ANDRESSA CRISTINA DE MOURA OLIVEIRA

EXTENSÃO DO ALGORITMO DE TURNBULL EM EXPERIMENTOS COM RÉPLICAS:

AVALIAÇÃO VIA SIMULAÇÃO E APLICAÇÃO EM DADOS ENTOMOLÓGICOS

LAVRAS - MG

2011

ANDRESSA CRISTINA DE MOURA OLIVEIRA

EXTENSÃO DO ALGORITMO DE TURNBULL EM EXPERIMENTOS COM RÉPLICAS: AVALIAÇÃO VIA SIMULAÇÃO E APLICAÇÃO EM DADOS ENTOMOLÓGICOS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador

Dr. Mário Javier Ferrua Vivanco

Co-orientador

Dr. Fortunato Silva de Menezes

LAVRAS - MG

2010

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Oliveira, Andressa Cristina de Moura.
Extensão do algoritmo de Turnbull em experimentos com réplicas : avaliação via simulação e aplicação em dados entomológicos / Andressa Cristina de Moura Oliveira. – Lavras : UFLA, 2010.
62p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2010.
Orientador: Mário Javier Ferrua Vivanco.
Bibliografia.

1. Estimação não-paramétrica. 2. Curvas de sobrevivência. 3. Auto-consistência. 4. Censura intervalar. I. Universidade Federal de Lavras. II. Título.

CDD – 519.54

ANDRESSA CRISTINA DE MOURA OLIVEIRA

EXTENSÃO DO ALGORITMO DE TURNBULL EM EXPERIMENTOS COM RÉPLICAS: AVALIAÇÃO VIA SIMULAÇÃO E APLICAÇÃO EM DADOS ENTOMOLÓGICOS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 20 de dezembro de 2010.

Dr. Telde Natel Custódio	UFSJ
Dr. Maurício Sérgio Zacarias	EMBRAPA
Dr. Fortunato Silva de Menezes	UFLA

Dr. Mário Javier Ferrua Vivanco

Orientador

LAVRAS - MG

2010

A André, meu pai; a Regina, minha mãe, pelo valor que dão à educação, e por nunca terem poupado esforços para que eu concluísse a graduação e o mestrado.

Os exemplos que me deram e ainda me dão, com caráter e honestidade, me fizeram como sou hoje; contribuindo significativamente pelas minhas conquistas e me impulsionando a seguir em frente.

Ao meu irmão, Hugo, pelo carinho, amizade e paciência de sempre.

DEDICO

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelas oportunidades que tem me dado, seja de estudo ou de trabalho, por abençoar-me sempre e principalmente pela família que deu a mim.

Aos meus pais, pelo companheirismo, por estarem sempre ao meu lado sendo presente nos momentos mais importantes de minha existência me apoiando sempre. Só tenho a agradecer a felicidade, a admiração o orgulho que têm com as minhas conquistas.

Ao meu irmão Hugo, com quem muito aprendo, por ser meu melhor amigo e pelo carinho de sempre.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, pela oportunidade de realizar o mestrado e o muito que aprendi no tempo que estive nesta instituição.

À CAPES, pela concessão da bolsa de estudos.

Ao professor Dr. Mário Javier Ferrua Vivanco, pela orientação, pelos ensinamentos e dedicação, com quem muito aprendi neste curto espaço de tempo. Sua contribuição para o meu crescimento e desenvolvimento profissional foi fundamental.

Ao professor Dr. Fortunato Silva de Menezes, pela co-orientação, e pelo imenso auxílio principalmente com a parte computacional do trabalho.

Ao professor Dr. Marcelo Ângelo Cirillo, por sua paciência e por acreditar tanto em mim.

Aos demais professores e funcionários do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária e do Departamento de Ciências Exatas da UFLA que muito contribuíram para o meu crescimento neste período.

Ao Paulo César e ao Fabrício que inúmeras vezes me atenderam com boa vontade e presteza para sanar minhas incontáveis dúvidas.

Aos colegas dos cursos de mestrado e doutorado em estatística, principalmente àqueles com os quais passei grande parte deste período e que não esquecerei: Alexandre, Luzia, Suzana, Maíra, Diogo, Lucas, Jair, Felipe, Carol, Thalita, Izabela e Vanessa.

Aos amigos de sempre que mesmo longe são presentes: Elza, Lucas, Diana, Elizângela, Gislane, Gilson, Cássio, Druzo, Tati, Layla e Juliana.

RESUMO

Este trabalho foi desenvolvido com o objetivo de estudar a aplicabilidade do algoritmo proposto por Turnbull (1976), para estimar curvas de sobrevivência, em experimentos com réplicas e comparar os resultados deste algoritmo com o método proposto por Gouvêa (2006). Para tal utilizou-se dados simulados e dados reais provenientes de um experimento realizado com abelhas da espécie *Apis mellifera L.*. Inicialmente aplicou-se o algoritmo em cada uma das réplicas individuais do experimento e posteriormente, à amostra conjunta de todas as réplicas, que neste trabalho chamou-se de "extensão", tanto para os dados simulados quanto para os dados reais. Foi observado que para dados simulados o algoritmo de Turnbull fornece boas estimativas da curva de sobrevivência simulada a partir da distribuição Weibull quando aplicado às réplicas individuais, mas quando o algoritmo é aplicado à amostra conjunta de todas as réplicas o mesmo não ocorre. Ao aplicar o algoritmo de Turnbull aos dados experimentais, verificou-se que existem diferenças em relação aos resultados encontrados por Gouvêa (2006). Verificou-se também que o algoritmo não estima bem as curvas de sobrevivência em amostras pequenas, ou quando o número de intervalos de censura é pequeno.

Palavras-chave: Censura Intervalar. Estimação não-paramétrica. Auto-consistência. Curvas de Sobrevivência.

ABSTRACT

This work was developed to study the applicability of the algorithm proposed by Turnbull (1976) to estimate survival curves in experiments with replicates, and to compare the results of this algorithm with the method proposed by Gouvêa (2006). For this we used simulated data and real data from an experiment with bees (*Apis mellifera L.*). Initially the algorithm was applied individually in each one of the replicates of the experiment and then to the joint sample with all replicates, which in this work was called "extension", for both simulated data and the real data. It was observed that for simulated data the Turnbull's algorithm provides good estimates of the survival curve simulated from the Weibull´s distribution when applied to the individual replicates, but when the algorithm is applied to the joint sample of all the replicates is not true. By applying the Turnbull´s algorithm to the experimental data, it was found that there are differences in the results reported by Gouvêa (2006). It was also noted that the algorithm does not estimate well the survival curves on small samples, or when the number of censoring intervals is small.

Keywords: Interval censoring. Nonparametric estimation. Self-consistency. Survival curves.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO

Uma importante característica dos dados de sobrevivência é a censura, que é a ocorrência de observações incompletas. Dados censurados são aqueles em que os indivíduos estudados não experimentaram o evento de interesse. Sua importância está no fato de mesmo que a observação tenha sido incompleta ou parcial, ainda assim tais observações são utilizadas na análise estatística, uma vez que contribuem na construção da função de verossimilhança, o que melhora a estimação dos parâmetros. São três os tipos de censura: à direita, à esquerda e intervalar.

A censura intervalar, foco deste estudo, é caracterizada pelo fato de não se conhecer o tempo exato de ocorrência do evento, sabe-se somente que este ocorreu em um determinado intervalo de tempo. Esse tipo de censura é muito comum, pois em muitos casos fica inviável monitorar as unidades experimentais constantemente. Em estudos clínicos, por exemplo, as observações aos pacientes estudados são feitas periodicamente, caracterizando o tempo até a ocorrência do evento como dados censurados em intervalos.

A Análise de Sobrevivência é amplamente aplicada nas áreas médica, entomológica, em engenharias, ciências sociais, ciências econômicas, ciências dos alimentos, entre outras. Alguns pesquisadores quando utilizam métodos de Análise de Sobrevivência assumem que o evento que ocorreu em um intervalo de tempo, tenha ocorrido no início, no final, ou no ponto médio deste intervalo, para viabilizar o uso de alguns pacotes estatísticos. Porém, tal decisão conduz a erros de medição, visto que não é conhecido o tempo certo em que aconteceu o evento. Tais erros conduzem à construção de funções de sobrevivência pouco confiáveis.

Existem experimentos em que, por natureza do fenômeno, é necessário fazer réplicas, isto é, dividir a amostra estudada em subgrupos, em que dentro de cada um desses subgrupos haja homogeneidade entre os indivíduos estudados, ou em que dentro de cada um desses subgrupos seja avaliado um diferente aspecto do objeto em estudo, dentre outras razões. Cada um desses subgrupos é chamado de réplica.

Turnbull (1976) desenvolveu um método não-paramétrico para estimar a função de distribuição acumulada em caso de censura intervalar. Este método, embora muito utilizado, ainda não foi aplicado em casos em que os experimentos são planejados e executados com o uso de réplicas. Gouvêa (2006) propôs um estimador não-paramétrico de curvas de sobrevivência para dados com censura intervalar particularmente aplicável em experimentos com réplicas.

Objetivou-se, neste trabalho, estudar a aplicabilidade do algoritmo auto-consistente de Turnbull aos dados de sobrevivência intervalar obtidos de experimentos replicados. Para isso, efetuou-se a aplicação do algoritmo a cada uma das réplicas individuais do experimento e ao agrupamento de todas as réplicas e se avaliou os resultados obtidos. Para assim, talvez, propor uma extensão do método de Turnbull para uma possível aplicação em experimentos com réplicas. Por último, comparou-se a curva de sobrevivência obtida através do algoritmo de Turnbull em um experimento com réplicas com a curva de sobrevivência obtida pelo Estimador Bootstrap proposto por Gouvêa (2006), uma vez que este último obteve excelentes estimativas para a função de sobrevivência com censura intervalar em um experimento replicado com dados entomológicos.

