para o modelo de mistura de longa duração Weibull forneceram os estimadores dos parâmetros, como mostra as Tabelas 4.1 e 4.2.

Tabela 4.1: Modelo Weibull padrão.

Parâmetros Estimados	Erros Padrões
$\hat{\beta}_0 = 1,0779$	$\pm 0,0577$
$\hat{\mu}_0 = 4299,2098$	$\pm 360,7034$

Para comparar a qualidade dos ajustes dos modelos, será utilizado o teste da diferença das deviances. As hipóteses a serem testadas serão: $H_0 : p = 1$ versus $H_1 : p \neq 1$. A não rejeição da hipótese H_0 indica que não existe evidência de que o modelo de mistura de longa duração Weibull ajusta-se melhor que o modelo Weibull padrão (não há evidência da existência de indivíduos imunes

Tabela 4.2: Modelo de mistura de longa duração Weibull.

Parâmetros Estimados	Erros Padrões
$\hat{q} = 0,3315$	$\pm 0,0199$
$\hat{\beta} = 1,6118$	$\pm 0,0913$
$\hat{\mu} = 942,9510$	$\pm 57,7382$

à reincidência). Em caso contrário, haverá evidência de ex-detento imune à reincidência ao crime.

Os valores das deviances para os dois modelos encontrados são: $2\log H_0 = -2486,072$; $2\log H_1 = -2462,312$.

$$\Lambda = [-2462,312 - (-2486,072)] = 23,759.$$

Como $23,759$ é um valor grande, comparado ao valor tabelado $3,84$ da distribuição de qui quadrado com um grau de liberdade, conclui-se que a hipótese H_0 é rejeitada. Ou seja, existe evidência de que parte dos ex-detentos da PEM não reincidirão ao crime. Essa proporção estimada é de 33%.

4.3.1 Influência das Covariáveis

Verificado que o modelo de mistura de longa duração ajusta-se melhor ao conjunto de dados dos ex-detentos da PEM faz-se necessário analisar se existe alguma covariável: Delito Cometido; Benefício Adquirido; Número de Cursos; Formação Escolar, que possa influenciar no ajuste do modelo obtido.

Como as covariávies apresentam muitas categorias e em algumas delas não existe observações repetida, faz-se necessário a agregação das categorias.

Para a análise da covariável Delito Cometido as categorias ficarm da seguinte maneira: um delito (tráfico; crime contra o patrimônio; crime contra a pessoa; crime contra a fé pública; crime contra liberdade sexual; porte de arma); dois delitos cometidos (tráfico/pessoa; tráfico/liberdade sexual; tráfico/fé publica; tráfico/patrimônio; pessoa/liberdade sexual; pessoa/fé pública; pessoa/patrimônio;

liberdade sexual/fé pública; liberdade sexual/patrimônio; fé pública/patrimônio; arma/tráfico; arma/patrimônio; arma/liberdade sexual) ; três delitos (tráfico/patrimônio/ sexual; tráfico/patrimônio/liberdade sexual; tráfico/patrimônio/fé pública; tráfico/patrim tráfico/pessoa/liberdade sexual; tráfico/pessoa/fé pública; tráfico/pessoa/arma; tráfico/fé pública/arma; tráfico/patrimônio/arma; tráfico/pessoa/arma; tráfico/fé pública/liberdade sexual. Para a realização da análise faz-se necessário agregar categorias: apenas um delito; dois delitoscometidos; três ou mais delitos.

Na análise da covariável Benefício adquirido as categorias ficaram: liberdade definitiva; liberdade condicional; regime semi-aberto; outros (regime aberto e indulto).

A covariavél número de cursos que o ex-detento realizou no decorrer do tempo que permaneceu na PEM está dividida seguintes categorias: de nenhum curso; apenas 1 curso; dois cursos; três cursos; quatro cursos; cinco cursos; mais de cinco cursos (6, 7, 8, 9, 10, 11, 12, 13, 14, 15 e 16).

Em relação a covariável Formação escolar, que trata-se do ensino que o ex-detento recebeu na PEM através do Centro de Ensino Supletivo (CES), realizou-se a seguinte agregação das categorias: não estudou; conclusão da 1ª fase do ensino fundamental (correspondente da 1ª a 4ª série - antigo primário); conclusão da 2ª fase do ensino fundamental (correspondente da 5ª a 8ª série -antigo ginásio); conclusão da 1ª fase do ensino fundamental e em curso a 2ª fase do ensino fundamental; conclusão do ensino fundamental (1ª e 2ª fases); outros (conclusão do ensino médio, conclusão da 2ª fase do ensino fundamental e em curso o ensino médio, conclusão da 2ª fase do ensino fundamental e do ensino médio, conclusão do ensino fundamental (1ª e 2ª fases) e conclusão do ensino médio.

