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

LUÍSA MARTINS FERNANDES

INFERÊNCIA BAYESIANA EM MODELOS DISCRETOS COM FRAÇÃO DE CURA

Brasília

2013

LUÍSA MARTINS FERNANDES

INFERÊNCIA BAYESIANA EM MODELOS DISCRETOS COM FRAÇÃO DE CURA

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

Orientador: Prof. Eduardo Yoshio Nakano

Universidade de Brasília

Brasília

2013

RESUMO

Este trabalho apresenta inferências do modelo Weibull discreto para dados de sobrevivência com fração de cura. As inferências foram realizadas dentro de um cenário bayesiano fazendo-se o uso das técnicas de MCMC (Markov Chain Monte Carlo). São apresentadas estimativas pontuais dos parâmetros do modelo e seus respectivos intervalos de credibilidade HPD (*Highest Posterior Density*), assim como um teste de significância genuinamente bayesiano – FBST (*Full Bayesian Significance Test*) como uma forma de seleção de modelos. A metodologia apresentada foi aplicada em dados simulados e ilustrada por dois problemas práticos: o primeiro sobre o tempo até a rehospitalização de pacientes com esquizofrenia, e o segundo sobre o tempo até a morte de homens com AIDS. O FBST se mostrou um procedimento simples e útil para seleção de modelos, motivando assim uma abordagem bayesiana na modelagem de dados discretos de sobrevivência.

Palavras-chave: Weibull discreto; fração de cura; inferência bayesiana; FBST.

Abstract

This work presents inferences of the discrete Weibull model for survival data with cure rate. The inferences were conducted within a Bayesian context, using the MCMC (Markov Chain Monte Carlo) techniques. Point estimates of model's parameters and their respective HPD (Highest Posterior Density) credible intervals are presented, as well as a Full Bayesian Significance Test (FBST) as a way to model selection. The methodology presented was applied on simulated data and illustrated by two practical problems: the time until re-hospitalization of patients with schizophrenia and the time until death of men with AIDS. The FBST proved being a simple and useful procedure for model selection, thus motivating a Bayesian approach in the modeling of discrete survival data.

Keywords*:* discrete Weibull; cure rate; Bayesian inference; FBST.

Agradecimentos

Muitas vezes perdemos oportunidades por não acreditarmos que podemos ir mais longe. Para mim, o Mestrado em Estatística foi um exemplo claro disso. Hoje, eu posso dizer que não perdi esta oportunidade por causa de pessoas importantes na minha vida, a elas que agradeço pelo incentivo no desenvolvimento e conclusão deste trabalho:

- Ao meu esposo, Jeovah Sena, pelo amor e paciência, nas muitas horas de estudo e trabalho compartilhadas durante todos esses anos, não só de Mestrado como também de Graduação, incentivando-me a não desistir;
- Ao meu orientador, Eduardo Nakano, que teve paciência, mostrou a "luz no fim do túnel", acreditou na minha capacidade e dividiu seu conhecimento comigo para o desenvolvimento e conclusão deste trabalho;
- Aos meus pais (Aloysio e Fátima) que me deram toda a base de educação para que essa conquista fosse possível e, desde já, me incentivam a seguir aprendendo;
- Aos familiares e amigos que me incentivaram e ajudaram de alguma forma no desenvolvimento e conclusão deste trabalho;
- E a Deus, pois se tenho tudo o que tenho, é porque Ele permitiu.

Índice de figuras

Índice de tabelas

Sumário

1. Introdução

A distribuição Weibull (Weibull, 1951) é uma das mais importantes distribuições utilizadas na modelagem de dados que representam o tempo até a ocorrência de um evento de interesse. Esse evento pode ser a morte de um paciente, reação a um medicamento, falha de um equipamento eletrônico, dentre outros eventos. Em geral, esses dados são analisados através de técnicas de Análise de Sobrevivência, e tem como principal característica a presença de censura, que consiste na observação parcial da resposta. Essa informação censurada, apesar de incompleta, é útil e importante para a análise. A distribuição Weibull é utilizada na análise de dados de sobrevivência quando os mesmos são contínuos. No entanto, em muitos casos os dados de sobrevivência não são contínuos. Dados discretos surgem, por exemplo, quando o tempo de sobrevivência é medido em meses, ciclos ou intervalos. Nakano & Carrasco (2006) estudaram as consequências do uso de um modelo contínuo em um conjunto de dados discretos e mostraram que nem sempre é razoável usar um modelo contínuo quando os dados são discretos.

Além disso, existem situações no qual os indivíduos podem se tornar imunes ao evento de interesse e serem considerados como "curados", ou seja, não suscetíveis ao evento de interesse. Assim, além do tempo de sobrevivência, é importante também estudar a parcela de indivíduos não suscetíveis ao evento de interesse. Modelos para a análise de dados com parcela de curados são frequentemente chamados de Modelos com Fração de Cura (Sobrevivência) ou Modelos de Longa Duração (Confiabilidade). Os modelos com fração de cura tradicionais são baseados em um modelo de mistura de duas distribuições: uma representando a distribuição dos tempos de falha ou sobrevida dos *não curados* (susceptíveis ao evento de interesse) e outra correspondendo a uma distribuição degenerada (que permita tempos de sobrevida em princípio, infinitos) para pacientes *curados* (não susceptíveis ao evento de interesse). Outra classe de modelos de mistura para dados discretos de sobrevivência pode ser visto também em Carrasco et al.

(2012), que consideram uma mistura de uma distribuição de tempos de falha discretos com uma distribuição degenerada no ponto zero (modelos com excessos de zeros) .

Neste contexto, o objetivo deste trabalho é propor uma abordagem bayesiana para o modelo Weibull com tempos de sobrevivência discretos com fração de curados. Desta forma, será feito aqui o uso do modelo Weibull discreto proposto por Nakagawa & Osaki(1975) e o modelo de mistura de Berkson & Gage(1952). Esse modelo Weibull discreto é correspondente ao modelo Weibull contínuo (Weibull, 1951) e tem como caso especial a distribuição Geométrica (que é o correspondente discreto do modelo Exponencial) quando o seu parâmetro de forma é igual a 1. O modelo discreto com fração de cura será proposto dentro de um contexto de análise de sobrevivência considerando dados censurados à direita e seus parâmetros serão estimados seguindo a abordagem bayesiana. Ademais, será proposto um teste de significância genuinamente bayesiano (FBST - *Full Bayesian Significance Test*) para testar o parâmetro de forma da distribuição Weibull discreta, assim como o parâmetro que modela a fração de curados. O FBST é um teste baseado no $e-valor$ (valor de evidência) e dito ser um teste genuinamente bayesiano, pois depende exclusivamente da distribuição *a posteriori* dos parâmetros (Pereira & Stern, 1999). Mais especificamente, o interesse será testar a hipótese do parâmetro de forma da distribuição ser igual a 1 e/ou o parâmetro que modela a fração de cura ser igual a zero. Casos em que essa hipótese não pode ser rejeitada indicam que o modelo mais simples pode ser utilizado. A metodologia proposta será avaliada através de dados simulados e ilustrada através de duas aplicações práticas reais.

No segundo capítulo, é feita uma rápida revisão dos conceitos de Análise de Sobrevivência, relembrando os tipos de censura, funções de sobrevivência, risco e principais relações entre elas; Estimador de Kaplan-Meier e Inferência Bayesiana, com principais conceitos, estimação pontual e intervalar e o teste FBST. A formulação do Modelo Weibull discreto com fração de cura, bem como alguns conceitos importantes, a construção de sua verossimilhança e obtenção da distribuição *a posteriori* são apresentadas no Capítulo 3. No quarto capítulo, são apresentados os dados simulados e respectivos resultados e avaliações da aplicação do modelo Weibull discreto com fração de cura. Finalmente no capítulo 5 a metodologia proposta é aplicada em dados reais e seus resultados são expostos.

2. Revisão Bibliográfica

2.1. Análise de Sobrevivência

A Análise de Sobrevivência é composta de um conjunto de métodos estatísticos para análise de dados para os quais a variável resposta é o tempo até a ocorrência de certo evento. A unidade de estudo, na maioria das vezes, é o indivíduo. O evento de interesse pode ser, dentre outros, a observação de morte, recuperação, reincidência de um fato, manifestação de uma doença, ocorrência de um sinistro, atraso no pagamento de um empréstimo, ou qualquer experiência de interesse que pode acontecer ao indivíduo. Há casos em que pode ser observado, na mesma análise, mais de um evento, podendo ser caracterizado tanto como um caso de eventos recorrentes ou um problema de riscos competitivos. Nenhum dos dois casos será analisado nesse trabalho.

A resposta, tempo até a ocorrência do evento de interesse, pode ser medida em anos, meses, semanas, dias, geralmente não negativa e medida em escala contínua. Alternativamente, por exemplo, pode referir-se à idade de um indivíduo no momento em que o evento de interesse é observado.

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

Colosimo & Giolo (2006) apresentam duas razões para o uso dos dados censurados na análise estatística: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida dos indivíduos; (ii) a omissão das censuras no cálculo das estatísticas de interesse pode originar conclusões errôneas.

Existem cinco tipos de censura (Lee & Wang, 2003):

- Censura tipo I (censura à direita): O período de estudo é pré-fixado e os tempos de sobrevivência dos indivíduos censurados são iguais ou maiores a esse período. Indivíduos perdidos antes do final do tempo de estudo também são observações censuradas.
- Censura tipo II (censura à direita): A quantidade de observações censuradas é pré-fixada e os tempos de sobrevivência dos indivíduos são observados. As observações censuradas são aquelas que se perderam durante o tempo de estudo ou não experimentaram o evento de interesse até o encerramento do estudo.
- Censura tipo III (censura aleatória): o período de estudo é pré-fixado e os indivíduos entram no estudo em diferentes momentos durante esse período. A observação censurada é aquela que é perdida antes que o evento de interesse seja observado. A censura aleatória assume independência entre o tempo de censura e tempo de falha.
- Censura à esquerda: Ocorre quando é sabido que o evento de interesse ocorreu antes de um determinado tempo T, porém o tempo exato inicial da ocorrência é desconhecido. Ou seja, o evento de interesse já ocorreu quando o indivíduo foi observado.
- Censura Intervalar: Ocorre quando o evento de interesse é conhecido por ter ocorrido entre certo intervalo de tempo.

Serão consideradas neste trabalho apenas as censuras à direita (tipos I, II e III), isto é, quando se sabe que o tempo de sobrevivência é sabidamente maior que o tempo censurado.

Em análise de sobrevivência, segundo Colosimo & Giolo (2006), o par (t_i, δ_i) representa os dados relativos ao $indivíduo\ i$ $(i = 1, \dots, n)$, sendo t_i o tempo de falha/censura do indivíduo e δ_i a variável indicadora de falha/censura, onde:

$$\delta_i = \begin{cases} 1\ , se\ t_i \text{ é } um\ tempo\ de\ falha; \\ 0\ , se\ t_i \text{ é } um\ tempo\ censurado. \end{cases} \qquad (\mathbf{2.1})$$

Os tempos de sobrevivência estão sujeitos a variações aleatórias e, como toda variável aleatória, possuem uma distribuição de probabilidade. Sendo T o tempo

transcorrido até a observação do evento de interesse, sua distribuição pode ser geralmente caracterizada pelas funções definidas a seguir.

2.1.1. Função de Sobrevivência

Probabilidade que a variável aleatória T exceda um determinado tempo t, ou seja, probabilidade que um indivíduo sobreviva mais que um tempo t.

Quando T é uma variável aleatória contínua, a Função de Sobrevivência $S(t)$ também é contínua e estritamente decrescente. Considerando $F(t)$ como a função de distribuição acumulada e $f(t)$ como a função densidade de probabilidade, onde $f(t) \geq 0$ para todo $t \geq 0$, a função de sobrevivência $S(t)$ pode ser definida por (Klein & Moeschberger, 2003):

$$\begin{aligned} S(t) &= P\,(T > t) \\ &= 1 - F(t) \\ &= 1 - P(T \leq t) \\ &= \int_t^{\infty} f(u)du \end{aligned} \qquad (\mathbf{2.2})$$

Quando T é uma variável aleatória discreta, que assume os valores $t = 0, 1, 2, \ldots$ e função de probabilidade $p(t) = P(T = t)$, a função de sobrevivência $S(t)$ pode ser definida por:

$$\begin{aligned} S(t) &= P\,(T > t) \\ &= \sum_{k=t+1} P(T = k)\,, \qquad t = 0,1,2,\ldots \end{aligned} \qquad (2.3)$$

2.1.2. Função de Risco

A função de risco, também conhecida como "taxa de falha condicional", é a probabilidade de falha de um indivíduo em um intervalo muito curto de tempo $(t + \Delta t)$, dado que ele sobreviveu até o tempo t, dividido pelo comprimento Δt do intervalo. A

função de risco, denotada por $h(t)$, é não negativa e representa a probabilidade instantânea de falha de um indivíduo. A função de risco é expressa por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P[\, indivíduo\ falhar\ no\ tempo\ (t + \Delta t)\ dado\ que\ ainda\ está\ em\ risco\ no\ tempo\ t]}{\Delta t}$$

Sendo T uma variável aleatória contínua, a função de risco pode ser expressa como:

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

$$h(t) = \frac{f(t)}{S(t)} \tag{2.4}$$

onde $h(t) \geq 0\ \forall t\ e\ \int_0^\infty h(t)\, dt = \infty$.

No caso discreto, a função de risco é igual a zero, exceto nos pontos onde pode ocorrer uma falha. É definida no intervalo $0 \leq h(t) \leq 1$. Pode ser expressa como:

$$h(t) = P(T = t \mid T \geq t) \tag{2.5}$$

$$= \frac{P(T = t)}{P\ (T \geq t)}$$

$$= \frac{P(T = t)}{P\ (T > t) + P(T = t)}$$

$$= \frac{p(t)}{S(t) + p(t)}, \qquad t = 0,1,2, \ldots$$

2.1.3. Relações importantes

As funções de sobrevivência, risco e probabilidade são matematicamente equivalentes, ou seja, conhecendo uma delas as demais podem ser derivadas.

As relações entre as funções expressas anteriormente, para um tempo de vida contínuo *T*, podem ser resumidas como:

$$S(t) = \int_t^{\infty} f(u)du \quad \Leftrightarrow \quad \frac{\partial}{\partial t}\left[1 - S(t)\right] = f(t) \Leftrightarrow$$

$$\Leftrightarrow -S'(t) = f(t)$$

$$= h(t)\, S(t) \quad \Leftrightarrow \quad -\frac{S'(t)}{S\,(t)} = h(t) \Leftrightarrow$$

$$\Leftrightarrow -\frac{\partial}{\partial t}\log S(t) = h(t). \tag{2.6}$$

Integrando de 0 a t , e utilizando $S(0) = 1$, tem-se:

$$\log S(t) = -\int_0^t h(u)du$$

$$\text{e } S(t) = exp\left[-\int_0^t h(u)du\right] \tag{2.7}$$

Para um tempo de vida T, a variável aleatória discreta que assume valores $t = 0, 1, 2, \ldots$, as relações podem ser resumidas da seguinte forma:

$$S(t) = S(t-1) - p(t) \Leftrightarrow \quad S(t) + p(t) = S(t-1) \Leftrightarrow$$

$$\Leftrightarrow \frac{1}{S(t) + p(t)} = \frac{1}{S(t-1)} \quad \Leftrightarrow \quad \frac{p(t)}{S(t) + p(t)} = \frac{p(t)}{S(t-1)} \Leftrightarrow$$

$$\Leftrightarrow h(t) = \frac{p(t)}{S(t-1)}$$

$$= \frac{S(t-1) - S(t)}{S(t-1)} \tag{2.8}$$

$$= 1 - \frac{S(t)}{S(t-1)} \quad , \quad t = 0,1,2, \ldots$$

Para $t = 0$, a função de risco é $h(0) = p(0) = P(T = 0)$.