2 REFERENCIAL TEÓRICO

Nesta seção, são apresentados conceitos e métodos que foram utilizados para atingir o objetivo deste trabalho.

2.1 Censura Intervalar

Neste trabalho, nos atemos apenas à censura intervalar, onde o tempo exato de ocorrência do evento de interesse é desconhecido; sabe-se apenas que pertence a um determinado intervalo de tempo. Esse tipo de situação aparece com frequência em estudos clínicos quando os indivíduos são acompanhados por um período pré-estabelecido de tempo e a ocorrência, ou não, do evento é monitorada em visitas periódicas. Assim, os tempos de falha T não são observados exatamente. Sabe-se somente que o evento de interesse ocorreu em algum momento dentro do intervalo de tempo (L,R] em que $L < T \leq R$.

Segundo Colosimo e Giolo (2006), censura à direita, censura à esquerda e tempos exatos de falha podem ser classificados como casos particulares da censura intervalar. Censura à direita pode ser representada pelo intervalo (L,∞), em que $T > L$. Censura à esquerda pode ser representada pelo intervalo (0,R), em que $0 < T < R$. E um tempo exato de falha pode ser representado pelo intervalo [T,T] em que T é o tempo exato de ocorrência do evento.

Novamente, segundo Colosimo e Giolo (2006), existem três mecanismos de censura diferenciados. O Tipo I é aquele em que o estudo será finalizado após um período de tempo estabelecido previamente pelo pesquisador. O Tipo II é aquele em que se finaliza o estudo quando o evento de interesse tiver ocorrido em um número pré-estabelecido de indivíduos. E o terceiro mecanismo, o Aleatório, ocorre quando um indivíduo é retirado no decorrer do estudo sem que

tenha ocorrido a falha. Estes três mecanismos de censura podem ocorrer concomitantemente com a censura intervalar.

Para realizar estudos em que os dados são censurados, é necessário utilizar métodos de análise de sobrevivência, visto que são esses métodos que possibilitam a análise de tais dados. A partir desses métodos estima-se a curva de sobrevivência que fornecerá as informações necessárias sobre os dados estudados.

2.2 Curvas de Sobrevivência

Nos textos básicos de estatística, uma análise descritiva consiste essencialmente em encontrar medidas de tendência central e variabilidade. Como a presença de censuras invalida este tipo de tratamento aos dados de sobrevivência, o principal componente da análise descritiva envolvendo dados de tempo de vida é a função de sobrevivência (COLOSIMO; GIOLO, 2006).

A função de sobrevivência, S(t), é definida como a probabilidade de um indivíduo observado sobreviver a um determinado tempo t. Em termos probabilísticos, podemos escrevê-la como:

$$S(t) = P(T \geq t);$$

em que P(T ≥ t) é a probabilidade de que o tempo T de vida de um indivíduo seja maior que um determinado tempo t.

Para construir curvas de sobrevivência podemos utilizar o estimador não-paramétrico da função de sobrevivência proposto por Kaplan e Meier (1958), denominado pelos autores de estimador limite-produto, que é muito útil e amplamente encontrado na literatura. Este estimador considera tantos

intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.

O estimador de Kaplan-Meier é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right)$$

em que,

- $\hat{S}(t)$ é a função de sobrevivência estimada;
- $t_1 < t_2 < ... < t_k$, são os k tempos distintos e ordenados de falha;
- d_j é o número de falhas em t_j, j = 1, 2, ..., k
- n_j é o número de indivíduos sob risco em t_j

A curva de sobrevivência obtida pelo estimador de Kaplan-Meier é uma função escada, com degraus nos tempos observados de falhas, que fornece informações sobre os dados analisados. Podemos, por exemplo, encontrar o percentual de indivíduos que ainda não experimentaram a falha até um determinado tempo de interesse.

O estimador de Kaplan-Meier é um estimador de máxima verossimilhança de *S(t)*, que por trabalhar com tempos de falha exatos, é utilizado quando ocorre censura à direita. Assim, o definimos porque será utilizado ao iniciar a execução do processo iterativo do algoritmo de Turnbull.

Segundo Peto (1973), com dados censurados em intervalos, a curva de sobrevivência estimada pode não ser única, pois a probabilidade de ocorrência do evento para um determinado indivíduo em estudo depende apenas da diferença entre os valores da função de sobrevivência nos pontos extremos do intervalo e não do comportamento detalhado da função no intervalo. Por isso, ao

estimar uma curva de sobrevivência com dados censurados intervalarmente deseja-se encontrar a estimativa de máxima verossimilhança.

2.3 Estimador Não-Paramétrico de Máxima Verossimilhança

Peto (1973) propôs um método para a análise de dados censurados em um intervalo, onde utilizou-se o algoritmo Newton-Raphson para o Estimador Não-Paramétrico de Máxima Verossimilhança (ENPMV). Tal método é descrito a seguir:

Considere n observações independentes, $x_1, x_2, ..., x_n$; em que sabe-se apenas que x_i está no intervalo $[L_i, R_i]$, ou seja, $L_i \leq x_i \leq R_i$ para i = 1, 2, ..., n. Dos conjuntos $\{L_i\}$ e $\{R_i\}$ podemos obter todos os intervalos fechados disjuntos cujos pontos de extremidades esquerda e direita pertençam aos conjuntos $\{L_i\}$ e $\{R_i\}$, respectivamente, e que não contêm em seu interior membros destes mesmos conjuntos; apenas nos extremos dos intervalos. Estes intervalos são descritos como:

$$[q_1, p_1], [q_2, p_2], ...,[q_m, p_m],$$

em que $q_j \in \{L_i\}$ e $p_j \in \{R_i\}$, e j = 1, 2, ..., m.

Peto (1973) expôs que a verossimilhança é uma função da curva de sobrevivência decrescente nos intervalos de censura e é independente de como esse decrescimento realmente ocorre, ou seja, a estimativa da função de sobrevivência é indefinida em cada intervalo $[q_j, p_j]$ que contêm observações, mas é bem definida entre estes intervalos.

O gráfico 1 ilustra um exemplo em que a função de sobrevivência estimada é indefinida nos intervalos $[q_j, p_j]$ que contêm as observações e é bem definida entre estes intervalos.

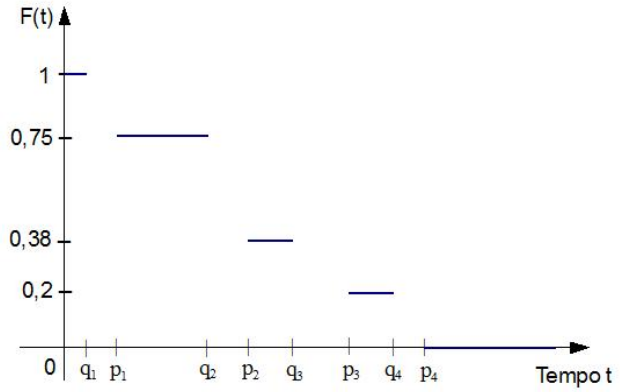

Gráfico 1 Estimativa de uma função de sobrevivência que é bem definida entre os intervalos $[q_j, p_j]$ e é indefinida nestes mesmos intervalos

A partir daí, denotando-se o tamanho desse decrescimento da curva de sobrevivência a cada intervalo (step) por $s_1, s_2, ..., s_m$, em que $\sum_{j=1}^{m} s_j = 1$; o autor afirma que a função de verossimilhança é uma função de $s_1, s_2, ..., s_{m-1}$, em que $s_j \geq 0$. Assim, o logaritmo da função de verossimilhança, conhecido como função suporte, também é uma função de $s_1, s_2, ..., s_{m-1}$.

Para encontrar os valores de s_j, j=1, 2, ...,m, que maximizam a função suporte, Peto (1973) utilizou o método de Newton-Raphson para encontrar o máximo absoluto da função suporte e, por conseguinte, determinar a estimativa não-paramétrica de máxima verossimilhança para dados com censura intervalar.

O raciocínio descrito por Peto (1973) foi utilizado por Turnbull (1976) ao propor um algoritmo auto-consistente para estimar de forma não paramétrica a função de distribuição acumulada de dados com censura intervalar.

2.4 Auto-Consistência

A auto-consistência de um estimador foi definida pela primeira vez por Efron (1967), ao comparar duas amostras com dados censurados.

Um estimador é dito auto-consistente se trata especificamente de censuras e é obtido através de um processo iterativo, específico para cada tipo de dados, realizado para encontrar uma função empírica que seja uma boa estimativa da função desconhecida (que poder ser uma função de distribuição acumulada ou uma função de sobrevivência).

Atribui-se inicialmente um valor para a estimativa da função desejada e através de um processo iterativo obtém-se uma sequência de estimativas que convergem para um valor. Este valor resultante será a estimativa auto-consistente da função desconhecida.

Segundo Efron (1967) o estimador auto-consistente obtido através do processo iterativo é equivalente ao estimador limite-produto de Kaplan e Meier (1958), que por sua vez é um estimador não paramétrico de máxima verossimilhança (ENPMV).