A Tabela 4.3 apresenta as deviances dos modelos de mistura de longa duração Weibull na presença de cada covariável.

Comparando os valores das deviances de cada modelo ajustado na presença de covariáveis com o modelo de mistura de longa duração Weibull, verifica-se que os modelos na presença de cada covariável apresentam diferenças significativas em relação ao modelo de mistura Weibull sem a presença de covariável, sendo assim pode-se afirmar que todas as covariáveis influenciam no modelo

Tabela 4.3: Modelo de mistura de longa duração Weibull com covariáveis.

Modelagem com a covariável	deviance	grau de liberdade	valor crítico
Delito cometido	$-2446,47$	2	$5,99$
Benefício adquirido	$-2432,27$	3	$7,81$
Número de cursos	$-2425,51$	6	$12,59$
Formação escolar	$-2446,06$	6	$12,59$

ajustado.

Apesar da análise em relação as covariavés ter sido significativa, neste trabalho somente será analisada a covariável Benefício Adquirido, pois as demais covariáveis apresentam um grande número de categorias agregadas.

Na covariável Benefício Adquirido pelos ex-detentos ao sair da PEM, foi constatado que:

a) dos 1172 ex-detentos 406 receberam o beneficio Liberdade Condicional apresentaram tempo médio de liberdade de 912 dias com desvio padrão de 619;

b) receberam o benefício Liberdade Definitiva 112 indivíduos apresentando tempo médio de liberdade de 1373 dias e desvio padrão de 738;

c) para o benefício Regime Semi-Aberto foram 577 indivíduos com tempo médio de liberdade de 1093 dias e desvio padrão 674;

d) enquadraram no benefício Outros (Regime Aberto e Indulto): 77 indivíduos que apresentaram tempo médio de liberdade de 1073 dias e desvio padrão 807.

Para cada tipo de benefício foi ajustado o modelo Weibull padrão e o modelo de mistura de longa duração Weibull.

4.3.2 Liberdade Condicional

O grupo do ex-detentos que receberam o benefício de liberdade condicional

era composto por 406 indivíduos dos quais apenas $9,36\%$ tinham observações completas e o restante $90,64\%$ eram observações censuradas.

Os parâmetros do modelo Weibull padrão e seus erros padrões para o grupo da Liberdade Condicional são apresentados na Tabela 4.4.

Tabela 4.4: Modelo Weibull padrão.

Parâmetros Estimados	Erros padrões
$\hat{\beta}_0 = 1,1559$	$\pm 0,1601$
$\hat{\mu}_0 = 7312,1546$	$\pm 2070,3359$

As estimativas e seus erros padrões para o modelo de mistura de longa duração Weibull são apresentados na Tabela 4.5

Tabela 4.5: Modelo de mistura de longa duração Weibull.

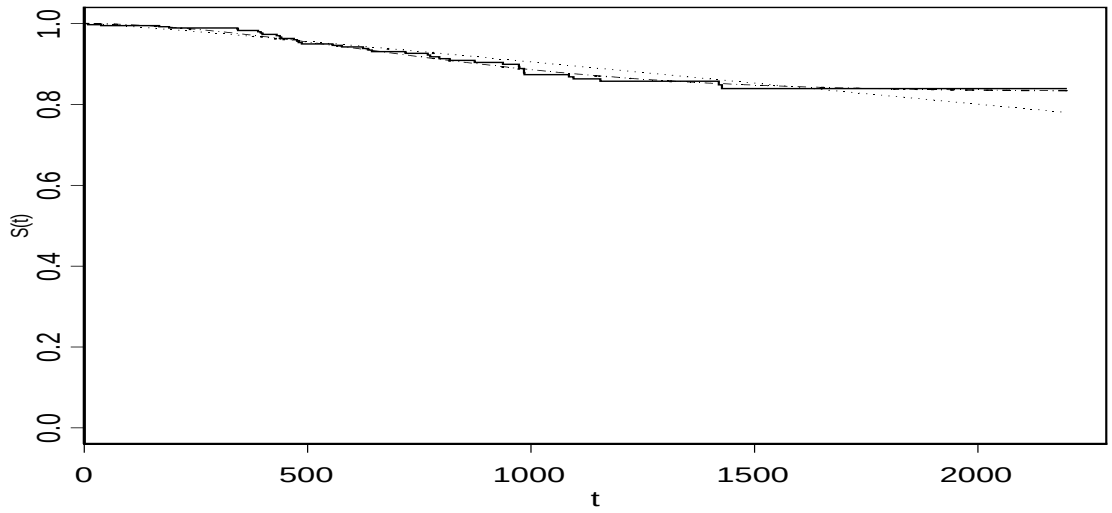
Parâmetros Estimados	Erros Padrões
$\hat{q} = 0,1670$	$\pm 0,0298$
$\hat{\beta}_1 = 1,8484$	$\pm 0,2957$
$\hat{\mu}_1 = 927,9567$	$\pm 141,6423$