A função de sobrevivência também pode ser escrita como o produto entre probabilidades de sobrevivência condicionais, e relacionada com a função de risco da seguinte maneira:

$$\begin{aligned} S(t) &= \frac{S(0)}{1}.\frac{S(1)}{S(0)}.\frac{S(2)}{S(1)}\ldots\frac{S(t-1)}{S(t-2)}.\frac{S(t)}{S(t-1)} \\ &= \prod_{k=0}^{t}\frac{S(k)}{S(k-1)} \\ &= \prod_{k=0}^{t}[1-h(k)] \end{aligned} \tag{2.9}$$

Aqui, $S(-1) = P(T > -1) = 1$.

2.2. Estimador de Kaplan-Meier

Conhecido na literatura também como estimador produto-limite, o estimador de Kaplan Meier (Kaplan & Meier, 1958) incorpora a informação de todas as observações disponíveis, tanto censuradas quanto não censuradas, ao contrário das estimativas apresentadas anteriormente.

Considere um estudo com n indivíduos, e seus tempos de sobrevivência distintos ordenados $t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(r)}$. A função de sobrevivência, $S(t)$, é estimada por:

$$\hat{S}_{KM}(t) = \prod_{j:\, t_{(j)} \leq t}\left[\frac{n_j - d_j}{n_j}\right] \tag{2.10}$$

onde, n_j é o número de indivíduos sob risco no tempo $t_{(j)}$ (inclusive) e d_j é o número de indivíduos que experimentam o evento de interesse no tempo $t_{(j)}$, $j = 1,2, \ldots, r$. Quando os dados apresentam algum valor repetido ou censurado, temos $r \leq n$; caso contrário, $r = n$.

A função de risco, $h(t)$, na presença de censuras é estimada por:

$$\hat{h}_{KM}(t) = \frac{nº\ indivíduos\ que\ experimentaram\ o\ evento\ de\ interesse\ em\ t_{(j)}}{\left(t_{(j+1)} - t_{(j)}\right) x (nº\ de\ indivíduos\ com\ tempos \geq t_{(j)})}$$

$$= \frac{d_j}{\Delta t_j\, x\, n_j} \qquad (2.11)$$

$$= \frac{1}{\Delta t_j}\left(1 - \frac{n_j - d_j}{n_j}\right)$$

para $t \in \left[t_{(j)}, t_{(j+1)}\right]$, onde $\Delta t_j = \left(t_{(j+1)} - t_{(j)}\right)$.

2.3. Inferência Bayesiana

Definindo a Estatística como uma ciência que relaciona dados para análise de questões específicas de interesse, esta inclui a elaboração de métodos de coleta, resumo, apresentação e delineamento de respostas às questões levantadas por esses dados. Esses dados podem conter incertezas, seja na seleção dos itens a serem mensurados ou na variabilidade do processo de mensuração.

De acordo com Christensen et al. (2010), a análise estatística bayesiana é baseada na premissa de que toda incerteza deve ser modelada usando probabilidades, e que inferências estatísticas devem ser conclusões lógicas com base nas leis da probabilidade. A inferência bayesiana procura modelar essa incerteza, utilizando modelos de probabilidade.

A análise dos dados é feita pela fórmula de Bayes, utilizado para quantificar o aumento da informação acerca do parâmetro desconhecido (θ). Essa informação pode ser aumentada relacionando esse parâmetro não observável a uma quantidade aleatória (Y) observável. Dado que a distribuição amostral $f(y|\theta)$ define a relação entre a variável aleatória e o parâmetro desconhecido, a fórmula de Bayes é definido da seguinte forma:

$$p\,(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{f(y|\theta)p(\theta)}{\int p(\theta,y)d\theta}$$

Sendo θ uma variável aleatória contínua.

Para θ discreta, a formula de Bayes é definida por:

$$p\ (\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{f(y|\theta)p(\theta)}{\sum_\theta p(\theta,y)}$$

Sendo o teorema um método para atualizar as probabilidades de eventos não observados, faz-se necessária uma probabilidade "*a priori*", definida acima como $p(\theta)$, para o evento despercebido. A distribuição "*a priori*" são informações coletadas previamente ou de forma independente dos dados estudados.

Para um valor fixo da variável aleatória θ, a função de verossimilhança dos prováveis valores de θ pode ser definida como:

$$\ell(\theta; y) = p(y|\theta)$$

Dada a ocorrência do evento relacionado, a função de verossimilhança e a distribuição "*a priori*" são combinadas e levam à distribuição "*a posteriori*". Como $1/p(y)$ não depende do parâmetro (θ), funciona como uma constante normalizadora para $p\ (\theta|y)$.

Dessa forma, a fórmula de Bayes pode ser definido como:

$$p\ (\theta|y)\ \propto \ell(\theta; y)\ p(\theta).$$

Isto é, a distribuição "*a posteriori*" de θ é proporcional a verossimilhança multiplicada pela distribuição "*a priori*".

2.3.1. Estimação pontual e intervalar

Partindo da necessidade de resumir a informação contida *a posteriori* através de poucos valores numéricos, tem-se como forma mais simples a estimação pontual. Nela, a distribuição *a posteriori* é resumida através de um único número, $\hat{\theta}$.

Da perspectiva de *Teoria da Decisão*, para escolher uma estimativa pontual $\hat{\theta}$ de alguma quantidade θ deve-se trabalhar como se $\hat{\theta}$ fosse θ. Para isso, é necessário especificar uma função de perda $l(\hat{\theta}\,|\theta)$, medindo as consequências de trabalhar como se o verdadeiro valor da quantidade de interesse θ fosse $\hat{\theta}$. A perda esperada *a posteriori* se $\hat{\theta}$ for usado é:

$$\ell(\widehat{\theta}\,|y) \;= \int_{\Theta} \ell(\widehat{\theta}\,|\theta)\, p\,(\theta|y)\; d\theta\,,$$

e a estimativa de Bayes é o valor de $\widehat{\theta}$ que minimiza $\ell(\widehat{\theta}\,|\theta)$ em Θ. O estimador de Bayes é a função dos dados $\theta^*(y) = \arg\min_{\theta\epsilon\Theta} \ell(\widehat{\theta}\,|y)$ (Dey & Rao, 2005).

A estimativa de Bayes depende da função de perda escolhida. A função de perda geralmente deve ser escolhida tendo como base as utilizações previstas da estimativa. No entanto, algumas funções de perda convencionais têm sido sugeridas para as situações em que não estão previstas utilizações particulares. Estas funções de perda produzem estimativas que podem ser consideradas como simples descrições da localização da distribuição posterior.

Se a função de perda é quadrática, então a estimativa de Bayes é a média *a posteriori,* $\hat{\theta} = E[\theta|y]$ com:

$$E[\theta_i|y] \;= \int_{\Theta} \theta_i p\,(\theta|y)\; d\theta\,, \qquad i = 1, \dots, k$$

Se a função de perda é a função 0-1 definida como:

$$\ell(\widehat{\theta}\,|\theta) \;= \begin{cases} 1\; se\; \left|\hat{\theta} - \;\theta\right| > \varepsilon \\ 0\; se\; \left|\hat{\theta} - \;\theta\right| < \varepsilon \end{cases}, \;\; \forall \varepsilon > 0$$

então, neste caso pode-se mostrar que a estimativa de Bayes é a moda *a posteriori*, tal que

$$p\,(\hat{\theta}|y) \;= \max_{\theta\epsilon\Theta} p\,(\theta|y)$$

$$= \max_{\theta\epsilon\Theta}\{p(\theta)l(\theta|y)\}$$

Se θ é univariada e a função de perda é linear, a estimativa de Bayes é a mediana *a posteriori* tal que o vetor de medianas *a posteriori:* $\widehat{\theta} \;= \left(\hat{\theta}_1, \dots, \hat{\theta}_k\,\right)$ é definido como:

$$P\{\theta_i \;\geq\; \hat{\theta}_i|\; y\} \;\geq \frac{1}{2}$$

$$P\{\theta_i \;\leq\; \hat{\theta}_i|\; y\} \;\geq \frac{1}{2}, \;\; i = 1, \dots, k$$

Segundo Ehlers (2011), a principal restrição da estimação pontual é que quando estimamos um parâmetro através de um único valor numérico, toda a informação presente na distribuição *a posteriori* é resumida através desse número. É importante também associar alguma informação sobre o quão precisa é a especificação desse número.

O intervalo de confiança bayesiano, ou intervalo de credibilidade, é obtido de uma região de Θ que contenha uma parte substancial da massa probabilística *a posteriori* (Paulino, et al., 2003).

Então, C é um intervalo de credibilidade de $100(1-\alpha)\%$ para θ se:

$$P[\theta \in C] \geq 1-\alpha$$

Intervalos de credibilidade são invariantes. Assim, para qualquer intervalo de credibilidade $100(1-\alpha)\%$ $C = [a,b]$ de θ, então $[\varphi(a), \varphi(b)]$ também é um intervalo de credibilidade $100(1-\alpha)\%$ para $\varphi(\theta)$, onde $\varphi(.)$ é uma transformação um a um.

Dados os infinitos intervalos de credibilidade $100(1-\alpha)\%$ que podem ser construídos, é interessante apenas aquele com o menor tamanho (comprimento, área, volume) possível. Obtidos em regiões onde todos os pontos da região tem maior densidade de probabilidade (HPD) *a posteriori* que os pontos exteriores à região.

Os intervalos HPD são invariantes a transformações lineares $\phi(.)$, mas não a transformações um a um $\varphi(.)$. Assim, para qualquer HPD $100(1-\alpha)\%$ $C = [a,b]$ de θ, então $[\phi(a), \phi(b)]$ também é um intervalo de HPD $100(1-\alpha)\%$ para $\phi(\theta)$, mas $[\varphi(a), \varphi(b)]$ não será um HPD $100(1-\alpha)\%$ para $\varphi(\theta)$. No entanto, mesmo perdendo a característica de ser HPD, $[\varphi(a), \varphi(b)]$ ainda será um intervalo de credibilidade $100(1-\alpha)\%$ para $\varphi(\theta)$.

2.3.2. FBST (Full Bayesian Significance Test)

A abordagem Clássica, tendo como base a distribuição amostral, utiliza como medida de evidência em testes de hipóteses o *nível descritivo* ou *p-valor,* com a finalidade de mensurar a evidência trazida pelos dados em favor da hipótese nula (H_0). A partir da observação desses dados amostrais, a estatística do teste é calculada e, com base no valor observado, é medida a evidência contra a hipótese H_0. O *p-valor* é a probabilidade

de se obter um valor mais extremo que aquele observado pela estatística do teste, ou seja, sob a hipótese H_0, obtém-se a probabilidade para os pontos do espaço amostral que são tão ou mais desfavoráveis para a hipótese H_0 do que o valor observado. Sedo assim, para aqueles dados observados que não favorecem a hipótese nula, o *p-valor* indicará valores pequenos, levando a decisão de rejeitar a hipótese H_0.

É observado que a inferência realizada pela abordagem "Clássica" calcula o nível descritivo levando em consideração a informação dos dados que poderiam ter sido observados, mas ainda não o foram, violando o Princípio da Verossimilhança, segundo o qual todo processo de decisão deve ser feito nos dados devidamente observados. Na abordagem Bayesiana, a medida de evidência é calculada com base na função de verossimilhança dos dados observados e na distribuição *a priori* para a quantidade desconhecida, obedecendo ao Princípio da Verossimilhança e levando em consideração a hipótese alternativa. Também vale ressaltar como vantagem dos testes bayesianos o Fator de Bayes e a Probabilidade *a posteriori* de H_0. No entanto, para análise de uma hipótese precisa, é necessário introduzir uma massa de probabilidade positiva no valor definido sob H_0, que fazem esses testes dependerem não somente da distribuição *a posteriori* dos parâmetros.

Tendo como objetivo a apresentação de uma medida de evidência bayesiana coerente acerca de hipóteses precisas, Pereira & Stern(1999) apresentaram um teste que consiste na análise do "conjuntos de credibilidade". Alternativo ao tradicional *p-valor*, o FBST (*Full Bayesian Significance Test*) necessita, apenas, da distribuição *a posteriori* do(s) parâmetro(s), razão pela qual é chamado de teste "genuinamente bayesiano".

O FBST também apresenta outras propriedades, dentre elas (Faria Júnior, 2006):

- Tem uma definição intrinsecamente geométrica, independente de qualquer aspecto não geométrico, ou seja, é um procedimento invariante;
- Obedece ao Princípio da Verossimilhança, isto é, fornece o mesmo resultado em dois experimentos cujas funções de verossimilhanças são proporcionais;
- Não requer adição de probabilidades positivas a conjuntos de medida nula, ou estabelecer razões de crença iniciais arbitrárias entre hipóteses;

- É um procedimento exato, isto é, não utiliza no cálculo do *e-valor* qualquer aproximação assintótica;
- Permite a incorporação de informação via distribuição *a priori.*

Seja Y a variável aleatória, que quando observada produz os dados y, e considere o espaço estatístico $(\Xi,\ \Delta,\ \Theta,\ F,\ \xi)$ onde: Ξ é o espaço amostral dos possíveis valores de y; Δ é uma família de subconjuntos mensuráveis de Ξ; Θ é o espaço paramétrico; F é uma classe de medidas de probabilidade em Δ, parametrizadas em Θ; ξ é a densidade *a priori* em Θ. Considere uma hipótese nula H_0: $\theta \in \Theta_0$, onde $\Theta_0 \subset \Theta$.

O FBST é construído da seguinte maneira (Pereira & Stern, 1999): inicialmente ,define-se T_φ como um subconjunto do espaço paramétrico onde a densidade *a posteriori* $\xi(\theta|y)$ é maior que φ:

$$T_\varphi = \{\theta \in \Theta \mid \xi(\theta|y) > \varphi\}.$$

A credibilidade de T_φ é a sua probabilidade *a posteriori*:

$$K\left(T_\varphi\right) = \int_{T_\varphi} \xi(\theta|y)d\theta.$$

Agora, definindo θ^* como o (ou um) argumento onde a densidade *a posteriori* atinge o valor máximo sob a hipótese H_0 e ξ^* como o valor dessa densidade:

$$\theta^* \in \arg max_{\theta \in \Theta_0} \xi(\theta|y)$$

$$\xi^* = \xi(\theta^*|y)$$

E definindo também T^* como o "conjunto tangente" à hipótese H_0:

$$T^* = \{\theta \in \Theta \mid \xi(\theta|y) > \xi^*\},$$

então a credibilidade do "conjunto tangente" à hipótese nula é definida por:

$$K^* = \xi(T^*|y) = \int_{T^*} \xi(\theta|y)d\theta$$

Então, a evidência a favor da hipótese H_0, decorrente da observação dos dados y é definida por:

$$Ev(H_0) = 1 - K^*$$

Se a probabilidade de T^* é "grande", isto significa que o conjunto de valores da hipótese H_0 pertence a uma região de baixa probabilidade e a evidência trazida pelos dados é contra a hipótese H_0. Por outro lado, se a probabilidade de T^* é "pequena", então o conjunto de valores da hipótese H_0 está em uma região de alta probabilidade e a evidência trazida pelos dados é em favor da hipótese H_0.