A auto-consistência de um estimador foi muito utilizada por vários autores, sobretudo por Turnbull (1974) ao estimar de forma não paramétrica a função de sobrevivência com dados duplamente censurados, (isto é, censurados à esquerda ou à direita); por Turnbull (1976) ao estimar a função de distribuição acumulada para dados arbitrariamente censurados e truncados, e também por Yu, Li e Wong (2000) que estudaram as propriedades de estimadores auto-consistentes de funções de sobrevivência com censura intervalar.

A seguir, detalhamos o método construído por Turnbull (1976) para estimar de forma não-paramétrica uma função de distribuição quando os dados são truncados e/ou censurados intervalarmente.

2.5 Algoritmo de Turnbull

O algoritmo proposto por Turnbull (1976) estima de forma não-paramétrica uma função de distribuição F, quando os dados são incompletos devido ao agrupamento, a censura e/ou truncamento, usando a ideia de auto-consistência. Um algoritmo simples é construído e mostrado a convergir monotonicamente para obter uma estimativa

de que um ponto isolado {x} é um intervalo fechado [x, x] e que um intervalo semi-infinito é apenas semi-fechado. Assim, podemos escrever:

$$A_i = \bigcup_{j=1}^{k_i} \left[L_{ij}, R_{ij} \right] \quad i = 1, 2, ..., n$$

em que $-\infty \leq L_{i1} \leq R_{i1} < L_{i2} \leq ... < L_{ik} \leq R_{ik} \leq \infty$ e $R_{i1} > -\infty$, $L_{ik} < \infty$.

Turnbull (1976) construiu um conjunto C de intervalos disjuntos cujos pontos de extremidade esquerda e direita estão nos conjuntos $\{L_{ij};\ 1 \leq j \leq k_i,\ 1 \leq i \leq N\}$ e $\{R_{ij};\ 1 \leq j \leq k_i,\ 1 \leq i \leq N\}$, respectivamente, e que não contenham outros membros do $\{L_{ij}\}$ ou $\{R_{ij}\}$, exceto nos seus pontos de extremidade. Assim,

$$C = \bigcup_{j=1}^{m} \left[q_j, p_j \right]$$

em que $q_1 \leq p_1 < q_2 \leq ... < q_m \leq p_m$.

Para estimar a função de distribuição acumulada F, a função de verossimilhança pode ser expressa por:

$$L^*(F) = \prod_{i=1}^{n} P(x_i \in A_i \,|\, x_i \in B_i)$$

$$= \prod_{i=1}^{n} \frac{P\left(x_i \in A_i \bigcap x_i \in B_i\right)}{P\left(x_i \in B_i\right)}$$

$$= \prod_{i=1}^{n} \frac{P\left(x_i \in A_i\right)}{P\left(x_i \in B_i\right)}$$

Assim, a função de verossimilhança é proporcional à

$$L^*(F) = \prod_{i=1}^{n} \left\{ \frac{\sum_{j=1}^{k_i} \left[F(R_{ij}+) - F(L_{ij}-) \right]}{P\left(x_i \in B_i\right)} \right\} \tag{2.1}$$

em que,

$F(R_{ij}+)$ é o valor da função de distribuição no ponto R_{ij} quando tendemos ao número pela direita;

$F(L_{ij}-)$ é o valor da função de distribuição no ponto L_{ij} quando tendemos ao número pela esquerda.

Como desconhecemos a função F que maximiza (2.1), nossa busca por essa função F é facilitada pelos seguintes dois lemas, mencionados por Turnbull (1976), sendo necessário enfatizá-los neste trabalho.

***Lema 2.1*:** Qualquer função de distribuição que cresça fora do conjunto C não pode ser uma estimativa de máxima verossimilhança de F, exceto no caso trivial quando $\forall$ i, temos $A_i \cap C = B_i \cap C$.

Lema 2.2: Para valores fixos de $F(p_j+)$, $F(q_j-)$, j=1, 2, ..., m, a função de verossimilhança é independente do comportamento de F em cada intervalo $[q_j, p_j]$.

Seja $s_j = F(p_j+) - F(q_j -)$, para j=1, 2, ..., m;

em que,

$F(p_j+)$ é o valor da função de distribuição no ponto p_j quando tendemos ao número pela direita;

$F(q_j -)$ é o valor da função de distribuição no ponto q_j quando tendemos ao número pela esquerda.

Os vetores $\boldsymbol{s} = (s_1, s_2, ..., s_m)$ em que $\sum_{j=1}^{m} s_j = 1$ e $s_j \geq 0$, definem classes de equivalência no espaço das funções de distribuição F. Duas funções são ditas equivalentes, se tiverem os mesmos vetores **s**. Todas as funções na mesma classe de equivalência terão a mesma função de verossimilhança pelo Lema 2.2. O Lema 2.1 mostra que pode-se restringir a busca por um EMV para essas classes.

Isto mostra que o problema de maximizar $L^*(F)$ reduz-se a um de maximizar

$$L^*(s_1, s_2, ..., s_m) = \prod_{i=1}^{N} \left[\frac{\sum_{j=1}^{m} \alpha_{ij} s_j}{\sum_{j=1}^{m} \beta_{ij} s_j} \right] \tag{2.2}$$

$$\alpha_{ij} = \begin{cases} 1, se \left[q_j, p_j \right] \subset A_i \\ 0, caso \quad contrário \end{cases} \quad \text{e} \quad \beta_{ij} = \begin{cases} 1, se \left[q_j, p_j \right] \subset B_i \\ 0, caso \quad contrário \end{cases}$$

em que $\sum_{j=1}^{m} s_j = 1$ e $s_j \geq 0$, j=1, 2, ..., m.

Porém, não há uma solução analítica fechada para maximizar $L^*(s_1, s_2, ..., s_m)$ e assim estimar $F(p_j+)$ e $F(q_j-)$. Então, precisa-se utilizar algum procedimento numérico, e neste caso o procedimento é o algoritmo de Turnbull.

Para solucionar o problema acima, Turnbull construiu o algoritmo auto-consistente que será apresentado a seguir.

O Algoritmo Auto-Consistente

O algoritmo auto-consistente proposto por Turnbull (1976) obtém o EMV de **s** com base na equivalência entre o método de máxima verossimilhança e a auto-consistência.

A auto-consistência, definida na seção (2.4), é uma extensão de uma ideia de Efron (1967) para dados censurados à direita e que depois foi utilizada por Turnbull (1974) para dados duplamente censurados.

Para $1 \leq i \leq n$, $1 \leq j \leq m$, considere

$$I_{ij} = \begin{cases} 1, & se \quad x_i \in \left[q_j, p_j \right] \\ 0, & caso \quad contrário \end{cases}$$

Por causa da censura o valor de I_{ij} é desconhecido, no entanto a sua esperança foi definida por Turnbull como,

$$E\left[I_{ij} \right] = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k} = \mu_{ij}\left(s\right), \tag{2.3}$$

em que,

α_{ij} é a constante indicadora se o intervalo $[q_j, p_j]$ pertence ao intervalo A_i;

s_j é a probabilidade do intervalo $[q_j, p_j]$ conter alguma observação.

Assim, $\mu_{ij}(s)$ representa a probabilidade de que a i-ésima observação pertença ao intervalo $[q_j, p_j]$ quando F pertence à classe de equivalência definida por $\mathbf{s} = (s_1, ..., s_m)$.

Além disso, por causa da truncagem, cada observação $X_i = x_i$, pode ser considerada pertencente a um conjunto G, cujo tamanho é desconhecido e todos os valores (exceto X_i) estão no complementar de B_i. J_{ij} será o número no conjunto G correspondente à i-ésima observação, que têm valores em $[q_j, p_j]$.

Exemplificando: considere um conjunto G que contenha a observação $X_i = x_i$ dentre outros elementos. De todos os elementos do conjunto G, apenas x_i pertence ao conjunto B_i, todos os demais elementos de G estão no complementar de B_i. Daí, J_{ij} será o número de elementos de G que pertencem ao intervalo $[q_j, p_j]$. J_{ij} é desconhecido, mas a sua esperança, em **s**, é dada por:

$$E\left[J_{ij}\right] = \frac{\left(1-\beta_{ij}\right)s_j}{\sum_{k=1}^{m}\beta_{ik}s_k} = v_{ij}\left(s\right) \tag{2.4}$$

Assim, $v_{ij}(s)$ representa a probabilidade de que outras observações, além de x_i, estejam no intervalo $[q_j, p_j]$ quando F pertence à classe de equivalência definida por **s**.

Tratando as esperanças como valores observados, a proporção de observações no intervalo $[q_j, p_j]$ é:

$$\pi_j(s) = \frac{\sum_{i=1}^{n}\left\{\mu_{ij}(s)+v_{ij}(s)\right\}}{M(s)} \tag{2.5}$$

em que,

$$M(s) = \sum_{i=1}^{N}\sum_{j=1}^{m}\left\{\mu_{ij}(s)+v_{ij}(s)\right\}$$

Dizemos que o vetor de probabilidades s é ***auto-consistente*** se

$$s_j = \pi_j(\boldsymbol{s}) \qquad (1 \le j \le m) \tag{2.6}$$

Turnbull (1976) definiu a estimativa *auto-consistente* de **s** como qualquer solução das equações simultâneas (2.6). O processo iterativo para encontrar a solução de (2.6) é descrito a seguir:

a) Obter estimativas iniciais s_j^0, pelo estimador de Kaplan-Meier;

b) Estimar $\mu_{ij}(s^0)$ e $v_{ij}(s^0)$ para $1 \le i \le N$ e $1 \le j \le m$, e assim $M(s^0)$ e $\pi_j(s^0)$;

c) Melhorar as estimativas s_j^1 definindo $s_j^1 = \pi_j(s^0)$ para $1 \le j \le m$;

d) Retornar para a etapa B, com s^1 substituindo s^0, e assim por diante;

e) Parar quando o rigor necessário tenha sido alcançado.