A Figura 4.1 mostra os ajustes das curvas de sobrevivência dos modelos ajustados juntamente com a estimativa de Kaplan-Meier.

A comparação da qualidade dos ajustes dos modelos poderia ser feita apenas visualmente, tendo em vista que o modelo mais bem ajustado aos dados será aquele cujos pontos da função de sobrevivência estiverem mais próximos dos valores obtidos pela estimativa de Kaplan-Meier, porém para não correr o risco de fazer conclusões precipitadas, será utilizado o teste da diferença das deviances.

O grupo dos ex-detentos que receberam o benefício Liberdade Condicional tiveram os valores de deviances apresentados na Tabela 4.6

Testando a hipótese $H_0 : p = 1$ versus $H_1 : p \neq 1$, para verificar se no grupo do benefício Liberdade Condicional existe evidência de elementos imune a reincidência, calculou-se a diferença das deviances



(a)

Figura 4.1: Curvas de sobrevivência. (——): Kaplan-Meier, ($\cdots\cdots$): Modelo Weibull Padrão e ($\cdot - \cdot - \cdot$): Modelo de mistura de longa duração Weibull.

Tabela 4.6: Liberdade Condicional.

Modelo	Deviance
Weibull Padrão	$L\left(\hat{\mu}_0, \hat{\beta}_0\right) = -386,4677$
Mistura de longa duração Weibull	$L\left(\hat{p}, \hat{\mu}_1, \hat{\beta}_1\right) = -383,2211$

$$\Lambda = 2\left[-383,2211 - (-386,4677)\right] = 6,4913,$$

ao comparar o valor de Λ com o valor de qui-quadrado com um grau de liberdade $(3,84)$, conclui-se que H_0 não é plausível. Pode-se, então, afirmar que existe evidência de indivíduos imunes no grupo dos ex-detentos que receberam o benefício Liberdade Condicional. Assim o modelo de mistura de longa duração Weibull está melhor ajustado e por este modelo, a proporção de ex-detentos que não reincidirão ao crime é de aproximadamente 17%.

4.3.3 Liberdade Definitiva

Dos 112 ex-detentos que receberam o benefício liberdade definitiva $17,86\%$ tinham observações completas e $82,14\%$ foram observações censuradas.

Para o ajuste do modelo Weibull padrão, os estimadores dos parâmetros e as estimativas dos seus erros padrões são apresentados na Tabela 4.7

Tabela 4.7: Modelo Weibull padrão.

Parâmetros Estimados	Erros Padrões
$\hat{\beta}_0 = 0,8764$	$\pm 0,1808$
$\hat{\mu}_0 = 9599,5602$	$\pm 4352,3228$

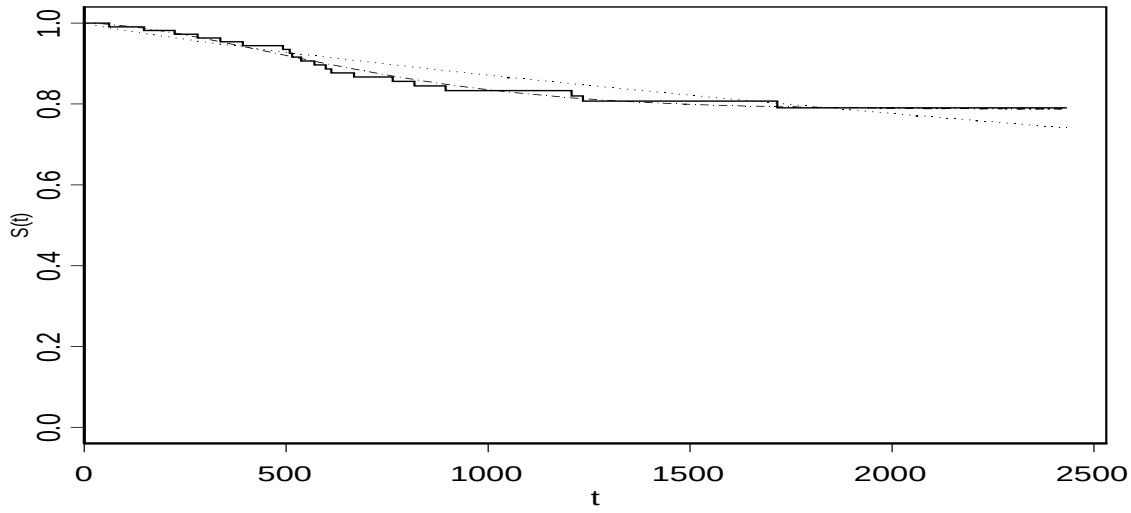
Para o modelo de mistura de longa duração Weibull ajustado, os estimadores de máxima verossimilhança com seus respectivos erros padrões são apresentados na Tabela 4.8