A definição acima é bastante geral, uma vez que foi criada com o objetivo de testar hipóteses precisas, ou seja, uma hipótese nula para a qual a sua dimensão seja menor que a dimensão do espaço paramétrico, isto é, $\dim(\Theta_0) < \dim(\Theta)$.

3. Modelo Weibull discreto com fração de cura

3.1. Distribuição Weibull

Proposta por Wallodi Weibull em 1951 (Weibull, 1951), a distribuição Weibull tem sido uma das distribuições de probabilidade mais utilizadas na modelagem de dados que representam o tempo até a ocorrência do evento de interesse. Isso se deve, em grande parte, à sua simplicidade, flexibilidade, variedade de formas, todas com a importante propriedade de ter a uma função de risco monótona.

3.1.1. Caso Contínuo

Quando a distribuição Weibull é utilizada na análise de dados contínuos, onde T ≥ 0, sua função de densidade dotada de dois parâmetros é definida por:

$$f(t) = \frac{\beta}{\lambda}\left(\frac{t}{\lambda}\right)^{\beta-1} exp\left\{-\left(\frac{t}{\lambda}\right)^{\beta}\right\}, t \geq 0 \qquad (3.1)$$

onde $\beta, \lambda > 0$ são os parâmetros de forma e escala, respectivamente. O parâmetro λ apresenta a mesma unidade de medida de T, enquanto o parâmetro β não tem unidade.

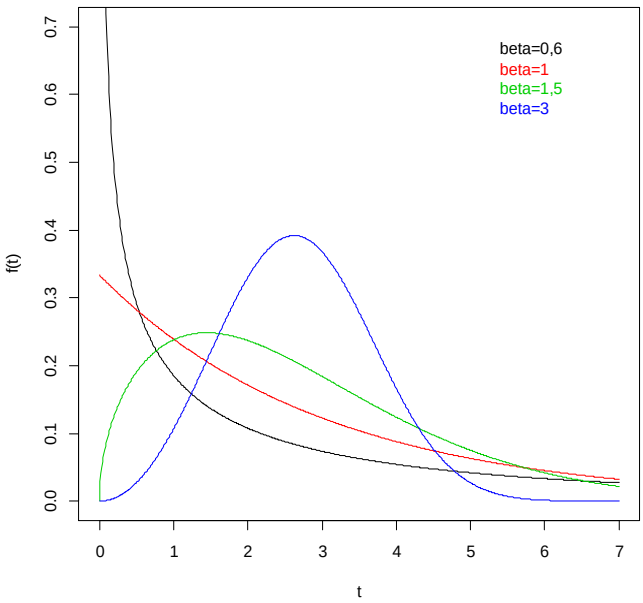

Figura 3.1 - Função densidade da distribuição Weibull contínua para $\lambda = 3$ e diversos valores do parâmetro β.

A função de Sobrevivência da distribuição Weibull contínua é definida por:

$$S(t) = exp\left\{-\left(\frac{t}{\lambda}\right)^{\beta}\right\}, t \geq 0 \qquad (3.2)$$

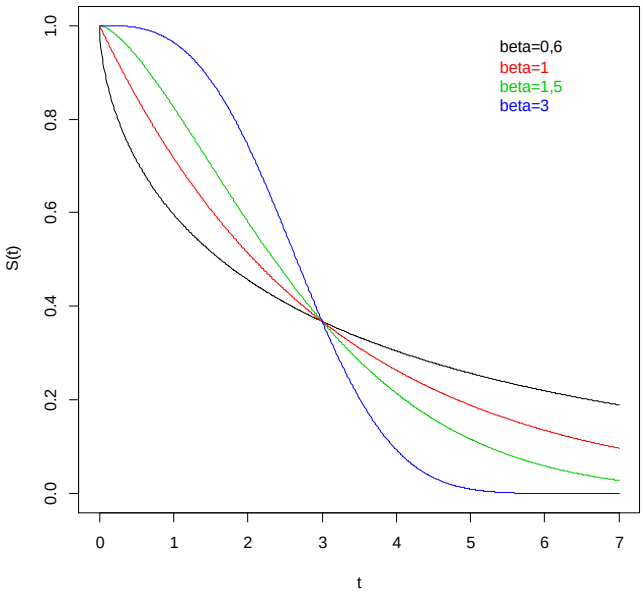

Figura 3.2 - Função de sobrevivência da distribuição Weibull contínua para λ=3 e diversos valores do parâmetro β.

Pela função de sobrevivência, e a relação $h(t) = \frac{f(t)}{S(t)}$, a função de risco é expressa como:

$$h(t) = \frac{f(t \mid \lambda, \beta)}{S(t \mid \lambda, \beta)}$$

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

$$= \frac{\beta}{\lambda}\left(\frac{t}{\lambda}\right)^{\beta-1} \tag{3.3}$$

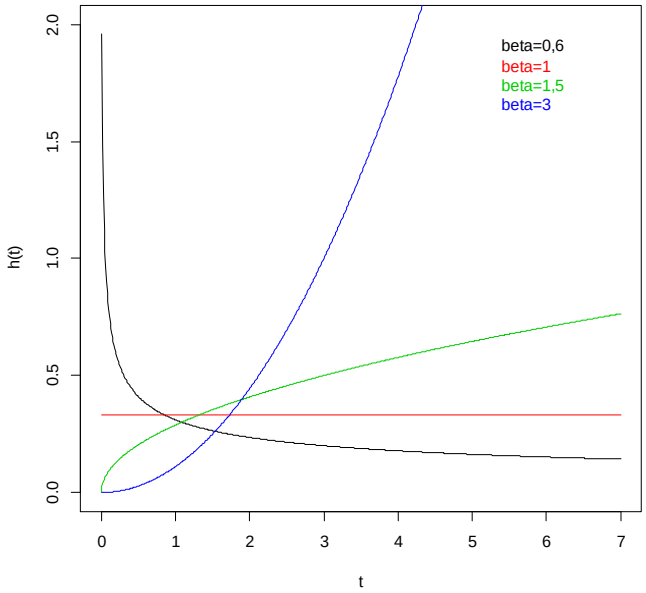

Figura 3.3 - Função de risco da distribuição Weibull contínua para λ=3 e diversos valores do parâmetro β.

Note que para o parâmetro de forma $\beta < 1$, tem-se funções de risco monótonas decrescentes; para $\beta > 1$, as funções de risco são monótonas crescentes; e para $\beta = 1$, tem-se a distribuição Exponencial com função de risco constante. Além disso, a função é côncava se $0 < \beta < 1$ e convexa se $\beta > 2$.

3.1.2. Caso Discreto

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

O modelo Weibull discreto proposto por Nakagawa & Osaki (1975) é equivalente ao modelo Weibull contínuo, e a variável aleatória discreta é obtida por $T = [Y]$, onde $[Y]$ representa a "parte inteira de Y" (Nakano & Carrasco, 2006). Se $Y \sim Weibull(\beta, \lambda)$, então:

$$\begin{aligned} p(t) &= P[T = t] \\ &= P[t \leq Y < t + 1] \\ &= S_Y(t) - S_Y(t + 1) \\ &= e^{-\left(\frac{t}{\lambda}\right)^{\beta}} - e^{-\left(\frac{t+1}{\lambda}\right)^{\beta}} \\ &= q^{t^{\beta}} - q^{(t+1)^{\beta}}, \qquad t = 0,1,2, \ldots \end{aligned} \tag{3.5}$$

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

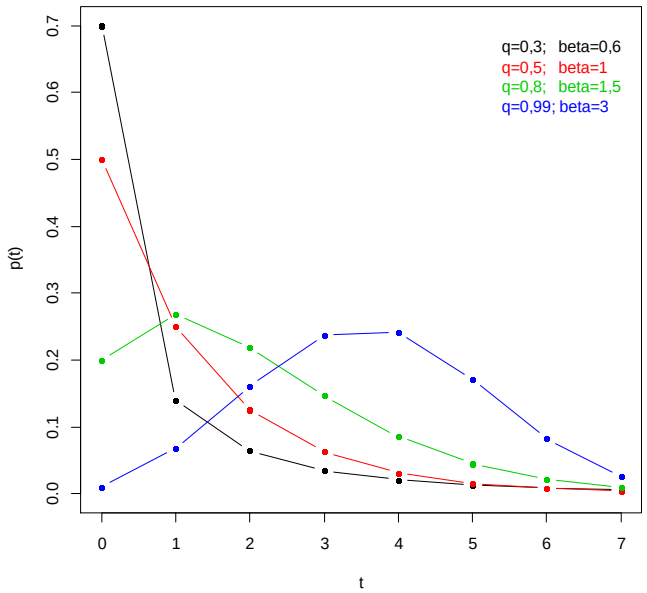

Figura 3.4 - Distribuição de probabilidades da Weibull discreta para diversos valores dos parâmetros λ e β.

A função de sobrevivência é definida como:

$$
\begin{aligned}
S(t) &= P[T > t] \\
&= \sum_{k=t+1}^{\infty} p(t) \\
&= \sum_{k=t+1}^{\infty} q^{t^{\beta}} - q^{(t+1)^{\beta}} \\
&= \left(q^{(t+1)^{\beta}} - q^{(t+2)^{\beta}}\right) + \left(q^{(t+2)^{\beta}} - q^{(t+3)^{\beta}}\right) + \cdots \\
&= q^{(t+1)^{\beta}} \qquad (3.5)
\end{aligned}
$$

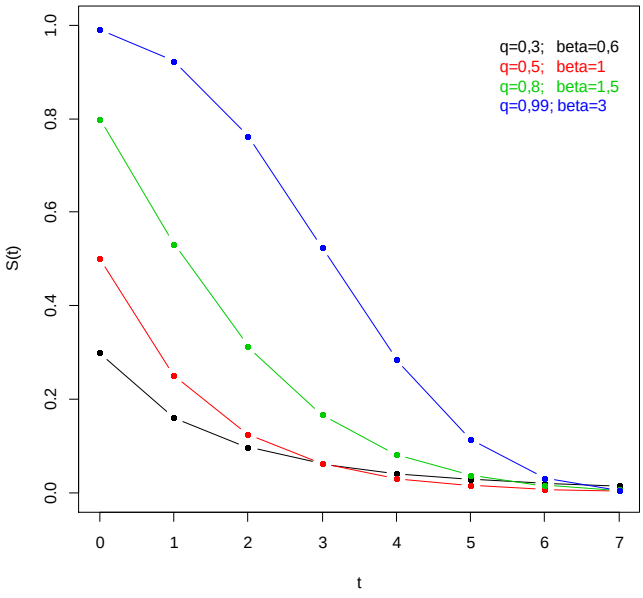

Figura 3.5 - Função de sobrevivência da Weibull discreta para diversos valores dos parâmetros λ e β.

E a função de risco pode ser expressa da seguinte forma:

$$h(t) = \frac{p(t)}{S(t) + p(t)}$$

$$= \frac{q^{t^\beta} - q^{(t+1)^\beta}}{\left[q^{(t+1)^\beta}\right] + \left[q^{t^\beta} - q^{(t+1)^\beta}\right]}$$

$$= 1 - q^{(t+1)^\beta - t^\beta}, \qquad t = 0,1,2, \ldots \tag{3.6}$$

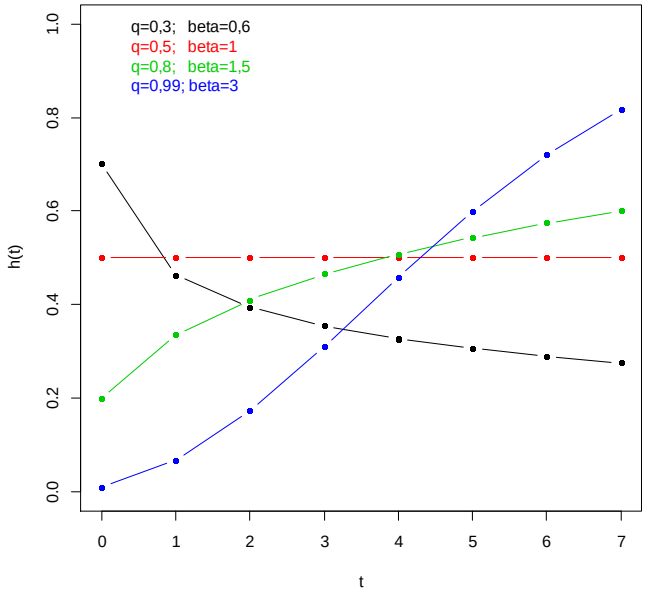

Figura 3.6 - Função de risco da Weibull discreta para diversos valores dos parâmetros λ e β.

A função de risco do modelo Weibull discreto é crescente quando $\beta > 1$ e decrescente quando $\beta < 1$. Note que se $\beta = 1$, o modelo se reduz à distribuição geométrica, que tem função de risco constante e igual a $(1 - q)$.

3.2. Fração de Cura

Em análise de sobrevivência, assume-se que em algum momento do estudo os indivíduos observados sofrerão o evento de interesse. Porém, podem existir indivíduos que, mesmo após um longo período de estudo, nunca apresentarão o evento.

O conjunto de dados aos quais esses indivíduos curados ou imunes pertencem possui uma "fração de curados" ou "indivíduos de longa duração", ou seja, aqueles indivíduos que, apesar de um tempo de estudo considerável, não apresentaram o evento de interesse.

A indicação de indivíduos imunes nos dados de sobrevivência geralmente é dada pela observação de uma proporção elevada de dados com presença de censura à direita. Ou seja, ao final do estudo, após o término do tempo limite, a incidência de indivíduos que não apresentaram o evento de interesse é elevada.

Modelar os dados sem levar em consideração esses quantidade de indivíduos curados ou imunes na população estudada pode levar a conclusões distorcidas. Para analisar tais dados de sobrevivência, Berkson & Gage (1952) propuseram a divisão da população estudada em duas subpopulações: a primeira composta pelos indivíduos não suceptíveis ao evento de interesse (fração de curados), e a outra composta pelos indivíduos ainda sob risco.

A modelagem consiste em uma mistura de duas distribuições paramétricas: uma função de sobrevivência própria associada aos indivíduos não curados (NC), com probabilidade $(1-\pi)$; e para a fração de curados (C), uma função de sobrevivência degenerada com probabilidade associada igual a π, onde $\pi \in (0,1)$. Como o tempo de falha para os indivíduos curados é suposto infinito, em princípio, sua função de sobrevivência, $S_C(t)$ é igual a 1.

Então, a função de sobrevivência com fração de cura, $S_{FC}(t)$, pode ser definida como:

$$\begin{aligned} S_{FC}(t) &= P(C)P(T > t \mid C) + P(NC)P(T > t \mid NC) \\ &= \pi\, S_C(t) + (1-\pi)\, S_{NC}(t) \\ &= \pi + (1-\pi) S_{NC}(t)\,, \end{aligned} \tag{3.7}$$

onde $S_{NC}(t)$ é a função de sobrevivência associada aos indivíduos não curados e $S_C(t)$, é a associada aos curados. A função de sobrevivência com fração de cura é uma função imprópria, pois como:

$$\lim_{t\to\infty} S_{NC}(t) = 0$$

tem-se que

$$\lim_{t\to\infty} S_{FC}(t) = \pi$$

Note que, se $\pi = 0$, então $S_{FC}(t) = S_{\mathrm{NC}}(t)$.