Por exemplo, $\left|s_j^u - s_j^{u-1}\right| < 0,001$.

Assim, a estimativa de máxima verossimilhança de F é dada por

$$\hat{F}(x) = \begin{cases} 0 & se \quad x < q_1 \\ \hat{s}_1 + \hat{s}_2 + \ldots + \hat{s}_j & se \quad p_j < x < q_{j+1} \\ 1 & se \quad x > p_m \end{cases}$$

e é indefinida para $x \in \left[q_j, p_j \right]$, para $1 \le j \le m$.

A partir da estimativa da função de distribuição acumulada encontra-se a estimativa da função de sobrevivência, *S(t),* uma vez que *S(t) = 1 - F(t).*

O algoritmo auto-consistente de Turnbull foi implementado no software R por Giolo (2004) em um estudo em que se compara a eficiência de dois tipos de tratamento de câncer de mama. Com esta implementação no software estatístico R, o algoritmo auto-consistente se tornou simples de aplicar. Isto contrasta com o método Newton-Raphson usado por Peto (1973) para encontrar o ENPMV, pois este último trabalha com derivadas de primeira e segunda

ordem de uma função de probabilidades, tornando o cálculo mais trabalhoso. Dempster, Laird e Rubin (1977) em sua teoria de máxima verossimilhança para variáveis latentes, desenvolveu o algoritmo "EM" e suas propriedades. O método descrito acima pode ser visto como um exemplo de um algoritmo EM (Esperança-Maximização).

2.6 Estimador Bootstrap de curvas de sobrevivência com censura intervalar

Ao analisar dados entomológicos com censura intervalar, Gouvêa (2006) tentou utilizar o algoritmo EM (Esperança-Maximização), com o intuito de comparar dois tipos de alimentos artificiais para abelhas. Este algoritmo é um método utilizado para computar estimativas não paramétricas de máxima verossimilhança.

Gouvêa (2006) mostrou que não há diferenças nas ENPMV dos saltos quando se aplica o algoritmo EM ou quando se determinam tais estimativas por meio de contagens, para o tipo de dados analisados.

Como eram 10 abelhas em cada réplica, a probabilidade de morte no intervalo que continha uma observação era de 1/10.

Com as 10 réplicas de cada alimento, Gouvêa (2006) obteve 10 ENPMV via contagem. As 10 estimativas foram consideradas como uma amostra a partir da qual se aplicou o método Bootstrap para melhorar as estimativas de cada salto P_j.

O método *Bootstrap* foi aplicado para estimar os saltos e, posteriormente, construir as curvas de sobrevivência. Além disso, esse método possibilita a construção de intervalos de confiança que permitem

comparar as curvas de sobrevivência sem a necessidade de se estabelecer um método especifico.

Foram retiradas 1000 amostras Bootstrap com reposição para estimar as curvas de sobrevivência por meio de saltos.

Assim, estimou-se a média de cada uma dessas 1000 amostras, denotada por $\overline{p}_{jb}$. Com essas médias, determinou-se a média Bootstrap dada pela expressão:

$$\overline{\overline{p}}_j = \frac{\sum_{b=1}^{1000} \overline{p}_{jb}}{1000}.$$

Com os saltos Bootstrap $\overline{\overline{p}}_j$, Gouvêa (2006) construiu as curvas de sobrevivência para cada alimento analisado.

O Bootstrap é um método computacional que determina medidas de precisão para cálculos estatísticos. Consiste de uma técnica de reamostragem, que aproxima a distribuição de uma função das observações pela distribuição empírica dos dados baseada em uma amostra finita. A amostragem é feita, com reposição, da amostra original, quando a distribuição é desconhecida (Bootstrap não-paramétrico). Neste caso, supõe-se que os dados observados são obtidos da função de distribuição empírica $\hat{F}(x)$, que é definida como uma distribuição discreta que dá probabilidade igual a 1/n a cada valor x_i, i= 1, ..., n. Intuitivamente, a suposição desta distribuição equiprovável, i.e., de peso 1/n para cada valor amostral x_i, tem suas raízes no desconhecimento da distribuição de probabilidade associada ao conjunto de dados. Podemos

fazer um paralelo primeiramente com Inferência Bayesiana que quando não se conhece a priori da distribuição, usa-se a chamada navalha de Occam, que pondera igualmente todos os dados e, posteriormente, com Física Estatística (mais intimamente ligado a Termodinâmica). Em Física Estatística, quando tratamos um sistema com muitas partículas livres (i.e., sem potencial de interação entre as mesmas) não sabemos as coordenadas generalizadas das mesmas no espaço de fases (posição r_i e velocidade v_i) do sistema de partículas quando o sistema apresenta energia fixa (sistema isolado). A hipótese então é considerar que cada ponto do espaço de fases é igualmente provável.

Tecnicamente, quando temos poucos dados observados e não sabemos a distribuição de probabilidade que os mesmos seguem (ou deveriam seguir), utilizamos o método Bootstrap para "gerar" mais dados amostrais e com isto realizar uma análise estatística mais confiável.

O estimador (não-paramétrico) Bootstrap é calculado a partir de um procedimento de contagem como indicado nos passos seguintes:

a) leitura de *n* dados de tempo de falha.

b) configuração dos *i* intervalos de censura.

c) contagem do número de dados no interior de cada intervalo *i* configurado no passo (b).

d) cálculo da probabilidade de ocorrência do evento no intervalo *i* na iteração 0 bootstrap.

e) cálculo dos "saltos" em cada intervalo *i* na iteração 0 bootstrap.

f) define o vetor bootstrap base (iteração 0).

g) cálculo dos valores das probabilidades em cada intervalo *i* nas iterações seguintes bootstrap *k*.

h) cálculo dos "saltos" em cada intervalo *i* na iteração bootstrap *k*.

Ao final deste processo obtêm-se a função de sobrevivência para um número especifico de iterações (*k1*) bootstrap.

A seguir, apresenta-se a distribuição Weibull que será usada posteriormente como referência para a avaliação do método de Turnbull em estudos com réplicas.

2.7 Distribuição Weibull

A distribuição Weibull é muito utilizada em estudos clínicos, entomológicos, industriais, dentre outras áreas. A grande aplicabilidade desta distribuição se deve ao fato dela apresentar muitas formas, em que sua função taxa de falha é sempre monótona, ou seja, crescente, decrescente ou constante.

Para uma variável aleatória T com distribuição Weibull, as funções de sobrevivência, de risco e a função densidade de probabilidade são, respectivamente,

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \tag{2.7}$$

$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \tag{2.8}$$

e

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}. \tag{2.9}$$

em que, γ é o parâmetro de forma, α é o parâmetro de escala e ambos são positivos. Além disso, $t \geq 0$. O parâmetro α tem a mesma unidade de medida de t, e γ não tem unidade.

A distribuição Weibull, com parâmetro $\gamma > 1$, tem função risco estritamente crescente. Tal distribuição representa bem um conjunto de dados como o que vamos utilizar. Se $\gamma < 1$ a função risco é estritamente decrescente. Para $\gamma=1$ a função risco é constante e temos a distribuição exponencial que é um caso particular da Weibull.

3 MATERIAL E MÉTODOS

Nesta seção descreveremos a metodologia e os dados utilizados neste trabalho.

3.1 Material

Gouvêa (2006) relata um experimento realizado com abelhas do Apiário Experimental da Universidade Federal de Lavras com o intuito de comparar dois tipos de alimentos artificiais. Neste experimento, duzentas abelhas, da espécie *Apis mellifera L.,* foram divididas em dois grupos de 100 abelhas, em que cada grupo recebeu um dos seguintes tratamentos:

a) Tratamento 1: solução aquosa de mel em 50%;

b) Tratamento 2: solução aquosa de frutose em 50%.

Para cada tratamento as abelhas foram separadas em 10 subgrupos, contendo 10 abelhas cada, que constituem as réplicas. Cada réplica consistiu de 10 abelhas em uma gaiola de PVC, que possuía a tampa perfurada, por onde foi inserido o alimento. As gaiolas foram mantidas em sala climatizada e continha, cada uma, um recipiente de vidro com capacidade para 20 ml do alimento artificial. Em cada réplica, foi oferecido às abelhas um chumaço de algodão embebido em água destilada contendo o alimento.

Os tempos de observação eram fixos, a cada 12 horas. Foram observados 22 intervalos de tempo. A cada observação, registrou-se a ocorrência ou não do evento, morte das abelhas. Assim, a cada ocorrência do evento sabia-se apenas que o tempo de vida do inseto pertencia ao intervalo, com duração de 12 horas, desde a última observação e a observação atual.

Os dados experimentais apresentam as seguintes médias e variâncias:

a) Tratamento 1: $\overline{X} = 144$ h e $s^2 = 3312$ h^2;

b) Tratamento 2: $\overline{X} = 99$ h e $s^2 = 2088$ h^2.

Com o propósito de avaliar a capacidade e/ou possibilidade de extensão do algoritmo de Turnbull, quando aplicado em experimentos com réplicas, foi realizado um estudo via simulação tomando como referência a função de sobrevivência de uma distribuição Weibull.