Tabela 4.8: Modelo de mistura de longa duração Weibull.

Parâmetros Estimados	Erros Padrões
$\hat{q} = 0,2122$	$\pm 0,0442$
$\hat{\beta}_1 = 1,6630$	$\pm 0,3186$
$\hat{\mu}_1 = 783,7475$	$\pm 129,1056$

A Figura 4.2 mostra as curvas de sobrevivência do benefício Liberdade

Condicional para os modelos ajustados juntamente com a estimativa de Kaplan-Meier.



(a)

Figura 4.2: Curvas de sobrevivência. (——): Kaplan-Meier, ($\cdots\cdots$): Modelo Weibull Padrão e ($\cdot - \cdot - \cdot$): Modelo de mistura de longa duração Weibull.

Para o grupo dos ex-detentos com benefício Liberdade Definitiva, os valores das deviances são apresentados na Tabela 4.9

Tabela 4.9: Liberdade Definitiva.

Modelo	Deviance
Weibull Padrão	$L\left(\hat{\mu}_0, \hat{\beta}_0\right) = -198,7370$
Mistura de longa duração Weibull	$L\left(\hat{p}, \hat{\mu}_1, \hat{\beta}_1\right) = -194,6454$

Novamente, testando as hipóteses a respeito da existência de proporção de imunes ao crime, tem-se que o valor da diferença das deviances é

$$\Lambda = 2\left[-194,6454 - (-198,7370)\right] = 8,1832.$$

Comparando Λ com o valor de qui quadrado com um grau de liberdade, concluiu-se que a hipótese H_0 não é plausível e, então, há evidência de in-

divíduos imunes ao crime no grupo dos ex-detentos que receberam o benefício Liberdade Definitiva; e a proporção dos ex-detentos que não reincidirão ao crime é estimada em 21%.

4.3.4 Regime Semi-Aberto

Dos 577 indivíduos que receberam o benefício Regime Semi-Aberto $33,45\%$ eram observações completas e $66,55\%$ observações censuradas.

Para o modelo Weibull padrão os parâmetros estimados via máxima verossimilhança e seus respectivos erros padrões são apresentados na Tabela 4.10.

Tabela 4.10: Modelo Weibull padrão.

Parâmetros Estimados	Erros Padrões
$\hat{\beta}_0 = 1,0755$	$\pm 0,0674$
$\hat{\mu}_0 = 3072,6459$	$\pm 257,6910$

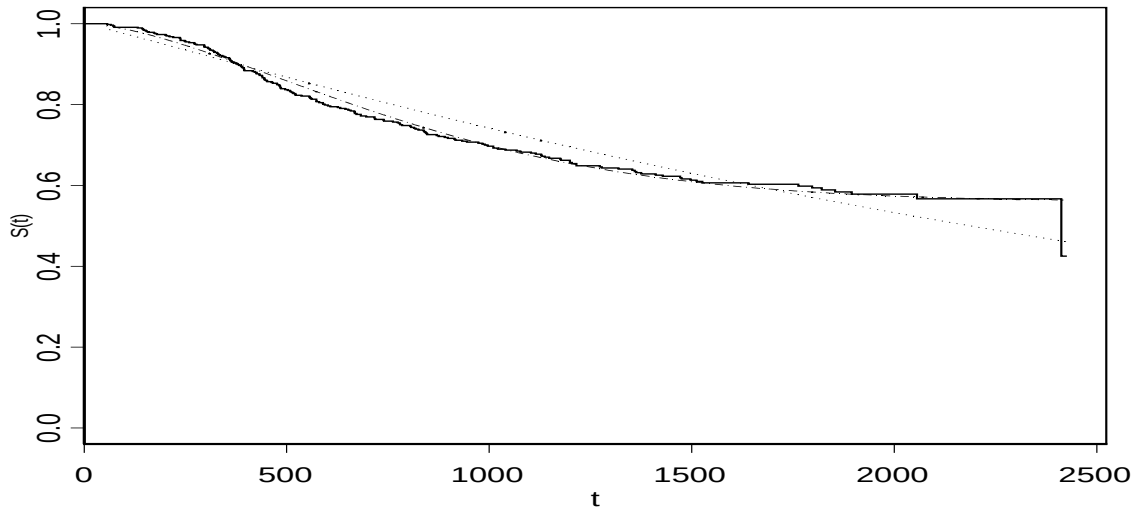
Para o modelo de mistura de longa duração Weibull, os parâmetros estimados via verossimilhança e seus erros padrões são apresentados na Tabela 4.11