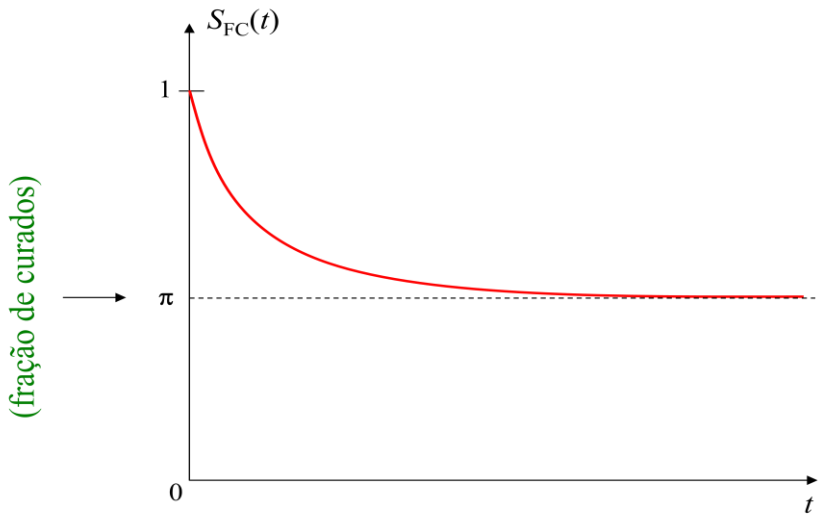

Figura 3.7 - Função de sobrevivência com fração de curados

3.3. Formulação do modelo Weibull com fração de cura

Assumindo T variável aleatória discreta que assume valores $t = 0, 1, 2, \ldots$ como o tempo de ocorrência do evento de interesse, com distribuição $Weibull(\beta, \lambda)$ discreta, tem-se a função de sobrevivência expressa por:

$$S_{FC}(t) = \pi + (1-\pi)\left(q^{(t+1)^{\beta}}\right), \qquad t = 0,1,2,\ldots \tag{3.8}$$

Assim, a distribuição de probabilidades de T é:

$$\begin{aligned} p_{FC}(t) &= S_{FC}(t-1) - S_{FC}(t) \\ &= \left[\pi + (1-\pi)q^{t^{\beta}}\right] - \left[\pi + (1-\pi)\left(q^{(t+1)^{\beta}}\right)\right] \\ &= (1-\pi)\left[q^{t^{\beta}} - q^{(t+1)^{\beta}}\right] \\ &= (1-\pi)\, p_{NC}(t), \end{aligned} \tag{3.9}$$

onde $p_{NC}(t) = q^{t^\beta} - q^{(t+1)^\beta}$ é a distribuição de probabilidade dos indivíduos não curados.

Definidas as funções de risco e sobrevivência para o modelo, a função de risco é definida como:

$$h_{FC}(t) = \frac{p_{FC}(t)}{S_{FC}(t-1)}$$

$$= \frac{(1-\pi)\left[q^{t^\beta} - q^{(t+1)^\beta}\right]}{\left[\pi + (1-\pi)\left(q^{t^\beta}\right)\right]} \tag{3.10}$$

3.4. Formulação da função de Verossimilhança

Em dados de sobrevivência, a densidade e a função de sobrevivência representam, respectivamente, as falhas e as censuras na função de verossimilhança. Sendo assim, a contribuição de um indivíduo que apresenta falha em um tempo t na função de verossimilhança é $p_{FC}(t) = (1-\pi)\, p_{NC}(t)$ e a contribuição do indivíduo cujo tempo foi censurado em t é $S_{FC}(t) = \pi + (1-\pi)S_{NC}(t)$.

Sendo $\Theta = (q, \beta, \pi)$ o vetor de parâmetros desconhecidos e δ_i a variável indicadora de censura do indivíduo i, a verossimilhança para modelos com fração de cura pode ser definido como:

$$L_{FC}(T, \delta | q, \beta, \pi) \propto \prod_{i=1}^{n} [p_{FC}(t_i)]^{\delta_i} [S_{FC}(t_i)]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left[(1-\pi)\left(q^{t_i{}^\beta} - q^{(t_i+1)^\beta}\right)\right]^{\delta_i} \left[\pi + (1-\pi)\left(q^{(t_i+1)^\beta}\right)\right]^{1-\delta_i} \tag{3.11}$$

$$= (1-\pi)^{\Sigma\, \delta_i} \prod_{i=1}^{n} \left[q^{t_i{}^\beta} - q^{(t_i+1)^\beta}\right]^{\delta_i} \prod_{i=1}^{n} \left[\pi + (1-\pi)\left(q^{(t_i+1)^\beta}\right)\right]^{1-\delta_i}$$

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

A fim de simplificar a obtenção da distribuição *a posteriori*, consideramos a introdução da variável latente μ_i que dado (T, δ, Θ) segue distribuição $Bernoulli(p_i)$, onde $p_i = \frac{\pi}{\pi+(1-\pi)q^{(t_i+1)^\beta}}$ (Tanner & Wong, 1987).

Sendo assim, a função de verossimilhança para os dados aumentados é definida por:

$$\begin{aligned} L_{FC}(T,\delta,\mu|q,\beta,\pi) &= L(T,\delta|q,\beta,\pi)L(\mu|q,\beta,\pi) \\ &= \pi^{\sum\mu_i(1-\delta_i)}(1-\pi)^{n-\sum\mu_i+\sum\mu_i\delta_i}\, q^{\sum t_i^\beta\delta_i+\sum[(t_i+1)^\beta(1-\mu_i)(1-\delta_i)]} \\ &\quad \times\ e^{\sum\delta_i log\left(1-q^{(t_i+1)^\beta+t_i^\beta}\right)} \end{aligned} \tag{3.12}$$

3.5. Obtenção da Distribuição *a posteriori*

Consideramos, *a priori*, que $q \sim Beta\,(a_1, b_1)$, $\beta \sim Gama(a_2, b_2)$ e $\pi \sim Beta(a_3, b_3)$, em que $a_1, a_2, a_3, b_1, b_2\ e\ b_3$ são hiper-parâmetros positivos e conhecidos. Então, supondo a independência dos parâmetros, sendo a distribuição *a priori* pode ser descrita da seguinte maneira:

$$h(\Theta) \propto \pi^{a_3-1}(1-\pi)^{b_3-1}q^{a_1-1}(1-q)^{b_1-1}\beta^{a_2-1}e^{b_2\beta} \tag{3.13}$$

Com $\Theta = (q, \beta, \pi)$ temos, de (2.12), (3.12) e (3.13) que a distribuição *a posteriori* dos parâmetros é proporcional a:

$$\begin{aligned} p(q,\beta,\pi|y) &\propto h(\Theta)L_{FC}(T,\delta,\mu|q,\beta,\pi) \\ &\propto \pi^{a_3-1}\pi^{\sum\mu_i(1-\delta_i)}(1-\pi)^{b_3-1}(1-\pi)^{n-\sum\mu_i+\sum\mu_i\delta_i} \\ &\quad \times\ q^{a_1-1}q^{\sum t_i^\beta\delta_i+\sum[(t_i+1)^\beta(1-\mu_i)(1-\delta_i)]}(1-q)^{b_1-1} \\ &\quad \times\ \beta^{a_2-1}e^{b_2\beta}e^{\sum\delta_i log\left(1-q^{(t_i+1)^\beta+t_i^\beta}\right)} \\ &\propto \pi^{a_3+\sum\mu_i(1-\delta_i)-1}(1-\pi)^{b_3+n\,[-\sum\mu_i+\sum\mu_i\delta_i\,]-1} \\ &\quad \times\ \mathrm{q}^{\mathrm{a}_1+\left[\sum \mathrm{t}_\mathrm{i}^\beta\delta_\mathrm{i}+\sum[(\mathrm{t}_\mathrm{i}+1)^\beta(1-\mu_\mathrm{i})(1-\delta_\mathrm{i})]\right]-1}(1-\mathrm{q})^{\mathrm{b}_1-1} \\ &\quad \times\ \beta^{a_2-1}e^{b_2\beta}e^{\sum\delta_i log\left(1-q^{(t_i+1)^\beta+t_i^\beta}\right)}\ . \end{aligned} \tag{3.14}$$

E as distribuições condicionais *a posteriori* são dadas por:

$$p(q|\beta,\pi,T,\delta,\mu) \propto Beta\left(a_1 + \left[\sum t_i^{\beta}\delta_i + \sum\left[(t_i+1)^{\beta}(1-\mu_i)(1-\delta_i)\right]\right], b_1\right) \times \Psi(q,\beta,\pi,T,\delta,\mu)$$

$$p(\beta|q,\pi,T,\delta,\mu) \propto Gama\,(a_2\,, b_2\,) \times \Psi(q,\beta,\pi,T,\delta,\mu)$$

$$p(\pi|q,\beta,T,\delta,\mu) \propto Beta\left(a_3 + \sum \mu_i(1-\delta_i)\;, b_3 + n - \sum \mu_i + \sum \mu_i\delta_i\;\right) \times \Psi(q,\beta,\pi,T,\delta,\mu)$$

em que,

$$\Psi(q,\beta,\pi,T,\delta,\mu) = e^{\sum \delta_i log\left(1-q^{(t_i+1)^{\beta}+t_i^{\beta}}\right)}.$$

Note que, *a posteriori* (3.14) não pode ser obtida analiticamente. No entanto, ela pode ser estimada empiricamente através dos métodos MCMC (Markov Chain Monte Carlo) através do amostrador de Gibbs (Geman & Geman, 1984; Gelfand & Smith, 1990) com passos do algorítmo Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970). Detalhes desses métodos são apresentados nos Anexos A e B.

4. Simulações

Neste capítulo são descritas as simulações computacionais realizadas via software R (versão 3.0.1) juntamente com os resultados numéricos obtidos. As simulações têm por objetivo gerar dados de sobrevivência com fração de curados a fim de testar a significância dos parâmetros do modelo de fração de cura.

Foram geradas amostras de tamanho $n = 80$, $n = 150$ e $n = 500$ da distribuição Weibull Discreta com fração de cura, variando os valores dos parâmetros (q, β) e da fração de cura (π). O percentual de censura (δ) foi modificado conforme a variação de π, assumindo os valores $\delta = 0{,}1$ (quando $\pi = 0{,}05$) e $\delta = 0{,}25$ (quando $\pi = 0{,}15$). O mecanismo de censura utilizado foi o aleatório à direita.

Foram consideradas as estimativas bayesianas para os parâmetros do modelo. Para tanto, foi adotado como *prioris* não informativas, $q \sim \text{Beta}(1,1)$, $\beta \sim \text{Gama}(10^{-5}, 10^{-5})$ e $\pi \sim \text{Beta}(1,1)$. Toda inferência dos parâmetros foi realizada via MCMC – Markov Chain Monte Carlo - através do pacote MCMCPack do R (R Core Team, 2013), que utiliza como núcleo de transição uma cadeia de passeio aleatório (Anexo B).

As hipóteses testadas são as seguintes:

- $H1$ $(\pi = 0)$: com essa hipótese, o objetivo é verificar a parcela da amostra que não está suscetível ao evento de interesse, ou seja, averiguar se um modelo padrão, sem fração de curados, se adequa melhor aos dados simulados;
- $H2$ $(\beta = 1)$: aqui, o interesse é verificar se há alguma perda na precisão das estimativas ao se utilizar um modelo reduzido da Weibull Discreta: Exponecial discreta ou Geométrica;
- $H3(\pi = 0\ e\ \beta = 1)$: unindo as duas hipóteses descritas anteriormente, esta hipótese testa se o modelo Exponencial discreto sem fração de curados pode ser aplicado sem perda na precisão das estimativas.

Ao realizar o teste dessas três hipóteses, existem cinco resultados possíveis, cujas decisões são descritas seguindo o princípio da parcimônia[1]:

1. Aceitar $H1$, Aceitar $H2$, Aceitar $H3$: Este resultado indica que um modelo Exponecial discreto (Geométrico) sem fração de curados é suficiente para ajustar os dados;
2. Rejeitar $H1$, Rejeitar $H2$, Rejeitar $H3$: Este resultado indica que o melhor modelo para ajustar os dados é o modelo Weibull discreto com fração de curados (modelo completo);
3. Rejeitar $H1$, Aceitar $H2$, Rejeitar $H3$: Este resultado indica que um modelo Exponencial discreto (Geométrico) com fração de curados ajusta bem os dados;
4. Aceitar $H1$, Rejeitar $H2$, Rejeitar $H3$: Este resultado indica que um modelo Weibull discreto sem fração de curados ajusta bem os dados;
5. Aceitar $H1$, Aceitar $H2$, Rejeitar $H3$: Este resultado indica que ambos modelos: Weibull discreto sem fração de curados ($H1$) ou Exponencial discreto (Geométrico) com fração de curados ($H2$) podem ser adequados. Como essas hipóteses não são aninhadas, não há como decidir por um modelo pelo princípio da parcimônia. Então, neste caso, um critério que pode ser utilizado é escolher aquele modelo cuja hipótese apresenta o maior $e-valor$.

Note que duas combinações de resultados não citadas acima:

- Aceitar $H1$, Rejeitar $H2$, Aceitar $H3$; e

- Rejeitar $H1$, Aceitar $H2$, Aceitar $H3$

não ocorrem no FBST, pois ambas hipóteses $H1$ e $H2$ contém $H3$ (hipóteses aninhadas), logo o $e-valor$ de $H3$ sempre será maior que os $e-valores$ de $H1$ e $H2$.

Os dados com as médias *a posteriori* (estimativas) e os respectivos intervalos HPD (*Highest Posterior Density*) das estimativas dos parâmetros, bem como o $e-valor$ dos testes $H1: \pi = 0$, $H2: \beta = 1$ e $H3: \pi = 0$ e $\beta = 1$ são apresentados nas tabelas a seguir.

[1] Se um modelo mais simples pode ser rejeitado, não faz sentido optar por um modelo mais complexo.

Tabela.4.1 - Inferência Bayesiana dos parâmetros da simulação com 25% de censura

n	Parâmetro	Valor fixado	Estimativas	IC HPD 95%	*e-valor*		
					$H1$: $\pi = 0$	$H2$: $\beta = 1$	$H3$: $\pi = 0$ $e\ \beta = 1$
80	**q**	0,80	0,82	(0,734 ; 0,895)	-	-	-
	β	1,00	0,81	(0,574 ; 1,038)	-	0,633	0,000
	π	0,15	0,19	(0,037 ; 0,328)	0,262	-	
150	**q**	0,80	0,80	(0,735 ; 0,859)	-	-	-
	β	1,00	0,89	(0,710 ; 1,065)	-	0,759	0,000
	π	0,15	0,18	(0,094 ; 0,262)	0,013	-	
500	**Q**	0,80	0,80	(0,760; 0,834)	-	-	-
	β	1,00	0,95	(0,858 ; 1,044)	-	0,799	0,000
	π	0,15	0,29	(0,247 ; 0,337)	0,000	-	

Nota: Situação 5 em destaque (Modelo Exponencial com fração de cura).