Na simulação, realizada no software R (R DEVELOPMENT CORE TEAM, 2009), gerou-se uma população de tamanho 3000 de uma distribuição Weibull para o tempo T até a ocorrência do evento, para cada um dos tratamentos. Optou-se por gerar uma população de tamanho 3000 porque foi suficiente para representar a função de sobrevivência da distribuição Weibull. Ao gerar uma população de tamanho 5000, verificou-se que fornece a mesma função de sobrevivência. Entretanto, gerar uma população de tamanho 100 forneceria uma função de sobrevivência diferente daquela construída com 3000 dados. Ou seja, com 100 dados não estaríamos representando bem a função de sobrevivência da distribuição Weibull.

Os parâmetros de forma e de escala da distribuição Weibull simulada foram estabelecidos considerando a média estimada e a variância estimada de cada tratamento, através das seguintes relações:

$$E[X] = \gamma^{-1/\alpha}\Gamma\left(1+\alpha^{-1}\right)$$

e

$$Var[X] = \gamma^{-2/\alpha}\left[\Gamma\left(1+2\alpha^{-1}\right) - \Gamma^2\left(1+\alpha^{-1}\right)\right]$$

em que, γ é o parâmetro de forma e α é o parâmetro de escala.

Assim, foram geradas populações de tamanho 3000, oriundas das seguintes distribuições:

a) Tratamento 1: T ~ Weibull ($\gamma = 2,69$; $\alpha = 161,93$);
b) Tratamento 2: T ~ Weibull ($\gamma = 2,29$; $\alpha = 111,75$).o

Para cada tipo de alimento fornecido às abelhas, retirou-se aleatoriamente da respectiva população obtida através de simulação uma amostra de tamanho 500. Cada amostra de tamanho 500 foi dividida em 10 subgrupos de tamanho 50, de maneira que, esses subgrupos sejam considerados réplicas. Assim, obteve-se através de simulação um experimento com 10 réplicas, contendo cada uma 50 dados.

3.2 Métodos

Para estudar a aplicabilidade do algoritmo de Turnbull em um experimento replicado, utilizar-se-á os dados da distribuição Weibull simulada, descrita anteriormente. Tal distribuição tem função risco semelhante ao gráfico 2.

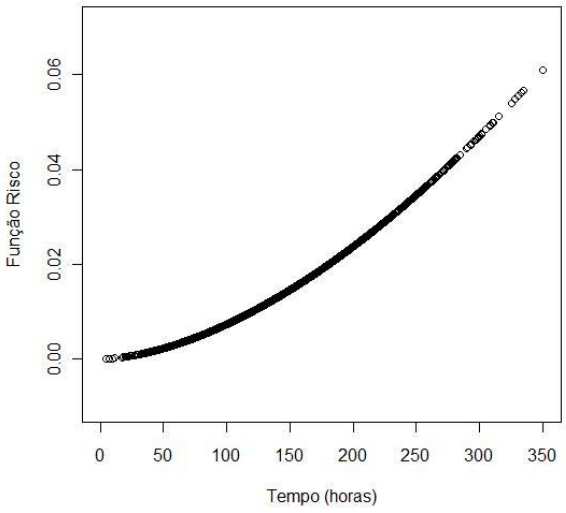

Gráfico 2 Função risco de uma distribuição Weibull simulada com 3000 dados e com os parâmetros de forma γ=2,69 e de escala α=161,93

Como o foco do nosso estudo é a estimação da função de sobrevivência com dados censurados intervalarmente em experimentos com réplicas, estipulou-se previamente que seriam 40 intervalos de censura para as réplicas geradas pela distribuição Weibull.

Para determinar tais intervalos arredonda-se o maior elemento da amostra retirada da população simulada e seleciona-se o menor número inteiro maior que este. Após isto, divide-se este número inteiro pelo número de intervalos desejados.

Exemplificando: Para a réplica 5 simulada, o maior valor encontrado foi 281,176. Assim, o limite superior do último intervalo de censura será 282. Para encontrar cada intervalo, dividimos esse valor pelo número de intervalos. Daí, 282/40 = 7,05, que será a amplitude de cada intervalo. A partir daí determina-se o número de observações existentes em cada um dos intervalos (figura 1).

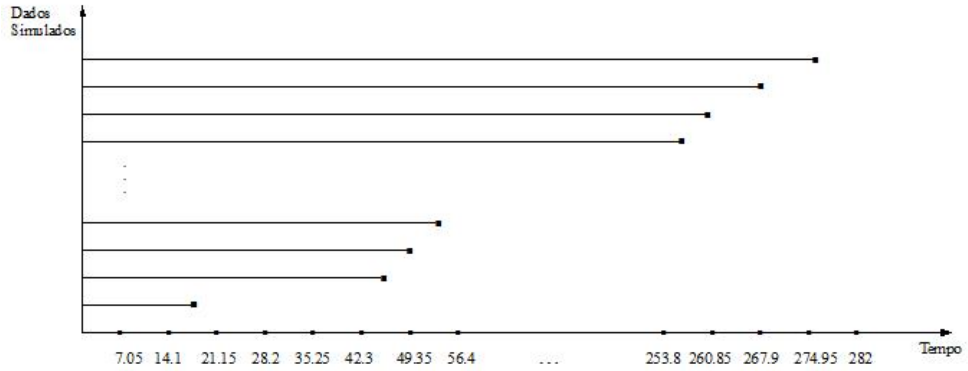

Figura 1 Esquematização do tempo até a ocorrência do evento para 8 dos 50 dados simulados para a réplica 5 gerada de uma distribuição de Weibull com os parâmetros 2,69 (forma) e 161,93 (escala)

Observa-se também na figura 1 que há uma observação no intervalo (14,1; 21,15], duas observações no intervalo (42,3; 49,35], uma no intervalo (49,35; 56,4], e assim por diante.

Para cada réplica foi necessário designar os 50 elementos aos intervalos que os contém e aplicar o algoritmo de Turnbull. Para iniciar o processo iterativo do algoritmo, calculou-se a estimativa de Kaplan-Meier nos extremos dos intervalos de censura, e conforme mencionado na seção 2.5, encontrou-se o valor dos saltos nos intervalos através da seguinte equação:

$$s_j = F(p_j+) - F(q_j -), \qquad \text{para } j = 1, 2, ..., m;$$

em que,

$F(p_j+)$ é o valor da função de distribuição no ponto p_j quando tendemos ao número pela direita;

$F(q_j-)$ é o valor da função de distribuição no ponto q_j quando tendemos ao número pela esquerda.

Partindo destas estimativas iniciais, determinou-se a estimativa não paramétrica pelo algoritmo auto-consistente de Turnbull, e determinou-se a curva de sobrevivência para cada réplica a fim de compará-la à curva simulada a partir da distribuição de Weibull.

Realizou-se este processo para as réplicas individuais e para a amostra conjunta. A partir, daí propôs-se uma extensão do método de Turnbull para uma possível aplicação em réplicas.

Repetiu-se o procedimento com os dados experimentais relativos ao tempo de vida das abelhas (Anexo). Ou seja, para os dados reais determinou-se a curva de sobrevivência das réplicas e da amostra conjunta através do algoritmo de Turnbull para compará-las à curva simulada a partir da distribuição Weibull.

Após essas etapas, comparou-se os resultados da aplicação do algoritmo de Turnbull aos experimentos com réplicas com os resultados obtidos pelo método proposto por Gouvêa (2006).

4 RESULTADOS E DISCUSSÃO

A partir da aplicação do algoritmo de Turnbull em um experimento com réplicas, obteve-se resultados sob diferentes aspectos, sejam eles relacionados ao número de intervalos de censura, ao número de observações, à estimativa da função de sobrevivência obtida pelo algoritmo, entre outros. Tais resultados são apresentados a seguir:

4.1 Avaliação do algoritmo de Turnbull extendido com dados simulados

No gráfico 3 está representada, em vermelho, uma curva de sobrevivência simulada a partir de uma distribuição de Weibull, com parâmetros de forma e de escala iguais a 2,69 e 161,93, respectivamente. A função escada, em preto, no gráfico 3 (a) e (b), corresponde à curva obtida através da aplicação do algoritmo de Turnbull, para duas quantidades diferentes de intervalos de censura.

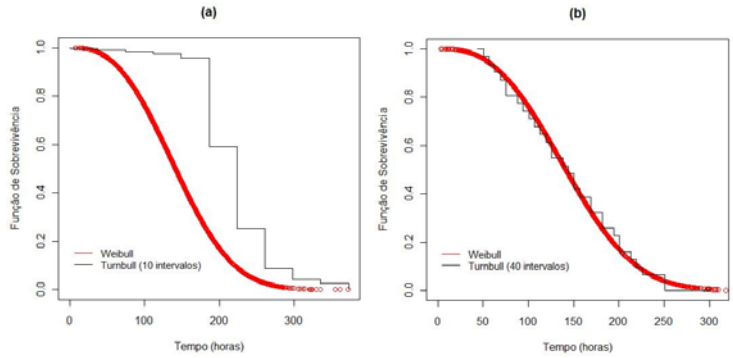

Gráfico 3 Curvas de sobrevivência, em vermelho, simulada a partir de uma distribuição Weibull em ambos os gráficos; e as curvas, em preto, estimadas pelo algoritmo de Turnbull com 10 intervalos no gráfico

(a), e com 40 intervalos no gráfico (b)

Pode-se observar no gráfico 3, que ao aplicar o algoritmo de Turnbull nos dados simulados a partir de uma distribuição Weibull, para 10 intervalos de censura o algoritmo não forneceu uma boa estimativa da curva de sobrevivência gerada pela distribuição Weibull. Em contrapartida, para 40 intervalos de censura o algoritmo forneceu uma boa estimativa da curva de sobrevivência simulada. Para valores maiores que 40, o algoritmo de Turnbull forneceu estimativas semelhantes. Assim, deve-se aplicar o algoritmo de Turnbull em experimentos cujo número de intervalos de censura não seja pequeno. Por isso, optou-se por fixar 40 intervalos de censura para os dados simulados.