Tabela 4.11: Modelo de mistura de longa duração Weibull.

Parâmetros Estimados	Erros Padrões
$\hat{q} = 0,440464$	$\pm 0,0279$
$\hat{\beta}_1 = 1,576968$	$\pm 0,1031$
$\hat{\mu}_1 = 912,834224$	$\pm 62,7280$

A Figura 4.3 mostra as curvas de sobrevivência para os modelos ajustados juntamente com a estimativa de Kaplan-Meier, do grupo dos ex-detentos que receberam o benefício Regime Semi-Aberto.

Para o grupo dos ex-detentos que receberam o benefício Regime Semi-Aberto os valores das deviances dos dois modelos ajustados são apresentados na Tabela 4.12



(a)

Figura 4.3: Curvas de sobrevivência. (——): Kaplan-Meier, ($\cdots\cdots$): Modelo Weibull Padrão e ($\cdot - \cdot - \cdot$): Modelo de mistura de longa duração Weibull.

Tabela 4.12: Regime Semi-Aberto

Modelo	Deviance
Weibull Padrão	$L\left(\hat{\mu}_0, \hat{\beta}_0\right) = -1754,0680$
Mistura de longa duração Weibull	$L\left(\hat{p}, \hat{\mu}_1, \hat{\beta}_1\right) = -1736,6952$

Testando as hipóteses $H_0 : p = 1$ versus $H_1 : p \neq 1$, encontrou-se o valor da diferença das deviances

$$\Lambda = 2\left[-1736,6952 - (-1754,0680)\right] = 34,7456.$$

Como o valor de Λ é muito grande, concluiu-se que a hipótese H_1 é aceita. Ou seja, existe evidência de que parte dos ex-detentos que receberam o benefício Regime Semi-Aberto não reincidirão ao crime. Essa proporção é estimada em 44%.

4.3.5 Outros benefícios (Indulto e Regime Aberto)

Para o grupo dos Outros, $15,58\%$ tinham observações completas e as observações censuradas perfaziam um total de $84,42\%$.

O ajuste do modelo Weibull padrão forneceu os seguintes parâmetros estimados com seus erros padrões apresentados na Tabela 4.13

Tabela 4.13: Modelo Weibull padrão.

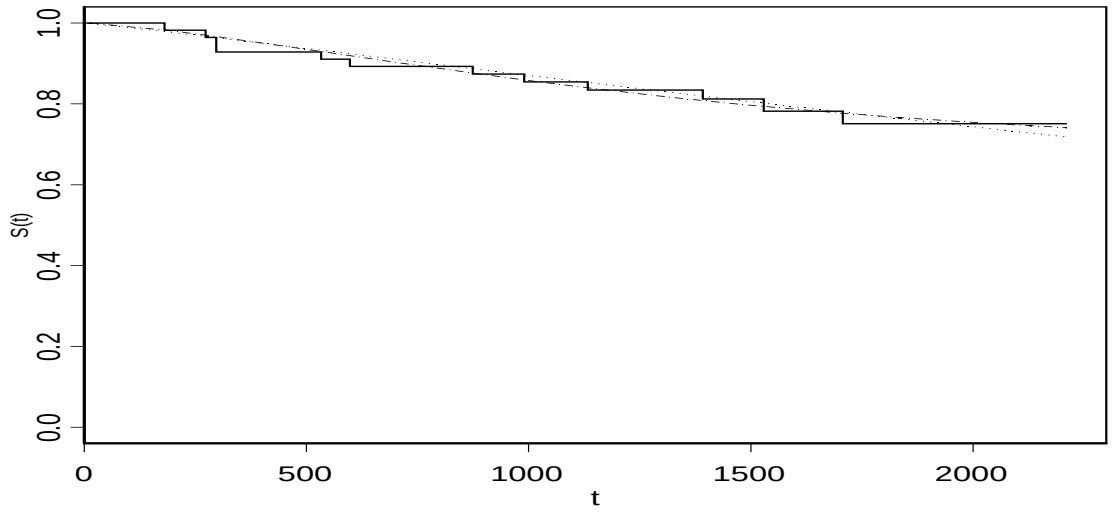
Parâmetros Estimados	Erros Padrões
$\hat{\beta}_0 = 1,0942$	$\pm 0,2936$
$\hat{\mu}_0 = 6070,5446$	$\pm 2709,1267$