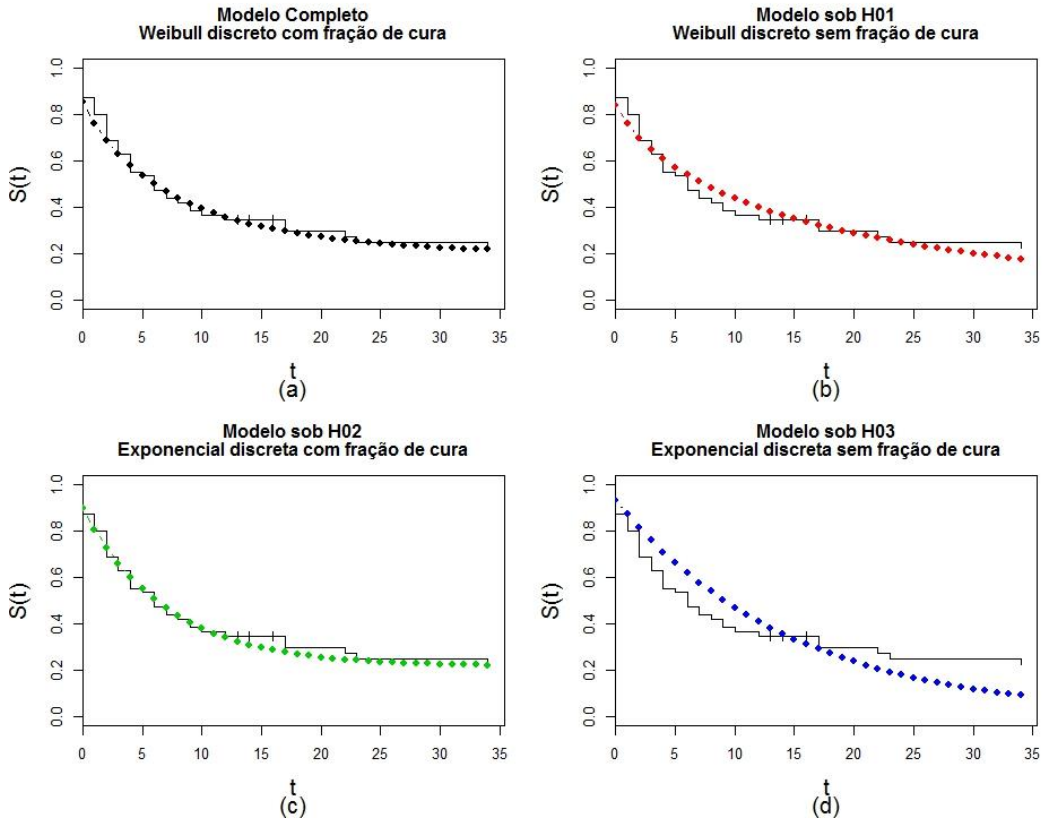

Figura 4.1 - Funções de sobrevivência estimadas para os resultados da amostra simulada de tamanho n=80 apresentada na Tabela 4.1. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese.

Analisando a amostra de tamanho $n = 80$ (em destaque na Tabela 4.1): a variabilidade da fração de cura é grande, o que torna esse valor não significativo $(e - valor = 0{,}262)$ ou seja, não rejeitamos a hipótese $H1$, de que o modelo Weibull discreto sem a fração de curados ajusta bem os dados. O mesmo ocorreu para a

estimativa de β, cujo teste ($e-valor = 0{,}633$) também não rejeitou um bom ajuste do modelo Exponencial discreto com fração de curados (hipótese $H2$). Ainda analisando a amostra de tamanho 80, o $e-valor$ resultante do teste FBST da hipótese $H3$ rejeita fortemente o modelo Exponencial discreto sem fração de cura ($e-valor = 0{,}000$). Este resultado nos leva à Situação 5, descrita anteriormente, que indica que ambos os modelos, Weibull discreto sem fração de cura ou Exponencial discreto com fração de cura, são adequados. Adotando o critério do maior $e-valor$, concluímos que o modelo escolhido é o Exponencial discreto com fração de curados. A Figura 4.1(c) mostra o bom ajuste desse modelo para os dados gerados.

Ademais, podemos notar que o FBST ganha poder à medida que aumentamos o tamanho da amostra, apresentando um $e-valor$ menor. Note que isso parece não ocorrer para a hipótese $H2$, que apresentou um aumento do $e-valor$ à medida que a amostra cresce. No entanto, esse comportamento é facilmente justificado ao notar que coincidentemente as estimativas de β estão mais próximas a 1 (hipótese $H2$) nas amostras grandes. Esse fenômeno ocorreu devido às flutuações existentes na geração da amostra.

Tabela 4.2 - Inferência Bayesiana dos parâmetros da simulação com 25% de censura

n	**Parâmetro**	**Valor fixado**	**Estimativas**	**IC HPD 95%**	***e-valor***		
					$\boldsymbol{H1}$: $\pi = 0$	$\boldsymbol{H2}$: $\beta = 1$	$\boldsymbol{H3}$: $\pi = 0$ $e\ \beta = 1$
80	**q**	0,80	0,83	(0,751 ; 0,904)	-	-	-
	$\boldsymbol{\beta}$	1,30	1,22	(0,895 ; 1,536)	-	0,480	0,133
	$\boldsymbol{\pi}$	0,15	0,18	(0,054 ; 0,304)	0,139	-	
150	**q**	0,80	0,80	(0,737 ; 0,862)	-	-	-
	$\boldsymbol{\beta}$	1,30	1,15	(0,917 ; 1,378)	-	0,539	0,004
	$\boldsymbol{\pi}$	0,15	0,18	(0,100 ; 0,268)	0,014	-	
500	**Q**	0,80	0,81	(0,778 ; 0,848)	-	-	-
	$\boldsymbol{\beta}$	1,30	1,29	(1,158 ; 1,426)	-	0,000	0,000
	$\boldsymbol{\pi}$	0,15	0,25	(0,201 ; 0,289)	0,000	-	

Nota: Situação 3 em destaque (Modelo Exponencial discreto sem fração de cura).

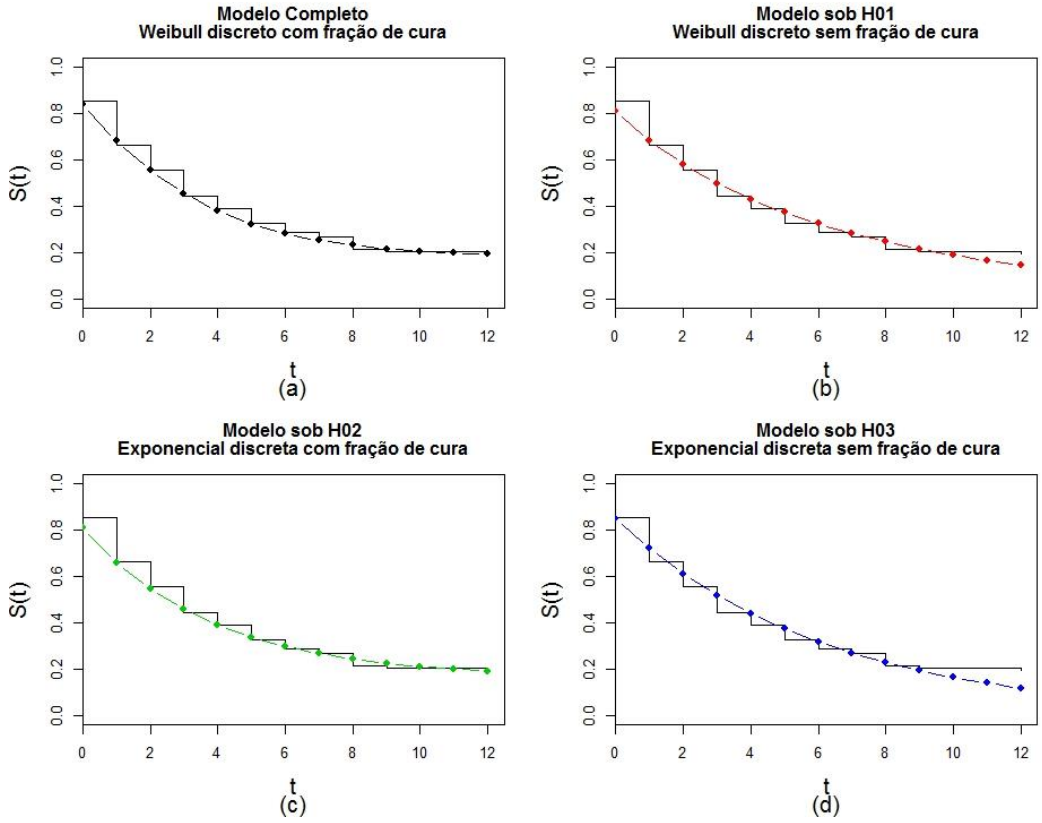

Figura 4.2 - Funções de sobrevivência estimadas para os resultados da amostra simulada de tamanho n=150 apresentada na Tabela 4.2. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese.

Alterado o valor do parâmetro β, com relação à simulação da tabela 4.1, analisamos agora a amostra de tamanho $n = 150$ (em destaque na Tabela 4.2): a estimativa de π apresentou uma baixa evidência de ser igual a zero ($e - valor = 0{,}014$), indicando que a Weibull discreta sem fração de cura não é um bom modelo para o ajuste desse conjunto de dados. A variabilidade da estimativa do parâmetro β é muito grande e portanto esse valor não é significativo ($e - valor = 0{,}539$), neste caso um modelo Exponencial discreto com fração de cura já apresenta um bom ajuste aos dados. Com o resultado do teste da hipótese $H3$, que rejeita fortemente a utilização do modelo Exponencial Discreto sem fração de cura, podemos afirmar que a Situação 3, descrita anteriormente, indica que um modelo Exponencial Discreto com fração de cura ajusta bem os dados. A Figura 4.2(c) ilustra esse ajuste.

Podemos notar também na tabela 4.2 que o FBST ganha poder à medida que o tamanho da amostra é aumentado, apresentando então um $e - valor$ menor. No entanto, não ocorre para a hipótese $H2$. Acreditamos que as flutuações existentes na geração da amostra também justifiquem esse fenômeno.

Tabela 4.3 - Inferência Bayesiana dos parâmetros da simulação com 10% de censura

n	Parâmetro	Valor fixado	Estimativas	IC HPD 95%	*e-valor*		
					$H1$: $\pi = 0$	$H2$: $\beta = 1$	$H3$: $\pi = 0$ $e\ \beta = 1$
80	**q**	0,80	0,83	(0,750 ; 0,899)	-	-	-
	β	1,90	2,00	(1,533 ; 2,466)	-	0,000	0,000
	π	0,05	0,06	(0,000 ; 0,124)	0,502	-	
150	**q**	0,80	0,82	(0,769 ; 0,881)	-	-	-
	β	1,90	1,84	(1,518 ; 2,181)	-	0,000	0,000
	π	0,05	0,05	(0,006 ; 0,103)	0,193	-	
500	**Q**	0,80	0,81	(0,779 ; 0,842)	-	-	-
	β	1,90	1,88	(1,724 ; 2,041)	-	0,000	0,000
	π	0,05	0,07	(0,043 ; 0,092)	0,000	-	

Nota: Situação 4 em destaque (Modelo Weibull discreto sem fração de cura)

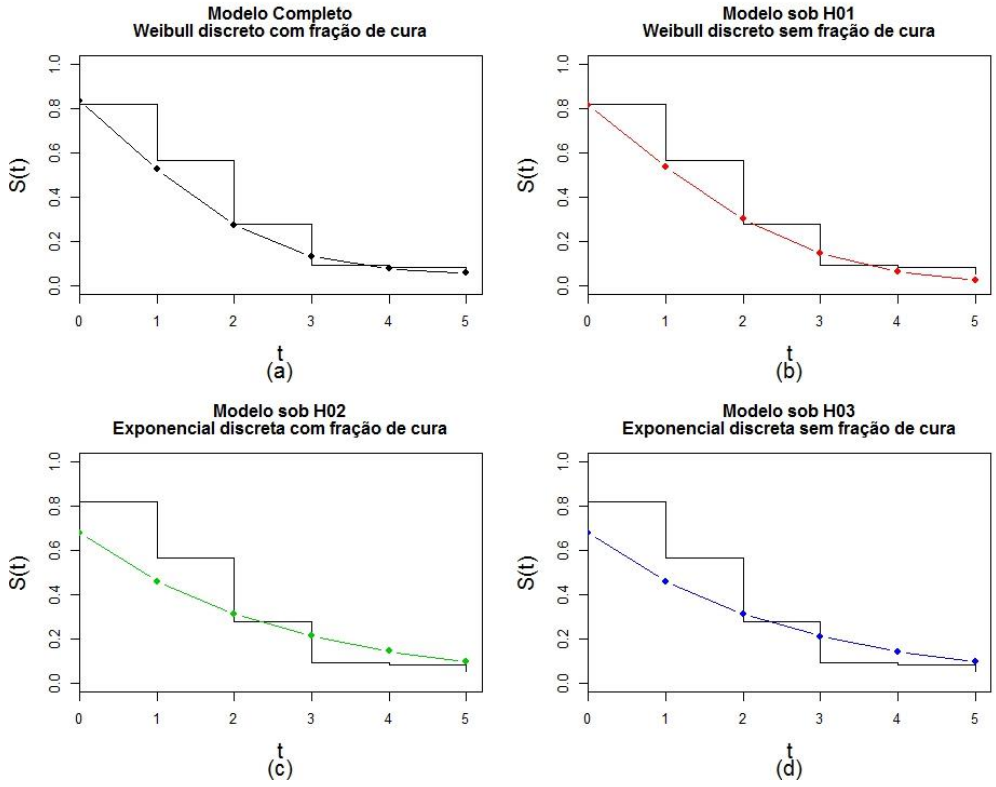

Figura 4.3 - Funções de sobrevivência estimadas para os resultados da amostra simulada de tamanho n=150 apresentada na Tabela 4.3. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese.

A Tabela 4.3 apresenta os resultados quando os valores dos parâmetros são alterados e o percentual de censura em 10% é reduzido. Mesmo com esse menor percentual de censuras, o modelo discriminou as observações censuradas em censuras propriamente ditas e indivíduos curados. Observando a Tabela 4.3 vemos, para n=150, que o modelo Weibull discreto sem fração de cura é um bom modelo para o ajuste do

conjunto de dados. Isso pode ser confirmado com a observação da Figura 4.3(b) e pela análise dos resultados dos testes FBST: o $e-valor$ calculado para testar o parâmetro β rejeita fortemente o modelo Exponencial com e sem fração de cura.