No gráfico 4, são apresentados os gráficos com a curva de sobrevivência construída a partir de uma simulação feita com 3000 dados originados de uma distribuição Weibull (em verde) e a curva estimada pelo algoritmo de Turnbull (em preto) para as réplicas 1, 2, 3 e 4.

No gráfico 5, são apresentados os gráficos com a curva de sobrevivência construída a partir de uma simulação feita com 3000 dados originados de uma distribuição Weibull (em verde) e a curva estimada pelo algoritmo de Turnbull (em preto) para as réplicas 5, 6, 7, 8, 9 e 10.

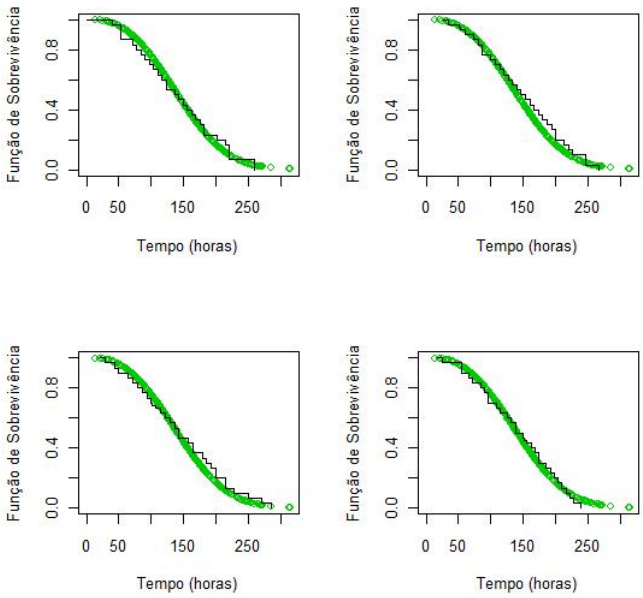

Gráfico 4 Curvas de sobrevivência, em verde, simulada a partir de uma distribuição Weibull com parâmetros de forma e escala iguais a 2,69 e 161,93, respectivamente, e as curvas estimadas pelo algoritmo de Turnbull (em preto) para as réplicas 1, 2, 3 e 4

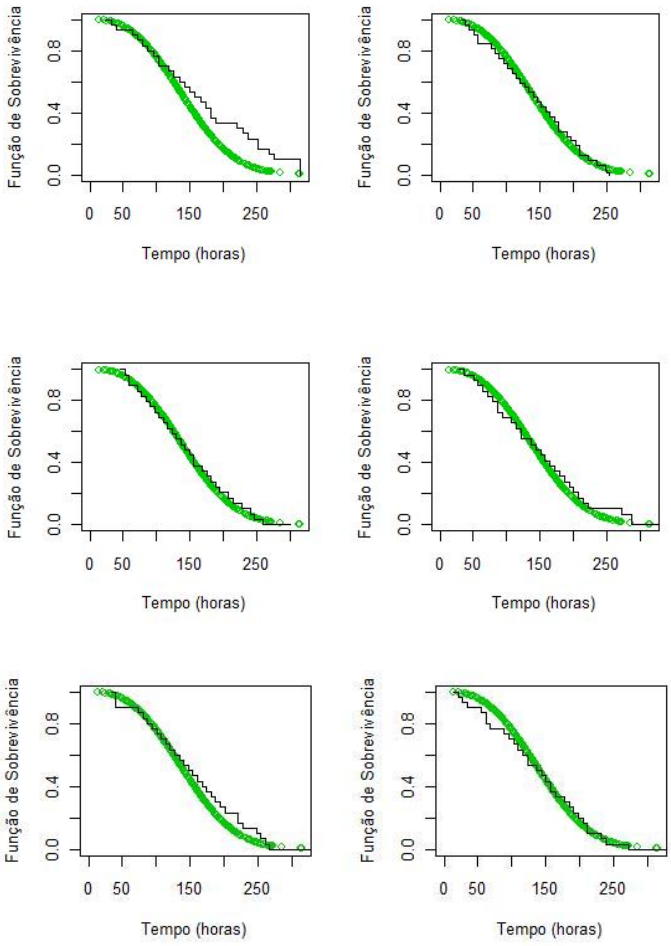

Gráfico 5 Curvas de sobrevivência, em verde, simulada a partir de uma distribuição Weibull com parâmetros de forma e escala iguais a 2,69 e 161,93, respectivamente, e as curvas estimadas pelo algoritmo de Turnbull (em preto) para as réplicas 5, 6, 7, 8, 9 e 10

Pode-se observar nos gráficos 4 e 5, que o algoritmo forneceu boas estimativas da curva de sobrevivência construída a partir da distribuição Weibull quando aplicado às réplicas. Pode-se observar que as curvas de sobrevivência estimadas pelo algoritmo se sobrepõem às curvas simuladas.

No gráfico 6, está representada a curva de sobrevivência construída a partir de uma simulação feita com 3000 dados originados de uma distribuição Weibull (em vermelho) e a curva estimada pelo algoritmo de Turnbull (em preto) para a amostra conjunta.

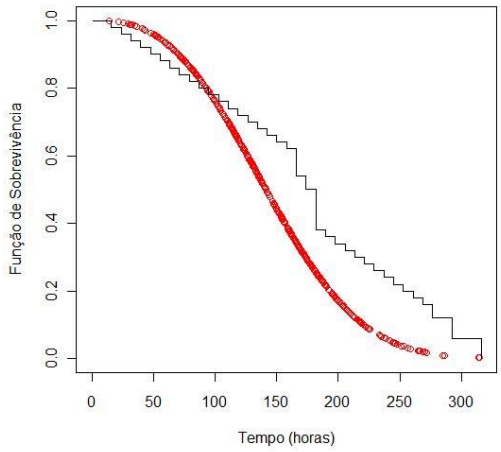

Gráfico 6 Curva de sobrevivência, em vermelho, simulada a partir de uma distribuição Weibull com parâmetros de forma e escala iguais a 2,69 e 161,93, respectivamente, e a curva estimada pelo algoritmo de Turnbull, em preto, para a amostra conjunta de todas as réplicas

Pode-se observar no gráfico 6 que o algoritmo de Turnbull não forneceu uma boa estimativa da curva de sobrevivência construída a partir da distribuição Weibull quando aplicado à amostra conjunta de todas as réplicas.

Assim, com 40 intervalos de censura, verificou-se que o algoritmo forneceu boas estimativas da curva de sobrevivência construída a partir da distribuição Weibull quando aplicado às réplicas. Entretanto, o mesmo não ocorre quando o algoritmo é aplicado à amostra conjunta.

4.2 Aplicação do algoritmo de Turnbull aos dados experimentais

No gráfico 7 estão representadas a curva de sobrevivência construída a partir da simulação feita com 3000 dados originados de uma distribuição Weibull em ambos os gráficos, e a curva estimada pelo algoritmo de Turnbull para a réplica 7 do alimento mel no gráfico 6a, e para a amostra conjunta de todas as réplicas, do mesmo alimento, no gráfico 6b.

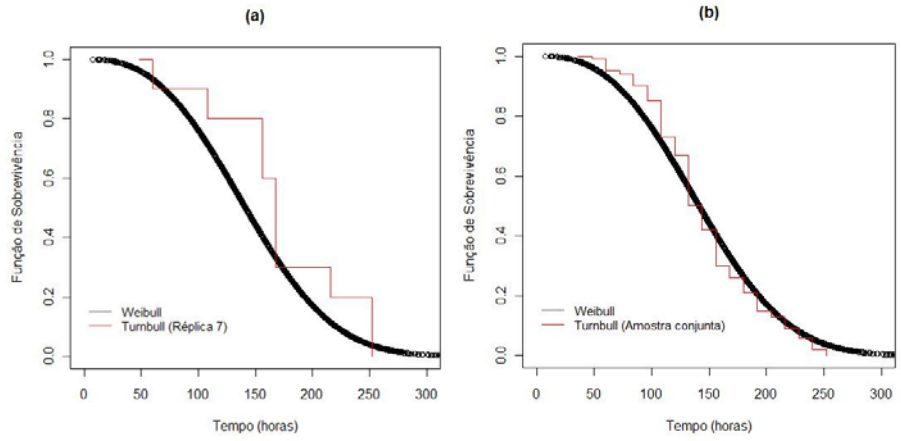

Gráfico 7 Curva de sobrevivência simulada a partir de uma distribuição Weibull com parâmetros 2,69 (forma) e 161,93 (escala), em ambos os gráficos. A função escada representa a estimativa obtida pelo algoritmo de Turnbull para a réplica 7 do alimento mel no gráfico (a), e para a amostra conjunta do mesmo alimento no gráfico (b)

Ao aplicar o algoritmo de Turnbull aos dados experimentais, verificou-se que o algoritmo não forneceu boas estimativas da curva de sobrevivência construída a partir da distribuição de Weibull quando aplicado às réplicas do experimento com o alimento mel.