O ajuste do modelo de mistura de longa duração Weibull forneceu os parâmetros estimados via verossimilhança, juntamente com as estimativas de seus erros padrões apresentados na Tabela 4.14.

Tabela 4.14: Modelo de mistura de longa duração Weibull.

Parâmetros Estimados	Erros Padrões
$\hat{q} = 0,3051$	$\pm 0,1668$
$\hat{\beta}_1 = 1,3837$	$\pm 0,4786$
$\hat{\mu}_1 = 1399,4594$	$\pm 984,6030$

A Figura 4.4 mostra as curvas de sobrevivência dos modelos ajustados juntamente com a estimativa de Kaplan-Meier para o grupo de ex-detentos que receberam outros benefícios.



(a)

Figura 4.4: Curvas de sobrevivência. (——): Kaplan-Meier, ($\cdots\cdots$): Modelo Weibull Padrão e ($\cdot - \cdot - \cdot$): Modelo de mistura de longa duração Weibull.

As deviances dos modelos Weibull padrão e mistura Weibull para o grupo de ex-detentos que receberam o benefício Outros são apresentados na Tabela 4.15

Tabela 4.15: Regime Aberto e Indulto (outros).

Modelo	Deviance
Weibull Padrão	$L\left(\hat{\mu}_0, \hat{\beta}_0\right) = -117,9921$
Mistura de longa duração Weibull	$L\left(\hat{p}, \hat{\mu}_1, \hat{\beta}_1\right) = -117,7086$

Calculando a diferença das deviances encontrou-se

$$\Lambda = 2\left[-117,7086 - (-117,9921)\right] = 0,567$$

Nesse grupo o teste aceita a hipótese H_0, concluindo que não existe evidência de que o grupo dos ex-detentos que receberam o benefício Regime Aberto e

Indulto (outros) tem uma proporção de ex-detentos imunes à reincidência ao crime.

Ressalta-se que com uso do modelo de mistura de longa duração Weibull não ficou evidente a existência de uma proporção de imunes no conjunto de ex-detentos da PEM, porém outro modelo poderia ter sido utilizado, tal como o modelo Log logística.

4.3.6 Proporção de Imunes por Tipo de Benefício

No ajuste do modelo de mistura de longa duração Weibull considerando a covariável benefício adquirido a proporção de ex-detentos que não reincidirão ao crime é estimada em 33%, entretanto considerando os níveis do benefícios adquiridos constata-se que os grupos dos ex-detentos que receberam os benefícios: Liberdade Condicional, Liberdade Definitiva e Regime Semi Aberto apresentaram evidências de que uma parte deles não reincidirão ao crime, e as estimativas das proporções de acordo com o benefício recebido ao sair da PEM são apresentadas na Tabela 4.16.

Tabela 4.16: Estimativa da proporção dos imunes de acordo com o tipo de benefício.

Tipo de Benefício	Estimativa de imunes
Liberdade Condicional	17%
Liberdade Definitiva	21%
Regime Semi Aberto	44%

Dos 406 ex-detentos que receberam o benefício Liberdade Condicional 69 deles não reincidirão ao crime, dos 112 que receberam o benefício Liberdade Definitiva 24 deles não reincidirão e para os 577 que adquiriram Regime Semi Aberto 254 não reincidirão ao crime.

Capítulo 5

Considerações Finais

Realizou-se, neste trabalho, o estudo da probabilidade de sobrevivência como uma função do tempo, abordando os tipos de censuras mais freqüentes num evento de interesse.

Mostrou-se o comportamento da variável resposta através das principais funções matemáticas: função densidade de probabilidade, função de sobrevivência e função de risco. E apresentou-se os modelos de distribuições mais utilizados em análise de sobrevivência dando a ênfase maior nos modelos exponencial e Weibull, descrevendo os procedimentos de estimação de seus parâmetros pelo método de estimação da máxima verossimilhança.