Tabela 4.4 - Inferência Bayesiana dos parâmetros da simulação com 10% de censura

n	Parâmetro	Valor fixado	Estimativas	IC HPD 95%	*e-valor*		
					$H1$: $\pi = 0$	$H2$: $\beta = 1$	$H3$: $\pi = 0$ $e\ \beta = 1$
80	**q**	0,80	0,83	(0,762 ; 0,902)	-	-	-
	$\boldsymbol{\beta}$	1,00	1,07	(0,839 ; 1,306)	-	0,970	0,920
	$\boldsymbol{\pi}$	0,05	0,05	(0,000 ; 0,109)	0,920	-	
150	**q**	0,80	0,87	(0,823 ; 0,913)	-	-	-
	$\boldsymbol{\beta}$	1,00	1,10	(0,927 ; 1,271)	-	0,656	0,006
	$\boldsymbol{\pi}$	0,05	0,07	(0,029 ; 0,124)	0,014	-	
500	**Q**	0,80	0,82	(0,792 ; 0,852)	-	-	-
	$\boldsymbol{\beta}$	1,00	0,98	(0,895 ; 1,072)	-	0,995	0,000
	$\boldsymbol{\pi}$	0,05	0,06	(0,035 ; 0,083)	0,000	-	

Nota: Situação 1 em destaque (Modelo Exponencial discreto sem fração de cura)

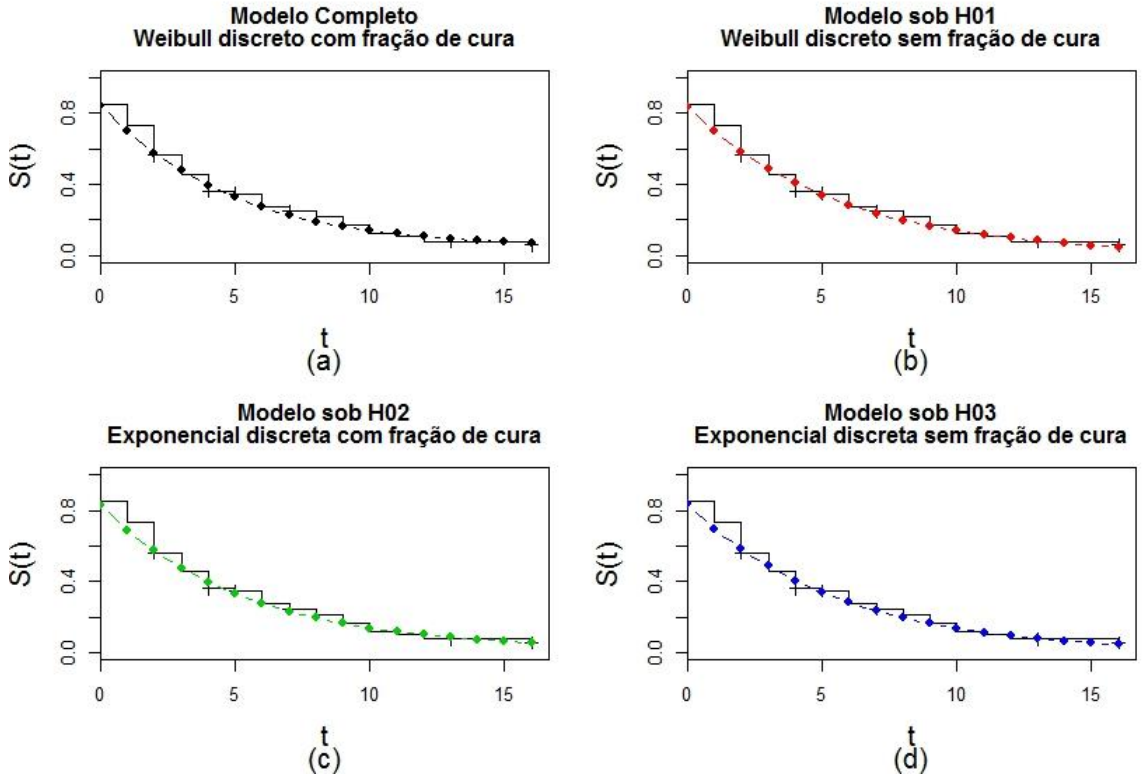

Figura 4.4 - Funções de sobrevivência estimadas para os resultados da amostra simulada de tamanho n=80 apresentada na Tabela 4.4. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese.

Nesta simulação, em destaque na Tabela 4.4, na análise dos $e-valores$ não houve evidência amostral para rejeitar as três hipóteses, remetendo a Situação 1 descrita anteriormente: apesar da estimativa do parâmetro π ser igual a 0,05, o FBST indica que um modelo sem fração de cura apresenta um bom ajuste dos dados; o parâmetro β com $e-valor = 0{,}970$ indica que o modelo Exponencial Discreto se ajusta bem aos dados; e o $e-valor$ da hipótese $H3$ confirma as duas anteriores, indicando que o modelo Exponencial Discreto sem fração de cura é o modelo que melhor se ajusta a este conjunto de dados. A Figura 4.4(d) mostra o bom ajuste desse modelo.

Nesta simulação, mais uma vez, podemos notar que o FBST ganha poder à medida que o tamanho da amostra é aumentado. Também podemos notar que isso parece não ocorrer para a hipótese $H2$ novamente. Esse fenômeno também pode ser justificado devido às flutuações existentes na geração da amostra.

Tabela 4.5 - Inferência Bayesiana dos parâmetros da simulação com 25% de censura

n	Parâmetro	Valor fixado	Estimativas	IC HPD 95%	*e-valor*		
					$\boldsymbol{H1}$: $\pi = 0$	$\boldsymbol{H2}$: $\beta = 1$	$\boldsymbol{H3}$: $\pi = 0$ $e\ \beta = 1$
80	**q**	0,80	0,84	(0,758 ; 0,911)	-	-	-
	$\boldsymbol{\beta}$	1,90	1,76	(1,275 ; 2,237)	-	0,010	0,010
	$\boldsymbol{\pi}$	0,15	0,19	(0,071 ; 0,321)	0,106	-	
150	**q**	0,80	0,81	(0,748 ; 0,878)	-	-	-
	$\boldsymbol{\beta}$	1,90	1,84	(1,462 ; 2,233)	-	0,000	0,000
	$\boldsymbol{\pi}$	0,15	0,23	(0,150 ; 0,312)	0,014	-	
500	**Q**	0,80	0,82	(0,781 ; 0,853)	-	-	-
	$\boldsymbol{\beta}$	1,90	1,90	(1,698 ; 2,099)	-	0,000	0,000
	$\boldsymbol{\pi}$	0,15	0,26	(0,214 ; 0,302)	0,000	-	

Nota: Situação 2 em destaque (Modelo Weibull discreto com fração de cura)

Nesta última simulação, com uma amostra de tamanho $n = 150$ (em destaque na tabela 4.5), as estimativas tanto de β quanto de π rejeitam fortemente as três hipóteses, ou seja, rejeitam o modelo Exponencial discreto com e sem fração de cura e o modelo Weibull sem fração de cura. Este resultado nos leva à Situação 2, descrita no início do capítulo, que indica que a utilização do modelo Weibull discreto com fração de cura (modelo completo) para um bom ajuste deste conjunto de dados é fundamental neste caso. A Figura 4.5(a) a seguir ilustra bem o ajuste do modelo Weibull discreto com fração de cura e um ajuste não tão bom para os demais modelos.

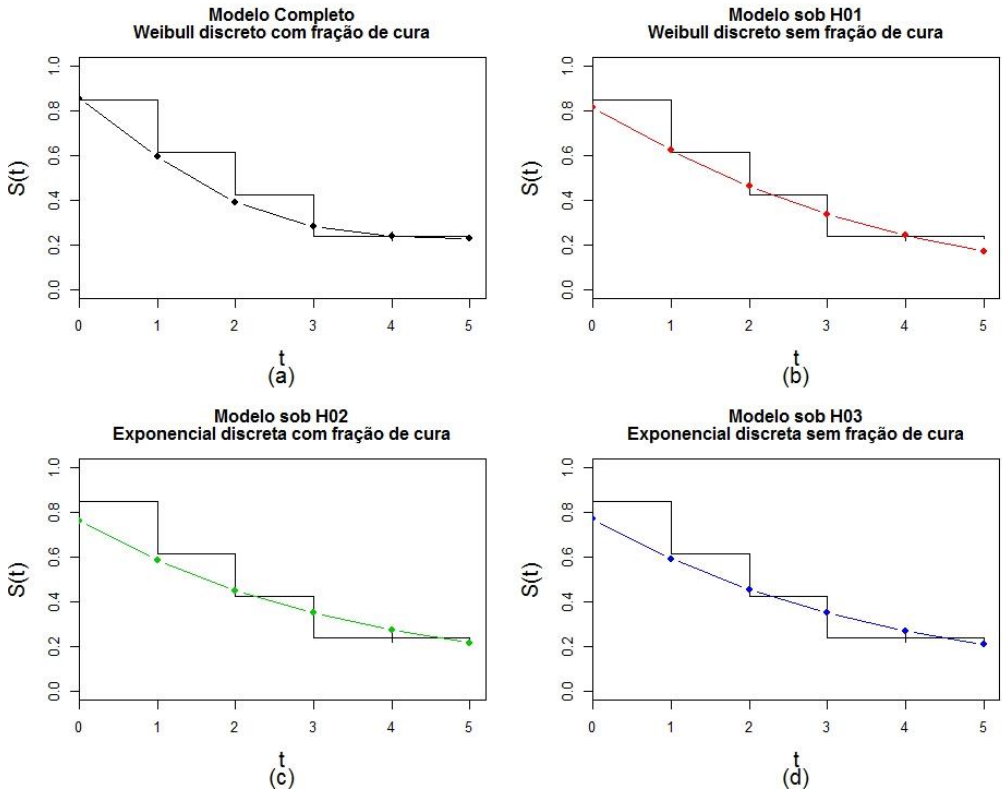

Figura 4.5 - Funções de sobrevivência estimadas para os resultados da amostra simulada de tamanho n=150 apresentada na Tabela 4.5. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese.

5. Aplicação em dados reais

Nesta seção são apresentados duas aplicações do modelo Weibull Discreto com fração de cura a dados reais. Na primeira aplicação, é apresentado um conjunto de dados sobre o tempo até a rehospitalização de pacientes com esquizofrenia que fazem uso de determinado medicamento. Na segunda aplicação, o conjunto de dados refere-se ao tempo até a morte de homens diagnosticados com Síndrome de Imunodeficiência Adquirida (AIDS).

Foram consideradas as estimativas bayesianas para os parâmetros do modelo. Para tanto, foi adotado como *prioris* não informativas, $q \sim Beta(1,1)$, $\beta \sim Gama(10^{-5}, 10^{-5})$ e $\pi \sim Beta(1,1)$. Toda inferência dos parâmetros foi realizada via MCMC – Markov Chain Monte Carlo - através do pacote MCMCPack do R (R Core Team, 2013), que utiliza como núcleo de transição uma cadeia de passeio aleatório (Anexo B).

5.1. Aplicação 1

O ajuste e teste de significância dos parâmetros do modelo Weibull discreto com fração de curados é ilustrado através de um conjunto de dados sobre o tempo até rehospitalização de pacientes com esquizofrenia e que fazem o uso do medicamento antipsicótico Risperidona. A rehospitalização foi definida como a readmissão do paciente por motivos psiquiátricos. Os dados foram baseados em uma amostra de $n = 63$ pacientes do Instituto de Psiquiatria da Universidade de São Paulo e é parte do estudo apresentado por Werneck et al. (2011). Aqui, a variável T representa o número de meses até a rehospitalização. Neste caso, $t = 0$ indica que o paciente retornou ao hospital menos de um mês após sua última visita. Os dados são apresentados na Tabela 5.1 abaixo.

Tabela 5.1 - Tempo até a rehospitalização de pacientes diagnosticados com esquizofrenia e que fazem uso do medicamento risperidona.

Tempo até rehospitalização (meses[1])	**Nº de indivíduos sob risco no início do mês**	**Nº de rehospitalizações no mês**	**Nº de censuras[2] no mês**
0	63	0	4
1	59	3	4
2	52	1	1
3	50	3	8
8	39	1	1
9	37	1	3
12	33	1	0
13	32	1	0
15	31	1	1
16	29	1	0
17	28	1	3
19	24	1	0
20	23	1	3
26	19	1	1
32	17	1	0
34	16	1	15

[1] Os tempos originais são em dias e aqui foram transformados em meses para ilustrar o modelo discreto.

Fonte: Werneck et. al. (2011).

[2] Censuras à direita.

No estudo, a observação cessou quando o paciente foi rehospitalizado, abandonou a medicação prescrita no momento da alta, teve seu medicamento trocado ou tinha chegado ao ponto de 4 anos no estudo sob a mesma medicação, sem ter sido rehospitalizado.

Podemos observar que o total de rehospitalizações (19) no período estudado é cerca de 50% menor que a quantidade total de censuras (44). Ainda observando o quantitativo de pacientes censurados, percebe-se que 23,5% dos pacientes permaneceram 4 anos com a mesma medicação.

Tabela 5.2 - Inferência Bayesiana dos parâmetros em estudo

n	**Parâmetro**	**Estimativas**	**IC HPD 95%**	***e-valor***		
				$\boldsymbol{H1}: \pi = 0$	$\boldsymbol{H2}: \beta = 1$	$\boldsymbol{H3}: \pi = 0$ $e\ \beta = 1$
63	**q**	0,954	(0,899 ; 0,993)	-	-	-
	$\boldsymbol{\beta}$	0,948	(0,572 ; 1,329)	-	0,997	0,931
	$\boldsymbol{\pi}$	0,332	(0,001 ; 0,588)	0,997	-	

Analisando os parâmetros estimados, nota-se que não houve resultado significativo para nenhum dos testes de hipótese. Para o teste $H1$, apesar da estimativa do parâmetro π ser igual a 0,332, não houve evidência amostral para rejeitar a hipótese de nulidade, ou seja, não há uma parcela significativa dos pacientes que permaneceram por mais de 3 anos de estudo sob a mesma medicação, e um modelo sem fração de cura representa bem esses dados neste período de 4 anos. Para o teste $H3$, e $-$ valor $= 0{,}931$ implica a aceitação das hipóteses H1 e H2 (pois H1 e H2 são aninhadas em H3). Assim, pelo princípio da parcimônia, temos que o modelo Exponencial Discreto sem fração de cura é um modelo adequado para ajustar a o tempo até a rehospitalização de pacientes com esquizofrenia que fazem o uso de Risperidona. A figura 5.1 a seguir ilustra de melhor maneira esses resultados.

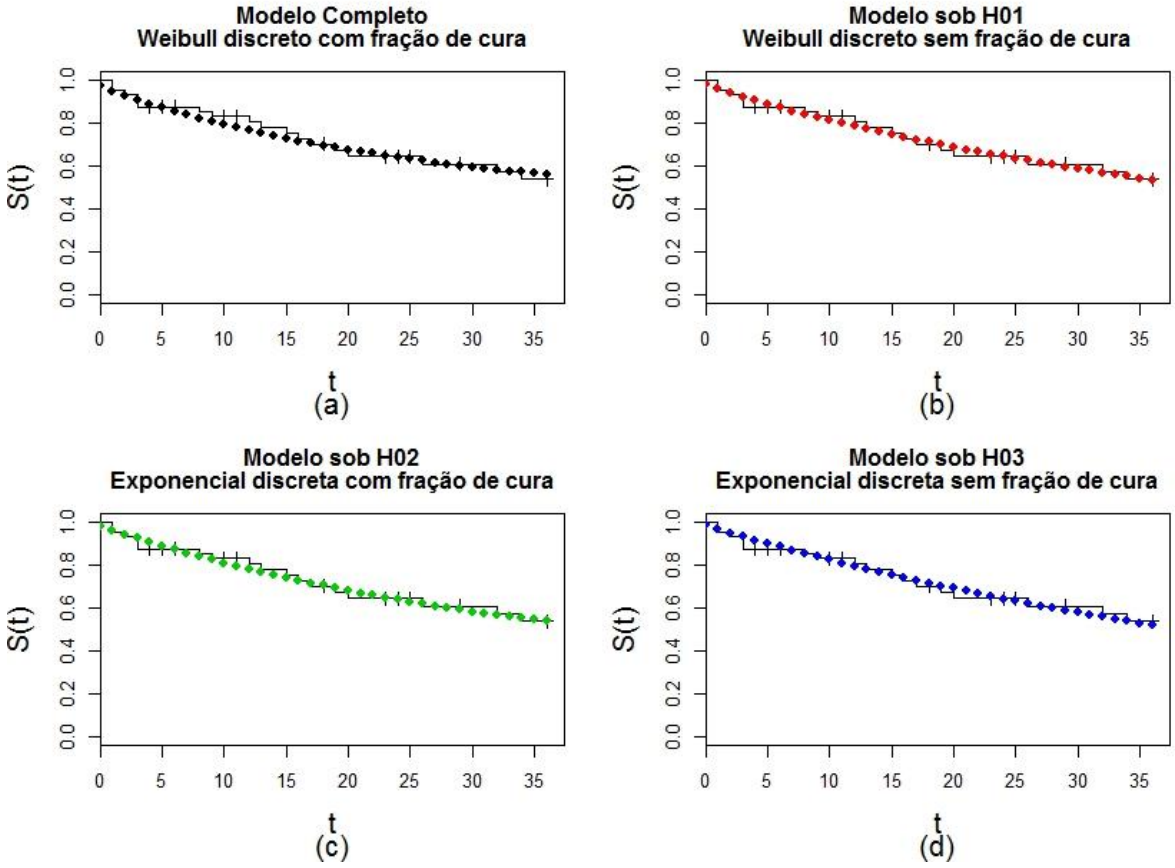

Figura 5.1 - Funções de sobrevivência estimadas para os resultados da amostra em estudo apresentada na Tabela 5.1. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese.