Pode-se observar no gráfico 7 que a curva de sobrevivência estimada pelo algoritmo não se sobrepõe à curvas simulada quando aplicado à réplica (gráfico 7a). Entretanto, o mesmo não ocorre quando o algoritmo é aplicado à amostra conjunta de todas às réplicas (gráfico 7b). Ou seja, quando aplicado à amostra conjunta o algoritmo forneceu uma estimativa que se aproxima da curva simulada.

No experimento com as abelhas, cada réplica era composta de 10 insetos apenas. É um número pequeno de amostra. Assim, notou-se que o algoritmo de Turnbull não se ajusta bem para amostras pequenas, principalmente se há uma concentração da maioria das observações em poucos intervalos.

Observa-se no gráfico 7a que a probabilidade de uma abelha sobreviver a t=150 horas obtida pelo algoritmo é de 80% ao passo que a probabilidade fornecida pela distribuição Weibull neste mesmo tempo é de aproximadamente 45%.

No gráfico 8 estão representadas a curva de sobrevivência construída a partir da simulação feita com 3000 dados originados de uma distribuição Weibull em ambos os gráficos, e a curva estimada pelo algoritmo de Turnbull para a réplicas 7 do alimento frutose no gráfico 8a, e para a amostra conjunta de todas as réplicas no gráfico 8b.

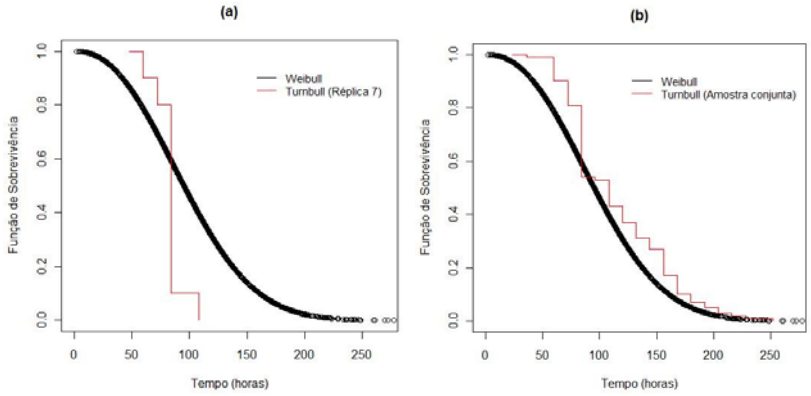

Gráfico 8 Curva de sobrevivência simulada a partir de uma distribuição Weibull com parâmetros 2,29 (forma) e 111,75 (escala), em ambos os gráficos. A função escada foi estimada pelo algoritmo de Turnbull para a réplica 7 do alimento frutose no gráfico (a), e para a amostra conjunta do mesmo alimento no gráfico (b)

Observa-se no gráfico 8 que os resultados obtidos realizando o mesmo procedimento com os dados do experimento com o alimento frutose foram análogos aos resultados obtidos para o alimento mel. A curva de sobrevivência estimada pelo algoritmo de Turnbull se assemelha à curva simulada para a amostra conjunta, porém, o mesmo não ocorreu para as réplicas individuais.

No gráfico 9, são apresentadas as curvas de sobrevivência estimadas através do algoritmo de Turnbull para os dois alimentos, mel e frutose, avaliados no experimento em questão.

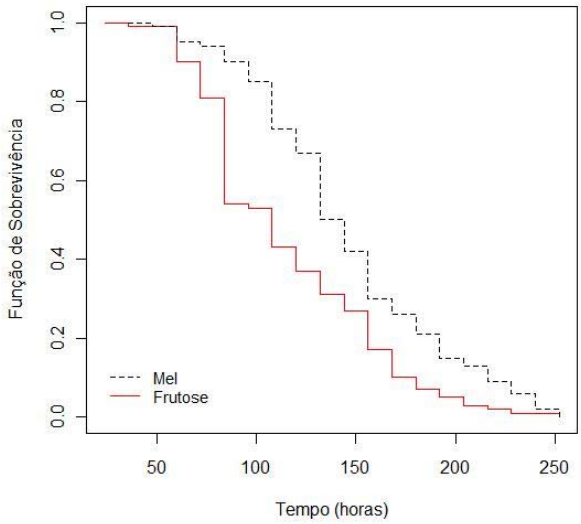

Gráfico 9 Curvas de sobrevivência estimadas pelo algoritmo de Turnbull; para o alimento 2 (frutose) representado pela linha contínua, e para o alimento 1 (mel) representado pela linha tracejada

De acordo com as estimativas obtidas pelo algoritmo de Turnbull para os dois alimentos fornecidos às abelhas no experimento, o alimento 1, solução aquosa de mel em 50%, proporciona maior tempo de sobrevida às abelhas que o alimento 2, solução aquosa de frutose em 50% (gráfico 9). Gouvêa (2006) encontrou este mesmo resultado ao aplicar o estimador Bootstrap não-paramétrico de curvas de sobrevivência por ela proposto, porém com curvas estimadas diferentes.

Entretanto, para verificar que tais estimativas são estatisticamente diferentes e afirmar que a probabilidade das abelhas alimentadas com mel sobreviverem a um determinado tempo “t” é maior que a probabilidade das abelhas alimentadas com frutose sobreviverem a este mesmo tempo “t” é

necessário realizar um teste de hipóteses ou encontrar intervalos de confiança para as funções de sobrevivência estimadas.

Embora o objetivo deste trabalho não seja a construção de testes de hipóteses ou intervalos de confiança para comparar curvas de sobrevivência estimadas, este assunto será discutido na seção 6.

4.3 Comparação das curvas de sobrevivência obtidas pelo algoritmo de Turnbull e pelo Estimador Bootstrap proposto por Gouvêa (2006)

Gouvêa (2006) utilizou o método Bootstrap a fim de melhorar as estimativas encontradas. Neste estudo, aplicamos o método Bootstrap com N_1=100, N_2=500 e N_3=1000 reamostragens. No gráfico 10, estão representadas a curva simulada a partir de uma distribuição de Weibull com os parâmetros de forma igual a 2,69 e de escala igual a 161,93; e as curvas de sobrevivência estimadas com os três tamanhos de amostras bootstrap diferentes.

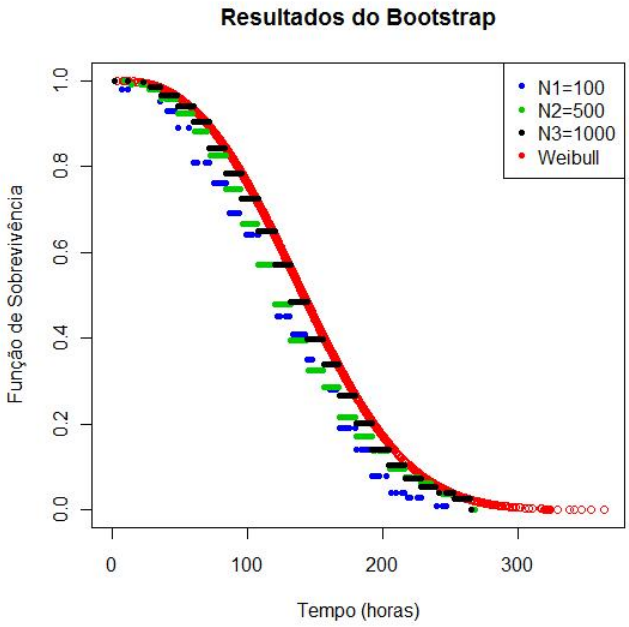

Gráfico 10 Curva de sobrevivência, em vermelho, simulada a partir de uma distribuição Weibull e as estimativas bootstrap obtidas com três tamanhos de amostras diferentes: N_1=100, em azul, N_2=500, em verde e N_3=1000 em preto

Pode-se observar no gráfico 10 que, quanto maior for o número de amostras bootstrap, melhor será a estimativa encontrada. Assim, o Bootstrap com N=1000 reamostragens foi que melhor estimou a distribuição Weibull com os parâmetros do alimento mel, no presente estudo.

A partir daí, comparou-se a estimativa Bootstrap com N=1000 reamostragens com a estimativa encontrada através do algoritmo de Turnbull para os dados simulados de uma distribuição Weibull.

No gráfico 11, são apresentadas as duas estimativas de uma curva simulada a partir da distribuição Weibull com parâmetros de forma e de escala

iguais a 2,69 e 161,93, respectivamente, através do Estimador Bootstrap e através do algoritmo de Turnbull.

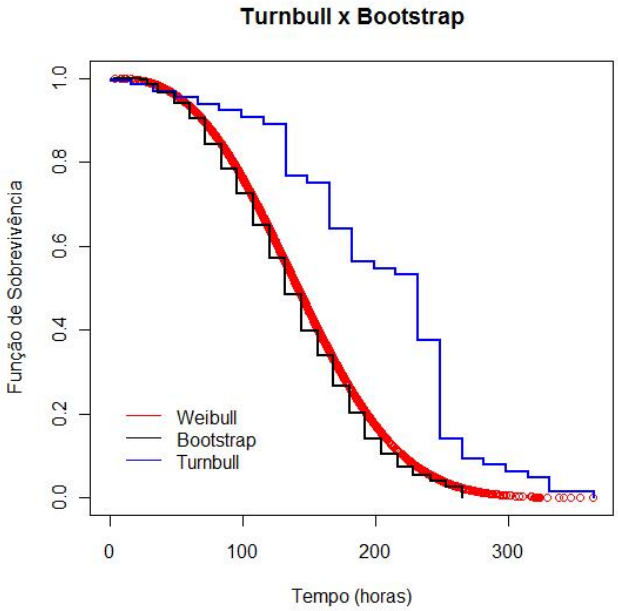

Gráfico 11 Curva de sobrevivência, em vermelho, simulada a partir de uma distribuição Weibull, a estimativa bootstrap obtida com N_3=1000, em preto, e a curva de sobrevivência estimada pelo algoritmo de Turnbull, em azul

Observa-se no gráfico 11, que a curva obtida através do Bootstrap (em preto) se sobrepõe à curva simulada a partir da distribuição Weibull (em vermelho), fornecendo uma boa estimativa. Entretanto, o algoritmo de Turnbull não fornece uma boa estimativa (em azul) da distribuição Weibull simulada.