Verificou-se, por meio de exemplos, ou por meio de uma aplicação que o modelo de mistura de longa duração Weibull foi o modelo mais apropriado para análise de conjunto de dados em que muitas observações censuradas estavam presentes, e também, sendo utilizados para ajustar dados em situações diferentes, tais como: sem a presença de covariável e com a presença de covariáveis.

Num estudo de caso sobre a reincidência de ex-detentos da PEM, os modelos de probabilidade Weibull padrão e mistura de longa duração Weibull foram aplicados em função da variável resposta tempo de liberdade. Também foi ajustado o modelo de mistura de longa duração considerando a presença de covariáveis. Finalizando, foram ajustados os modelos Weibull padrão e de mistura de longa duração Weibull para cada grupo de ex-detentos, de acordo com o tipo de benefício recebido para sair da PEM.

Para fazer as comparações sobre os ajustes dos modelos, além das técnicas gráficas, foi utilizado o teste da razão de máxima verossimilhança ou diferença das deviances.

Nas comparações realizadas com os modelos Weibull padrão e mistura de longa duração Weibull, os grupos dos ex-detentos que receberam os benefícios Liberdade Condicional, Liberdade Definitiva e Regime Semi-Aberto tiveram os ajustes do modelo de mistura de longa duração Weibull melhores do que os modelos Weibull padrão, enquanto que para o grupo dos ex-detentos que receberam o benefício Outros, não houve evidência de que o modelo de longa duração produzisse melhor ajuste que o modelo Weibull padrão (sem imunes).

Dos indivíduos considerados imunes, o grupo dos ex-detentos que recebeu o benefício Liberdade Condicional apresentou uma proporção de imunes de 17%; a proporção de imunes do grupo dos ex-detentos que receberam o benefício Liberdade Definitiva foi de 21%; e a proporção dos imunes do grupo que recebeu o benefício Regime Semi Aberto foi de 44%.

5.1 Sugestões para Novas Pesquisas

A modelagem realizada baseou-se no modelo de mistura de longa duração, outra análise envolvendo as covariáveis poderia ter sido feita através da Regressão.

Reaplicação do estudo em outras penitenciárias (para verificar, por exemplo, se o percentual de imunes por benefício é aproximadamente constante. Se os percentuais forem muito diferentes, relacioná-los com as carcterísticas das penitenciárias, etc).

Referências Bibliográficas

ANSCOMBE, F. J. *Estimating a mixed-exponential response law.* Journal of the American Statistical Association**,** v56, p493-502, 1961.

BERKSON, J. and GAGE, R. P. *Survival curve for cancer patients following treatment.* Journal of the American Statistical Association, v47, p501-515, 1952.

BITHEL, J. F. and UPTON, R. G. *A mixed model for survival applied to British children with neuroblastoma, In: Recent Developments in Statistics*. Barra. J. R. et al. (Eds), North-Holland, Amsterdam, p.635-646, 1977.

BLOOM, J. W. *Evaluating human service and correctional programs by modelling the timing of recidivism.* Sociological Methods and Research, v8, 179-208, 1979.

BOAG, J. *Maximun likelihood estimation of the proportion of patients cured by Cancer therapy*. Journal of the Royal Statistical Society, B.11, p15-53, 1949.

BORGES, W. S., COLOSIMO, E. A. e FREITAS, M. A. **Métodos estatísticos para melhoria da qualidade: construindo confiabilidade em produtos**. 12º Simpósio Nacional de Probabilidade e Estatística, 1996. ABE

BROADHURST, R. G. and MALLER, R. A. *Estimating the numbers of terms in criminal careers from one-step probabilities of recidivism*, **J**. Quant. Crim., v7, p275-290, 1991.

COLLETT, D. **Modelling Survival Data in Medical Research**. Chapman and Hall, New York, 1994.

COX, D. R. and HINKLEY, D. V. **Theoretical Statistics**. Chapman and Hall,

New York, 1974.

CORDEIRO, G. M. **Introdução à Teoria de Verossimilhança**. Rio de Janeiro. 10º Simpósio Nacional de probabilidade e estatística, 1992.

DE ANGELIS, R., CAPOCACCIA, R., HAKULINEN, T., SODERMAN, B. and VERDECCHIA, A. *Mixture models for cancer survival analysis: application to population-based data with covariates*. Statistics in Medicine, v18, p441-454, 1999.

DUNSMUIR, W., TWEEDIE, R., FLACK, L. and MENGERSEN, K. *Modelling of transitions between employment states for young Australians*. Austral. J. Statist., v31A, p165-196, 1989.