Podemos ver que todos os modelos se ajustam bem ao tempo de rehospitalização. Pelo princípio da parcimônia, concluímos que o modelo Exponencial discreto é adequado para o ajuste. Assim, por apresentar um risco constante, podemos

dizer que o risco de rehospitalização de um paciente com esquizofrenia que usa o medicamento Risperidona é constante ao longo dos meses.

5.2. Aplicação 2

O ajuste e teste de significância dos parâmetros do modelo Weibull discreto com fração de curados é ilustrado novamente em um conjunto de dados sobre o tempo até a morte de homens diagnosticados com AIDS (Síndrome de Imunodeficiência Adquirida). Os dados foram baseados em uma amostra de tamanho $n = 174$ homens brancos que viveram em uma região altamente afetada da cidade de São Francisco, no estado da Califórnia (Selvin, 2008). A variável T representa o número de meses desde o diagnóstico da AIDS até a morte do indivíduo. Neste caso, $t = 0$ indica que o indivíduo morreu antes de completar 1 mês de diagnóstico. Os dados são apresentados na Tabela 5.3 a seguir.

Tabela 5.3 - Tempo até a morte de homens com AIDS.

Tempo até a morte (meses)	Nº de indivíduos sob risco no início do mês	Nº de mortes no mês	Nº de censuras no mês
0	174	6	0
1	168	7	0
2	161	1	0
3	160	4	0
4	156	3	0
5	153	2	0
6	151	7	0
7	144	14	0
8	130	6	0
9	124	7	0
10	117	3	0
11	114	3	1
12	110	2	0
13	108	3	0
14	105	9	0
15	96	6	0
16	90	4	0
17	86	3	0
18	83	7	0
19	76	5	1
20	70	8	0
21	62	3	0
22	59	5	1
23	53	6	1
24	46	4	1
25	41	4	0
26	37	3	1
27	33	3	0
28	30	2	0
29	28	1	0
30	27	2	0
32	25	1	0
33	24	0	1
34	23	2	0
37	21	2	0
41	19	1	1
43	17	2	1
47	14	1	0
51	13	1	1
56	11	1	1
60	9	1	2
65	6	0	1
66	5	0	1
69	4	0	1
72	3	0	1
94	2	0	1
107	1	0	1

Fonte: Selvin(2008) , pág 248.

Observando os dados da tabela, o estudo tem a duração de 108 meses (tempo até a morte do último indivíduo). Aproximadamente 77,5% dos homens observados

morreram até 26 meses de estudo, sugerindo uma possível fração de curados existente no conjunto de dados. Podemos observar também que a quantidade de censuras é baixa, aproximadamente 10,9% dos indivíduos observados. No entanto, percebe-se que as censuras se concentram mais no final do estudo, quando não são observadas mais mortes. Isso é um forte indício de um percentual de indivíduos "curados". Neste caso, "indivíduos curados" pode ser definido como aquele indivíduo que percebe menos os sintomas do vírus HIV.

Tabela 5.4 - Inferência Bayesiana dos parâmetros em estudo

n	**Parâmetro**	**Estimativas**	**IC HPD 95%**	***e-valor***		
				$\boldsymbol{H1}: \pi = 0$	$\boldsymbol{H2}: \beta = 1$	$\boldsymbol{H3}: \pi = 0$ $e\ \beta = 1$
174	**q**	0,980	(0,968 ; 0,991)	-	-	-
	$\boldsymbol{\beta}$	1,341	(1,159 ; 1,527)	-	0,001	0,000
	$\boldsymbol{\pi}$	0,075	(0,033 ; 0,119)	0,000	-	

Como pode ser observado na Tabela 5.4, a estimativa do parâmetro π, apesar de baixa, rejeita fortemente um modelo sem fração de cura para o conjunto de dados estudado. Da mesma forma, a estimativa do parâmetro β rejeita a hipótese de um possível ajuste do modelo Exponencial Discreto. Desta maneira, é possível afirmar que é fundamental a utilização do modelo Weibull discreto com fração de cura para um bom ajuste deste conjunto de dados. Esse fato pode ser facilmente observado na Figura 5.2 a seguir, que mostra um bom ajuste somente para o modelo Weibull discreto com fração de curados. Ademais, a estimativa de $\beta = 1{,}341$ sugere que o risco de morte desses homens aumenta ao longo do tempo.

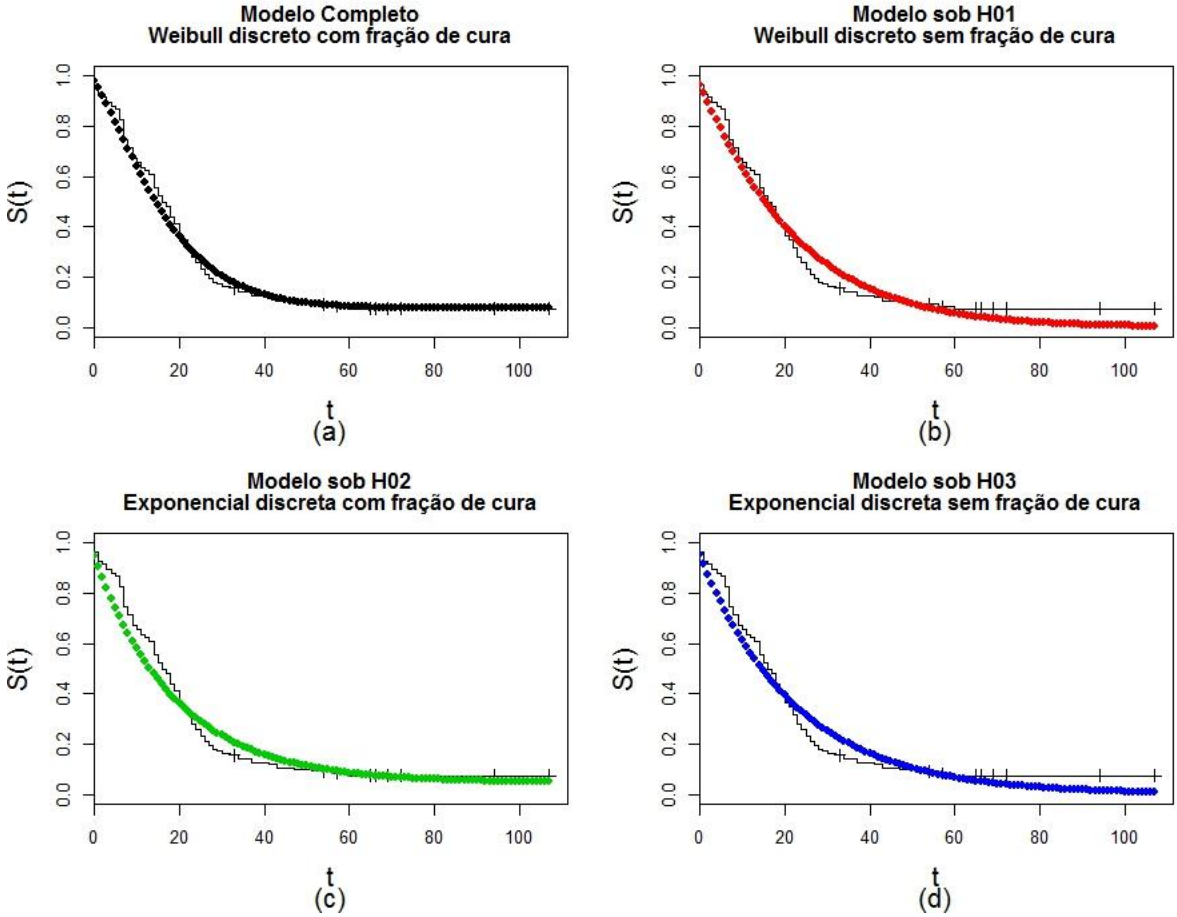

Figura 5.2 - Funções de sobrevivência estimadas para os resultados da amostra em estudo apresentada na Tabela 5.4. A função escada é a estimativa de Kaplan-Meier e os pontos são as estimativas dos modelos definidos por cada hipótese

Comparando os gráficos (a) e (d), podemos observar que o conjunto de dados estudado se ajusta melhor ao modelo Weibull Discreto com fração de cura que ao modelo Exponencial discreto sem fração de cura.

6. Considerações finais

O objetivo deste trabalho foi formular um modelo que considerasse tempos de sobrevivência discretos com fração de curados. O modelo utilizado foi o Weibull discreto proposto por Nakagawa & Osaki (1975), que corresponde ao modelo Weibull contínuo tipicamente utilizado na modelagem de dados que representam o tempo até a ocorrência do evento de interesse. Para analisar as situações em que há uma parcela de indivíduos não susceptíveis ao evento de interesse, nos baseamos no modelo de mistura proposto por Berkson & Gage(1952), onde temos duas distribuições: uma com o tempo de sobrevida dos n*ão curados* e uma distribuição degenerada para a parcela de *curados*.

Os conceitos básicos de tempo de falha e censura também foram introduzidos neste trabalho uma vez que é necessário o entendimento destes conceitos para a análise de sobrevivência.

Tendo como objetivo a apresentação de uma medida de evidência bayesiana coerente acerca de hipóteses precisas, utilizamos o procedimento *Full Bayesian Significance Test* (FBST) proposto por Pereira & Stern (1999) para calcular a evidência em favor das hipóteses nulas H1 ($\pi=0$), para averiguar se um modelo sem fração de curados se adequa aos dados, H2 ($\beta=1$), para verificar se um modelo mais simples (Exponencial discreto sem fração de curados) se adequa aos dados, e H3($\pi=0$ e $\beta=1$), unindo as hipóteses anteriores, para verificar se o modelo Exponencial discreto sem fração de curados pode ser aplicado sem perda de precisão.

No capítulo 4, foram descritas as simulações computacionais realizadas via software R, onde as hipóteses descritas anteriormente foram testadas para amostras de tamanhos diferentes (80, 150, 500). Ao realizar o teste dessas hipóteses, surgiram cinco resultados possíveis, ilustrando situações de seleção de modelos, que foram escolhidos pelo resultado do $e-valor$ do FBST e seguindo o princípio da parcimônia. Na implementação do FBST, foram geradas amostras da distribuição *a posteriori* dos parâmetros via métodos MCMC. Como esperado, o FBST ganha poder a medida que aumentamos o tamanho da amostra, apresentando um $e-valor$ cada vez menor.

No capítulo 5, aplicamos o modelo Weibull Discreto com fração de cura em dados reais. Na primeira aplicação, em que foi medido tempo até a rehospitalização de pacientes com esquizofrenia, não houve resultado significativo para nenhum dos testes, mostrando que o modelo Exponencial Discreto sem fração de cura já é adequado para ajustar os dados, sugerindo um risco de rehospitalização constante ao longo do tempo. Na segunda aplicação, em que foi medido o tempo até a morte de pacientes diagnosticados com AIDS, o teste FBST mostra que é fundamental a utilização do modelo Weibull discreto com fração de cura neste conjunto de dados.

Assim, o FBST mostrou-se uma ferramenta eficaz no estudo do ajuste do modelo Weibull discreto com fração de cura, podendo ser utilizado de forma eficaz como um critério de seleção de modelos. Apesar da distribuição *a posteriori* conjunta dos parâmetros não ser uma distribuição de probabilidades conhecida, valores que seguem a mesma distribuição podem ser facilmente obtidos através dos métodos MCMC. Além disso, o modelo Weibull discreto mostrou-se um modelo bem flexível para modelar tempos de sobrevivência discretos quando os mesmos não apresentam um risco constante. A combinação desses dois resultados pode incentivar a adoção de uma abordagem bayesiana na modelagem de dados discretos de sobrevivência. Poder contar com uma medida de evidência conceitualmente simples e de fácil implementação para a seleção de modelos pode encurtar a aplicação de métodos bayesianos nas pesquisas que envolvam tempos de sobrevivência discretos.

7. Propostas Futuras

Com base nos resultados obtidos neste trabalho, é interessante dar continuidade nos seguintes assuntos:

- Incluir covariáveis no modelo Weibull e no parâmetro de fração de cura.
- Comparar os resultados com a metodologia Clássica.
- Aplicar a metodologia proposta no trabalho a outras distribuições discretas.

8. Bibliografia

Berkson, J., & Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, pp. 47, 501-515.

Carrasco, C. G., Titia, M. H., & Nakano, E. Y. (2012). Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. *Tendências em Matemática Aplicada e Computacional - TEMA*, v.13, n.3, 247-255.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.* USA: CRC Press .

Colosimo, E. A., & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada.* São Paulo: Edgard Blücher Ltda.

Dey, D. K., & Rao, C. R. (2005). *Handbook of Statistics 25: Bayesian Thinking, Modeling and Computation.* Elsevier.

Ehlers, R. S. (2011). *Inferência Bayesiana.* São Paulo: USP.

Faria Júnior, S. (2006). Um ambiente computacional para um teste de significância bayesiano. *Dissertação de Mestrado*. Universidade de São Paulo, Brasil: Instituto de Matemática e Estatística.

Gamerman, D. (1996). *Simulação Estocástica Via Cadeias de Markov.* São Paulo: Associação Brasileira de Estatística.

Gelfand, A., & Smith, A. (1990). Sampling - based approaches to calculating marginal densities. *Journal of the American Statistical Association*, (85):398-409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6): 721-741.

Hastings, W. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, (57):97-109.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American*, (53):457-481.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis : techniques for censored and truncated data.* New York: Springer.

Lee, E. T., & Wang, J. W. (2003). *Statistical methods for survival data analysis.* New Jersey: John Wiley & Sons.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, (21): 1087-1092.

Nakagawa, T., & Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, v. R-24, n. 5, 300-301.

Nakano, E. Y., & Carrasco, C. G. (2006). Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Tendências em Matemática Aplicada e Computacional - TEMA*, v.7, n.1, 91-100.

Paulino, C. D., Turkman, M. A., & Murteira, B. (2003). *Estatística Bayesiana.* Lisboa: Fundação Calouste Gulbenkian.

Pereira, C. A., & Stern, J. M. (1999). Evidence and credibility: full Bayesian significance teste of precise hypothesis. *Entropy*, 1, 99-110.

R Core Team. (2013). *R: A Langage and Environment for Statistical Computing*. Fonte: R Foundation for Statistical Computing, Vienna, Austria: http://www.R-project.org/

Selvin, S. (2008). *Survival analysis for epidemiologic and medical research: A practical guide.* New York: Cambridge University Press.