5 CONCLUSÕES

Para dados obtidos através de simulação verificou-se que o algoritmo de Turnbull não fornece uma boa estimativa da curva de sobrevivência simulada quando o número de intervalos de censura é pequeno.

Para 40 ou mais intervalos de censura, o algoritmo de Turnbull fornece boas estimativas da curva de sobrevivência construída a partir da distribuição Weibull quando aplicado às réplicas individuais. Porém, o mesmo não ocorre quando o algoritmo é aplicado à amostra conjunta de todas às réplicas, pois a curva estimada pelo algoritmo de Turnbull não se assemelha à curva simulada.

Ao aplicar o algoritmo de Turnbull aos dados experimentais, o algoritmo não fornece boas estimativas da curva de sobrevivência quando aplicado às réplicas. Entretanto, quando aplicado à amostra conjunta de todas às réplicas a estimativa melhora.

Ao comparar a curva de sobrevivência estimada pelo algoritmo de Turnbull com a curva simulada pela distribuição Weibull para a amostra conjunta, os resultados obtidos por simulação não foram satisfatórios. O estimador Bootstrap proposto por Gouvêa (2006) forneceu uma excelente estimativa da curva simulada a partir da distribuição Weibull para um experimento com réplicas. Assim o método proposto por Gouvêa (2006) fornece estimativas melhores que o algoritmo de Turnbull para experimentos com réplicas.

Para obter resultados mais confiáveis e exatos tanto na comparação entre as curvas de sobrevivência estimadas para cada alimento quanto na comparação da curva de sobrevivência estimada através do algoritmo de Turnbull com a curva simulada a partir de uma distribuição Weibull é necessária a construção de intervalos de confiança para cada uma das curvas estimadas.

6 TRABALHOS FUTUROS

Este trabalho não se encerra com os resultados já obtidos. Com o intuito de aperfeiçoá-lo, pretende-se ainda:

1) Desenvolver uma metodologia para testar se as curvas geradas pela distribuição Weibull e pelo algoritmo de Turnbull são estatisticamente iguais, ou diferentes.

2) Construir intervalos de confiança ou um teste de hipóteses para comparação de curvas de sobrevivência estimadas a partir do método de Turnbull.

3) Simular para várias distribuições diferentes a metodologia proposta para estimar a curva de sobrevivência obtida através do método de Turnbull para várias réplicas.

4) Extrair da população gerada por simulação várias amostras de mesmo tamanho, aplicar ao algoritmo de Turnbull a cada uma das amostras e assim tentar obter intervalos de confiança.

REFERÊNCIAS

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: E. Blücher, 2006. 370 p.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society, Series B**, Oxford, v. 39, n. 1, p. 1-38, 1977.

EFRON, B. The two sample problem with censored data. In: SYMPOSIUM ON MATHEMATICS STATISTICS PROBABILITY, 5., 1967, Berkeley. **Proceedings…** Berkeley: SSP, 1967. p. 831-853.

GIOLO, S. R. **Turnbull´s nonparametric estimator for interval-censored data:** technical reports. Curitiba: UFPR, 2004. Disponível em: <http://www.est.ufpr.br/rt/suely04a.pdf>. Acesso em: 10 maio 2010.

GOUVEA, G. D. R. **Estimador bootstrap não-paramétrico de curvas de sobrevivência para dados entomológicos com censura intervalar**. 2006. 61 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2006.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, New York, v. 53, n. 282, p. 457-481, 1958.

PETO, R. Experimental survival curves for interval-censored data. **Applied Statistics**, London, v. 22, p. 86-91, 1973.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <http://www.R-project.org>. Acesso em: 10 abr. 2010.

TURNBULL, B. W. Empirical distribution function with arbitrarily grouped, censored and truncated data. **Journal of the Royal Statistical Society, Series B**, Oxford, v. 38, p. 290-295, 1976.

_____. Nonparametric estimation of a survivorship function with doubly censored data. **Journal American Statistical Association**, New York, v. 69, p. 169-173, 1974.

YU, Q.; LI, L.; WONG, G. Y. C. On consistency of the self-consistent estimator of survival functions with interval-censored data. **Scandinavian Journal of Statistics**, Oxford, v. 27, p. 35-44, 2000.

ANEXO

ANEXO A – Dados Experimentais

Tabela 1 – Dados reais, referentes ao tempo de vida de abelhas coletadas no Apiário Experimental da Universidade Federal de Lavras – UFLA.

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
12	1	0	0
12	2	0	0
12	3	0	0
12	4	0	0
12	5	0	0
12	6	0	0
12	7	0	0
12	8	0	0
12	9	0	0
12	10	0	0
24	1	0	0
24	2	0	0
24	3	0	0
24	4	0	0
24	5	0	0
24	6	0	0
24	7	0	0
24	8	0	0
24	9	0	0
24	10	0	0

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
36	1	0	1
36	2	0	0
36	3	0	0
36	4	0	0
36	5	0	0
36	6	0	0
36	7	0	0
36	8	0	0
36	9	0	0
36	10	0	0
48	1	0	0
48	2	1	0
48	3	0	0
48	4	0	0
48	5	0	0
48	6	0	0
48	7	0	0
48	8	0	0
48	9	0	0
48	10	0	0
60	1	1	1
60	2	0	0
60	3	0	1
60	4	0	2
60	5	0	2
60	6	1	0
60	7	1	1

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
60	8	0	1
60	9	0	0
60	10	1	1
72	1	0	2
72	2	0	2
72	3	1	1
72	4	0	2
72	5	0	0
72	6	0	0
72	7	0	1
72	8	0	0
72	9	0	0
72	10	0	1
84	1	0	2
84	2	0	1
84	3	1	1
84	4	0	2
84	5	1	1
84	6	0	5
84	7	0	7
84	8	0	2
84	9	2	3
84	10	0	3
96	1	0	0
96	2	1	1
96	3	1	0
96	4	0	0

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
96	5	0	0
96	6	0	0
96	7	0	0
96	8	2	0
96	9	0	0
96	10	1	0
108	1	2	1
108	2	2	1
108	3	0	1
108	4	2	0
108	5	2	1
108	6	0	0
108	7	1	1
108	8	1	5
108	9	1	0
108	10	1	0
120	1	1	1
120	2	0	1
120	3	3	0
120	4	0	0
120	5	0	0
120	6	2	1
120	7	0	0
120	8	0	0
120	9	0	1
120	10	0	2
132	1	3	1

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
132	2	2	0
132	3	0	0
132	4	2	1
132	5	3	3
132	6	3	1
132	7	0	0
132	8	2	0
132	9	1	0
132	10	1	0
144	1	0	0
144	2	0	1
144	3	1	1
144	4	0	0
144	5	2	0
144	6	1	1
144	7	0	0
144	8	1	0
144	9	2	1
144	10	1	0
156	1	1	0
156	2	0	1
156	3	0	3
156	4	2	0
156	5	1	0
156	6	1	1
156	7	2	0
156	8	2	1

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
156	9	3	2
156	10	0	2
168	1	0	0
168	2	1	2
168	3	0	1
168	4	0	0
168	5	0	2
168	6	0	0
168	7	3	0
168	8	0	0
168	9	0	1
168	10	0	1
180	1	0	1
180	2	0	0
180	3	1	0
180	4	2	0
180	5	0	0
180	6	0	0
180	7	0	0
180	8	0	0
180	9	1	1
180	10	1	0
192	1	0	0
192	2	1	0
192	3	2	0
192	4	0	1
192	5	0	0

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
192	6	1	1
192	7	0	0
192	8	0	0
192	9	0	0
192	10	2	0
204	1	1	0
204	2	0	0
204	3	0	0
204	4	0	1
204	5	0	0
204	6	0	0
204	7	0	0
204	8	0	0
204	9	0	1
204	10	0	0
216	1	0	0
216	2	0	0
216	3	0	0
216	4	0	0
216	5	1	0
216	6	0	0
216	7	1	0
216	8	0	1
216	9	0	0
216	10	0	0
228	1	1	0
228	2	0	0

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
228	3	0	0
228	4	1	1
228	5	0	0
228	6	0	0
228	7	0	0
228	8	1	0
228	9	0	0
228	10	0	0
240	1	0	0
240	2	0	0
240	3	0	0
240	4	1	0
240	5	0	0
240	6	0	0
240	7	0	0
240	8	1	0
240	9	0	0
240	10	2	0
252	1	0	0
252	2	0	0
252	3	0	1
252	4	0	0
252	5	0	0
252	6	0	0
252	7	2	0
252	8	0	0
252	9	0	0

Tabela 1 - Continuação

		Nº de abelhas mortas para cada alimento	
Tempo observado (horas)	Réplica	Mel	Frutose
252	10	0	0
264	1	0	0
264	2	0	0
264	3	0	0
264	4	0	0
264	5	0	0
264	6	0	0
264	7	0	0
264	8	0	0
264	9	0	0
264	10	0	0