FAREWELL, V. T. *The use of mixture models for the analysis of survival data with long-term survivors*. Biometrics, v38, p1041-1046, 1982.

FAREWELL, V. T. *Mixture models in survival analysis: are they worth the risk?* Canad. J. Statist., v14, p257-262, 1986.

GAMEL, J. W., MCLEAN, I. W. and ROSENBERG, S. H. *Proportion cured and mean log survival time as functions of tumor size*. Statistics in Medicine, v9, p999-1006, 1990.

GOLDMAN, A. I. *Survivorship analysis when cure is possible: A Monte Carlo study*. Statistics in Medicine, v 3, p153-163, 1984.

GULLAND, J. A. *On the estimation of population parameters from marked members.* Biometrika, v42, p269-270, 1955.

HAYBITTLE, J. L. *The estimation of the proportion of patients cured after treatnent for cancer of the breast*. Brit. J. Radiology, v32, p725-733, 1959.

KAPLAN, E. L. and MEIER, P. *Nomparametric estimation from incomplete observations*. Journal of the American Statistical Association, v53, p457-481, 1958.

KIMBER, H. and CROWDWER, M. *An analysis of resistance times to infection under treatment*. Statistics in Medicine, v3, p165-171, 1984.

KLEIN, J. P. and MOESCHBERGER, M. L. **Survival Analysis: Techiniques for Censored and Truncated Data**. Springer-Verlag, New York, 1997.

LANGLANDS, A. O.,POCOCK, S. J., KERR, G. R. and GORE, S. M. *Long-term survival of patients with breast cancer: a study of the curability of the disease*. Brit. Med. J**.,**v2, p1247-1251, 1979.

LAWLESS, J. L. **Statistical Models and Methods for Lifetime Data**. John Wiley and Sons, New York, 1982.

LEE, T. E. **Statistical Methods for Survival Data Analysis**. Jhon Wiley and Sons, New York, 1992.

LOUZADA NETO, F., MAZUCHELI, J. e ACHAR, J. A. **Introdução à Análise de Sobrevivência e Confiabilidade**. III Jornada de Estatística e II Semana da Estatística. 04 a 08 de novembro de 2002.

MALLER, R.A. and ZHOU, S. *Estimating the proportion of imunes in a censored sample*. Biometrika, v79, p731-739, 1992.

MALLER, R.A. and ZHOU, S. **Survival Analysis with Long-Term Survivors**. John Wiley and Sons, New York, 1996.

MALTZ, M.D. and MACCLEARY, R. *The mathematics of behavioral change: recidivism and construct validity. Evaluation Quarterly, v1, p421-438, 1977.*

MAZUCHELI, J. **Modelos de Múltiplos Riscos e de Mistura em Análise de Sobrevivência**. Tese de doutoramento em Engenharia de Produção, UFRJ. Rio de Janeiro, 2002.

MCLACHLAN, G. J., NG, S. K., ADAMAS, P., MCGILLIN, D. and GALBRAITH, A. J. *An algorithm for fitting mixtures of Gompertz distributions to censored survival data*. Journal of Statistical Software, v2, 1998.

MEEKER, W. Q. and LUVALLE, M. J. *An accelerated life test model based on reliability kinetics.* Technometrics, v37, p133-146, 1995.

MOULD, R. and BOAG, J. W. *A test of several parametric statistical models for estiming success rate in the treatment of carcinoma cervix uteri*. Brit. J.

Cancer, v32, p529-550, 1975.

NELSON, W. **Applied Life Data Analysis**. John Wiley and Sons, New York, 19982.

PARTANEN, J. *On waiting time distributions.* Acta Sociologica, v12, p132-143, 1969.

PENG, Y. and DEAR, K. B. *A nonparametric mixture model for cure rate estimation*. Biometrics, v56, p237-243, 2000.

REGAL, R. R. and LARNTZ, K. *Likelihood methods for testing group problem solving models with censored data*. Psychometrika, v45. p353-366, 1978.

SCHMIDT, P. and WITTE, A. D. **Predicting Recidivism using Survival Models**, Springer-Verlag, NewYork,1988.

SY, J. P. and TAYLOR, J. M. G. *Estimation in a Cox proportional hazards cure model*. Biometrics, v56, p227-236, 2000.

WEIBULL, W. A. *A statistical distribution of wide applicability*. Journal of Applied Mechanics, v18, p293-297, 1951.

YAMAGUCHI, K. *Accelerated failure time regression models with a regression model of surviving fraction: An application to the analysis of 'permanent employment" . Journal of the American Statistical Association, v87, p284-292, 1992.*