Tanner, M., & Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82, 528-550.

Weibull, W. (1951). A statistical distribution function of wide applicability. *J. Appl. Mech.-Trans*, ASME 18 (3): 293-297.

Werneck, A., Hallak, J., Nakano, E., & Elkis, H. (2011). Time to rehospitalization in patients with discharged on first generation antipsychotics, non-clozapine second generation antipsychotics, or clozapine. *Psychiatry Research*, pp. v.188, n.3, p.315-319.

APÊNDICE A: Scripts desenvolvidos

```
library(TeachingDemos)
library(splines)
library(survival)
library(MCMCpack)
require(coda)
require(lattice)
require(MASS)

set.seed(131528)

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

######## FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
sob.weidc<-function(x,q,b){ (q)^((x+1)^b) }

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

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

######## FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA COM FRAÇÃO DE CURADOS
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
###   f --- parâmetro que modela a fração de curados
sob.weidc.fc<-function(x,q,b,f){ f + (1-f)*( (q)^((x+1)^b) )}
######## GERAR VALORES DA WEIBULL DISCRETA COM FRAÇÃO DE CURADOS
q<-0.95                    ## parâmetro do modelo
b<-0.85                    ## parâmetro do modelo
n<-80          ## tamanho da amostra
```

```
p.cura<-0.05                ## proporção de curados
p.cens<-0.1                 ## proporção de censura (dos não curados)

tempo<-numeric(n)
censura<-numeric(n)
n.suscept<-n-rbinom(1,n,p.cura)    # gerando a quantidade de NÃO-CURADOS
x<-rweidc(n.suscept,q,b)
censura.suscept<-rbinom(n.suscept,1,1-p.cens)
tempo<-c(x,rep(max(x),n-n.suscept))
censura<-c(censura.suscept,rep(0,n-n.suscept))


########  ESTIMADOR EMPIRICO DE K-M
require(survival)
data<-Surv(tempo,censura)
km<-survfit(data~1)
plot(km,conf.int=F)

#### FUNÇÃO DE VEROSSIMILHANÇA
VERO<-function(p,tempo,censura){
q<-p[1]
b<-p[2]
f<-p[3]
if ( (q>0) && (q<1) && (b>0) && (f>0) && (f<1)  )

return (-1*(
      sum( censura*log(1-f) )
      +sum( censura * log( q^(tempo^b) - q^((tempo+1)^b) ) )
      +sum(   (1-censura)*log(f + (1-f)* q^((tempo+1)^b) ) )
 ))
else return (-Inf)
}

##### OBTENÇÃO DAS ESTIMATIVAS
a<-optim(c(0.9,1,.1),VERO,tempo=tempo,censura=censura)
q.est<-a$par[1]
b.est<-a$par[2]
f.est<-a$par[3]

##### GRÁFICOS
xx<-sort(tempo)
sobWei2<-sob.weidc.fc(xx,q.est,b.est,f.est)
points(xx,sobWei2,type="b",col=4)

posteriori<-function(p,tempo,censura){
q<-p[1]
b<-p[2]
f<-p[3]
if ( (q>0) && (q<1) && (b>0) && (f>0) && (f<1)  )

return (1*(
      sum( censura*log(1-f) )
      +sum( censura * log( q^(tempo^b) - q^((tempo+1)^b) ) )
      +sum(   (1-censura)*log(f + (1-f)* q^((tempo+1)^b) ) )
      + dbeta(q,1,1,log=T)
      + dgamma(b,10^-5,10^-5,log=T)
      + dbeta(f,1,1,log=T)
```

```
 ))
else return (-Inf)
}

M<-100000

theta<-MCMCmetrop1R(posteriori,c(0.9,1,.1),burnin=10000,mcmc=M,tempo=tempo,censura=censura)
q.est2<-mean(theta[,1])
b.est2<-mean(theta[,2])
f.est2<-mean(theta[,3])

##### GRÁFICOS
xx<-sort(tempo)
sobWei3<-sob.weidc.fc(xx,q.est2,b.est2,f.est2)
points(xx,sobWei3,type="b",col=2)

##### ESTIMATIVAS PONTUAIS E HPD
### q
mean(theta[,1])
emp.hpd(theta[,1])

### b
mean(theta[,2])
emp.hpd(theta[,2])

### f
mean(theta[,3])
emp.hpd(theta[,3])


########### FBST H1:f=0
## POSTERIORI SOB Ho1
posteriori.H01<-function(p,tempo,censura){
q<-p[1]
b<-p[2]
f<-0
if ( (q>0) && (q<1) && (b>0)  )

return (-1*(
      sum( censura*log(1-f) )
      +sum( censura * log( q^(tempo^b) - q^((tempo+1)^b) ) )
      +sum(   (1-censura)*log(f + (1-f)* q^((tempo+1)^b) ) )
      + dbeta(q,1,1,log=T)
      + dgamma(b,10^-5,10^-5,log=T)
      + dbeta(f,1,1,log=T)
 ))
else return (-Inf)
}

a.H01<-optim(c(.1,1),posteriori.H01,tempo=tempo,censura=censura)

#### MAXIMO DA POSTERIORI SOB H01
max.post.H01<- -1*a.H01$value

post.mcmc<-numeric(M)
for (j in 1:M) { post.mcmc[j]<-posteriori(theta[j,],tempo,censura) }
```

```
e.valor1<-sum(post.mcmc<max.post.H01)/M
e.valor1


############ FBST H2:b=1
## POSTERIORI SOB Ho2
posteriori.H02<-function(p,tempo,censura){
q<-p[1]
b<-1
f<-p[2]
if ( (q>0) && (q<1) && (f>0) && (f<1) )

return (-1*(
      sum( censura*log(1-f) )
      +sum( censura * log( q^(tempo^b) - q^((tempo+1)^b) ) )
      +sum(   (1-censura)*log(f + (1-f)* q^((tempo+1)^b) ) )
      + dbeta(q,1,1,log=T)
      + dgamma(b,10^-5,10^-5,log=T)
      + dbeta(f,1,1,log=T)
 ))
else return (-Inf)
}

a.H02<-optim(c(.1,.5),posteriori.H02,tempo=tempo,censura=censura)

#### MAXIMO DA POSTERIORI SOB H02
max.post.H02<- -1*a.H02$value

e.valor2<-sum(post.mcmc<max.post.H02)/M
e.valor2


############ FBST H3: f=0 e b=1
## POSTERIORI SOB Ho3
posteriori.H03<-function(p,tempo,censura){
q<-p[1]
b<-1
f<-0
if ( (q>0) && (q<1) )

return (-1*(
      sum( censura*log(1-f) )
      +sum( censura * log( q^(tempo^b) - q^((tempo+1)^b) ) )
      +sum(   (1-censura)*log(f + (1-f)* q^((tempo+1)^b) ) )
      + dbeta(q,1,1,log=T)
      + dgamma(b,10^-5,10^-5,log=T)
      + dbeta(f,1,1,log=T)
 ))
else return (-Inf)
}

#### MAXIMO DA POSTERIORI SOB H03 (OPTIM)
a.H03a<-optim(.5,posteriori.H03,tempo=tempo,censura=censura)
a.H03a
max.post.H03a<- -1*a.H03a$value
e.valor3a<-sum(post.mcmc<max.post.H03a)/M
e.valor3a
```

```
#### MAXIMO DA POSTERIORI SOB H03 (NLM)
#a.H03b<-nlm(posteriori.H03,0.5,tempo=tempo,censura=censura)
#a.H03b
#max.post.H03b<- -1*a.H03b$minimum
#e.valor3b<-sum(post.mcmc<max.post.H03b)/M
#e.valor3b

##########   GRÁFICOS DAS ESTIMATIVAS
data<-Surv(tempo,censura)
km<-survfit(data~1)
plot(km,conf.int=F,xlab="t",ylab="S(t)")

### q
q.full<-mean(theta[,1])
q.H01<-a.H01$par[1]
q.H02<-a.H02$par[1]
q.H03<-a.H03a$par

### b
b.full<-mean(theta[,2])
b.H01<-a.H01$par[2]
b.H02<-1
b.H03<-1

### f
f.full<-mean(theta[,3])
f.H01<-0
f.H02<-a.H02$par[2]
f.H03<-0

xxx<-seq(0,max(tempo))
sob.full<-sob.weidc.fc(xxx,q.full,b.full,f.full)
sob.H01<-sob.weidc.fc(xxx,q.H01,b.H01,f.H01)
sob.H02<-sob.weidc.fc(xxx,q.H02,b.H02,f.H02)
sob.H03<-sob.weidc.fc(xxx,q.H03,b.H03,f.H03)

par(mfrow=c(2,2))

plot(km,conf.int=F,xlab="t",ylab="S(t)",main=c("Modelo Completo","Weibull discreto com fração de cura"))
points(xxx,sob.full,type="b",col=1,pch=16)

plot(km,conf.int=F,xlab="t",ylab="S(t)",main=c("Modelo sob H01","Weibull discreto sem fração de cura"))
points(xxx,sob.H01,type="b",col=2,pch=16)

plot(km,conf.int=F,xlab="t",ylab="S(t)",main=c("Modelo sob H02","Exponencial discreta com fração de cura"))
points(xxx,sob.H02,type="b",col=3,pch=16)

plot(km,conf.int=F,xlab="t",ylab="S(t)",main=c("Modelo sob H03","Exponencial discreta sem fração de cura"))
points(xxx,sob.H03,type="b",col=4,pch=16)
```

ANEXO A : Amostrador de Gibbs

O amostardos de Gibbs (Geman and Geman,1984; Gelfand and Smith,1990) é um esquema de simulação estocástica utilizando cadeias de Markov, cuja função geradora é formada pelas densidades condicionais completas. Torna-se possível gerar amostras de uma distribuição marginal sem a necessidade de se calcular analiticamente a sua densidade.

Seja $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ um vetor paramétrico k-dimensional, onde θ_i, $i = 1,2, \dots, k$, são variáveis aleatórias cuja densidade *a posteriori* é dada por

$$h(\theta|y) = h(\theta_1, \theta_2, \dots, \theta_k) ,$$

e y representa o conjunto de dados observados.

Suponha que o interesse esteja na geração de uma amostra de $h(\theta|y)$ e que a geração direta da *posteriori* conjunta é extremamente complicada/custosa, mas que as gerações das condicionais $h(\theta_i|\theta_{(i)}, y)$ (onde $\theta_{(i)}$ é o vetor de θ sem o i-ésimo componente) são possíveis de ser realizadas. O algoritmo de Gibbs, então, fornece uma alternativa de geração baseada em sucessivas gerações das distribuições condicionais $h(\theta_i|\theta_{(i)}, y)$, $i = 1,2, \dots, k$. O algoritmo é descrito da seguinte forma (Gamerman, 1996):

1. Inicializa-se o contador com iterações da cadeia $j = 1$, e escolhe-se arbitrariamente os valores iniciais $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}\right)$;
2. Obtem-se um novo valor $\theta^{(j)} = \left(\theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_k^{(j)}\right)$ a partir de $\theta^{(j-1)}$ através de sucessivas gerações de valores

$$\theta_1^{(j)} \sim h\left(\theta_1|\theta_2^{(j-1)}, \dots, \theta_k^{(j-1)}, y\right)$$

$$\theta_2^{(j)} \sim h\left(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_k^{(j-1)}, y\right)$$

$$\vdots$$

$$\theta_k^{(j)} \sim h\left(\theta_k|\theta_1^{(j)}, \dots, \theta_{k-1}^{(j)}, y\right) ;$$

3. O contador é atualizado de j para $j + 1$ e retorna-se a (2) até a convergência. Assume-se que a convergência é atingida em uma iteração cuja distribuição

esteja arbitrariamente próxima da distribuição de equilíbrio $h(\theta|y)$ e não no sentido formal inatingível de número de iterações tendendo a infinito.

A forma de obter-se uma amostra de tamanho n é replicar a cadeia m vezes até a convergência (período de burn-in). Após a convergência, todas gerações de uma mesma cadeia são gerações da distribuição de equilíbrio e a amostra pode ser retirada tomando-se saltos entre os valores gerados, de forma a evitar a dependência entre o valor gerado e o valor anterior.

Taxas e diagnósticos de convergência para as cadeias podem ser utilizados tais como Geman and Geman (1984), Gelfand and Smith(1990), entre outros.

ANEXO B : Metropolis-Hastings

O algoritmo de Metropolis-Hastings (Metropolis et al. , 1953 , e Hastings, 1970) é utilizado para gerar amostras de uma distribuição conjunta nos casos onde as densidades condicionais apresentam formas (distribuições) conhecidas.

O interesse está em gerar valores de uma densidade $h(\theta_i|\theta_{(i)}, y)$. Para simplificar a notação, seja $\theta_i = \theta$ e $h(\theta)$ a distribuição marginal desejada. Suponha que a cadeia esteja no estado $\theta^{(j-1)}$ e um valor θ' é gerado de uma distribuição proposta $q(.|\theta^{(j-1)})$ (núcleo de transição que define a função geradora de novo estado de cadeia). O novo valor θ' é aceito com probabilidade

$$\alpha(\theta^{(j-1)}, \theta') = min\left(1, \frac{h(\theta')q(\theta^{(j-1)}|\theta')}{h(\theta^{(j-1)})q(\theta'|\theta^{(j-1)})}\right),$$

onde $h(.)$ é o núcleo da distribuição posterior desejada. O algoritmo de Metropolis-Hastings é dado pelos seguintes passos:

1. Inicia-se arbitrariamente com um ponto qualquer $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}\right)$ e também o contador $j = 1$;
2. Gera-se um novo valor θ' da distribuição $q(\theta'|\theta^{(j-1)})$;
3. Calcula-se a probabilidade de aceitação $\alpha(\theta^{(j-1)}, \theta')$ e simula-se g da Distribuição Uniforme Contínua no intervalo [0,1], ou seja, $g \sim U(0,1)$;
4. Se $g \leq \alpha(\theta^{(j-1)}, \theta')$, aceita-se o novo valor $\theta' = \theta^{(j)}$ e faz-se $j = j + 1$. Caso contrário, a cadeia permance em $\theta^{(j-1)}$ e reinicia-se o processo a partir do passo 2 até a convergência.

O núcleo de transição $q(.)$ define apenas uma proposta de movimento que pode ou não ser confirmado por α. Por esse motivo, $q(.)$ é normalmente chamado de proposta e quando olhado como uma densidade (ou distribuição) condicional, é chamado de densidade (distribuição) proposta.

Escolha de $q(.|\boldsymbol{\theta}^{(j-1)})$

(a) "Cadeias de passeio aleatório": $q(\theta|\theta') = q_1(|\theta' - \theta|)$, onde $q_1(.)$ é uma densidade multivariada. Neste caso, $\theta' = \theta + \epsilon$, onde ϵ é a variável incremento com distribuição $q_1(.)$. No caso em que $q_1(\epsilon) = -q_1(\epsilon)$, tem-se

$$\alpha = min\left(1, \frac{h(\theta')}{h(\theta^{(j-1)})}\right).$$

(b) "Cadeias independentes": se $h(\theta) \propto c(\theta)\psi(\theta)$, onde $c(\theta)$ é uma densidade que pode ser amostrada e ψ uniformemente limitada, toma-se

$$q(\theta|\theta') = c(\theta') .$$

Neste caso particular (o mais eficiente na prática), tem-se que

$$\alpha = min\left(1, \frac{\psi(\theta')}{\psi(\theta^{(j-1)})}\right).$$