VIABILIDADE DA APLICAÇÃO DE MECANISMOS DE CENSURA TIPO I E ALEATÓRIA EM DADOS ENTOMOLÓGICOS

PAULO JOSÉ PEREIRA[1]
MÁRIO JAVIER FERRUA VIVANCO[2]

RESUMO – Apesar da grande quantidade de pesquisas e trabalhos publicados sobre o tempo de vida na área da entomologia, não é comum encontrar algum que tenha utilizado as técnicas de Análise de Sobrevivência. Com este trabalho, objetivou-se avaliar a possibilidade de aplicar mecanismos de censura em dados entomológicos e estudar a viabilidade da aplicação do teste log-rank. Foram utilizados dados reais entomológicos sem censura de machos e fêmeas da *Ceraeochrysa cubana* e, a partir deles, montou-se uma sub-rotina em SAS®, na qual, por meio de simulação, introduziram-se as censuras aleatória e do tipo I, aplicando, assim, o teste log-rank para comparação das curvas resultantes e para diferentes porcentagens de censura. Utilizando os dados reais e os gerados por simulação, foi detectado que, com um aumento da presença de censura, a precisão do teste log-rank foi diminuindo.

TERMOS PARA INDEXAÇÃO: Análise de sobrevivência, censura, tempo de vida.

VIABILITY OF THE APPLICATION OF TYPE I CENSORING MECHANISMS AND RANDOM IN ENTOMOLOGICAL DATA

ABSTRACT – In spite of the great amount of research and works published about the lifetime in the area of entomology, it is not common to find some which had utilized techniques of Survival Analysis. This work has the objectives to evaluate the possibility of applying censoring mechanisms in the entomological data and to study the viability of the application of log-rank test. Real entomological data without censoring of males and females of *Ceraeochrysa cubana* were utilized. By means of simulation a sub-routine was set up in SAS® in which, the random and type I censoring were introduced. Applying the log-rank test for comparing the resulting curves, this for different percentage of censure. Utilizing the real and simulation the data, it was detected with an increase of censoring presence, the accuracy of the log-rank test has been decreasing.

INDEX TERMS: Censoring, lifetime, survival analysis.

INTRODUÇÃO

Em um experimento que analisa o tempo de vida de insetos, o descarte da informação de interesse pela perda de indivíduos é uma forma incorreta de analisar os dados. Isso ocorre porque a presença das observações incompletas no estudo justifica-se pelo fato de elas fornecerem, ainda, informações sobre o tempo de vida. Sem a presença desses dados, não seria possível realizar uma boa estimação de parâmetros e, assim, fazer uma análise adequada.

Para realizar uma análise correta e confiável de um experimento como o citado acima, é fundamental a utilização das técnicas de Análise de Sobrevivência. Essa é a área da estatística que tem como principal interesse analisar dados de tempos de vida que envolvem algumas observações cujas informações são incompletas (dados censurados). Entre as diversas áreas de aplicação desses métodos, pode-se citar a medicina, a engenharia e as ciências sociais.

Áreas menos exploradas estão relacionadas com estudos entomológicos nos quais avaliam-se, entre outras, variáveis relacionadas com o tempo de sobrevivência de insetos-pragas quando o ambiente natural desses é, de alguma forma, manipulado. Entre os diversos trabalhos que analisam variáveis do tipo mencionado an-

1. Professor da Universidade Tiradentes – UNIT - paulojosepe@bol.com.br
2. Professor Adjunto do Departamento de Ciências Exatas da UNIVERSIDADE FEDERAL DE LAVRAS/UFLA – Caixa Postal 37 – 37200-000 – Lavras, MG. ferrua@ufla.Br

teriormente, pode-se citar: Gonçalves (1990), que fez um estudo comparando o ciclo total em dias de fêmeas e machos de *Podisus nigrolimbatus* e *Podisus connexivus*, alimentadas com lagartas de *Bombyx mori*, em que foi detectada uma diferença significativa entre as fêmeas dessas espécies. Silva (1991) utilizou um delineamento inteiramente casualizado para testar se há diferença entre os tempos de vida de um crisopídeo submetidos a diferentes temperaturas. Micheletti (1999), utilizando dados de longevidade retirados de levantamentos realizados em campo, estimou a sobrevivência de fêmeas e machos de *Cerconota anonella* ajustando a distribuição Weibull aos dados. Esses estudos anteriormente mencionados foram realizados utilizando-se a informação completa. Isso é, cada indivíduo foi acompanhado até a sua morte. Nesses estudos, observações com informação incompleta são geralmente descartadas.

Já na Análise de Sobrevivência, a principal característica dos dados em estudo é a observação incompleta sobre o tempo de falha do indivíduo, ou seja, a **censura**.

Segundo Lawless (1982), existem três tipos de censura que são mais utilizadas e que são encontradas tanto nos casos de censura à direita quanto de censura à esquerda. São elas:

Tipo I – aquela em que o estudo é encerrado após um período preestabelecido de tempo;

Tipo II – aquela em que o estudo é encerrado após ter ocorrido a falha em um número preestabelecido de indivíduos;

Aleatória – ocorre quando o indivíduo é retirado do estudo por causas alheias ao experimento.

Um tipo de censura que não é exatamente como as mencionadas por Lawless (1982), porém semelhante com a Censura Tipo II, foi aplicada em estudo desenvolvido por Miramontes & Souza (1996). Nesse estudo foi analisada a sobrevivência de cupins colocados em tubos de ensaio hermeticamente fechados e reunidos em grupos de 1, 2, 3, 4, 8, 16 indivíduos. O experimento foi encerrado quando todos os cupins solitários morreram. Encontrou-se que a sobrevida foi significativamente maior para grupos maiores em relação a indivíduos isolados.

Realizou-se este trabalho com o objetivo de avaliar a possibilidade de aplicar mecanismos de censura, tanto a aleatória como a do tipo I, em dados entomológicos e estudar a viabilidade da aplicação do teste log-rank na comparação de duas curvas de sobrevivência à medida que se introduz uma porcentagem cada vez maior de censura ao experimento.

MATERIAL E MÉTODOS

Para alcançar os objetivos propostos neste trabalho, foram utilizados dados reais de longevidade, ou seja, do início da fase adulta até a sua morte, de machos e fêmeas, da *Ceraeochrysa cubana*, acompanhada em ambiente laboratorial, em um estudo realizado por Silva (1991). esse inseto é um importante crisopídeo utilizado para o controle biológico natural de pragas de plantas cultivadas. Conforme Pereira & Vivanco (2001) esses dados seguem uma distribuição Weibull.

Utilizando uma sub-rotina desenvolvida em SAS® por Vivanco (1994), foi possível estimar os parâmetros da distribuição dos tempos de falha pelo método da máxima verossimilhança, mediante o método de Newton-Raphson. Com os estimadores encontrados, geraram-se, por meio do método de simulação de Monte Carlo, 5.000 amostras de tamanho igual ao das amostras originais, ou seja, 40 e 31 indivíduos machos e fêmeas, respectivamente. Posteriormente, nas amostras simuladas, foram introduzidas porcentagens de censuras, aleatória e tipo I iguais a 10, 20, 30, 40 e 50%.

Por fim, aplicou-se o teste log-rank para comparar as curvas de sobrevivência de machos e fêmeas e, assim poder verificar a eficiência do teste à medida que se introduziam porcentagens cada vez maiores de censura aos dados.

Máxima Verossimilhança

Uma estimativa pontual confiável, conforme Spiegel (1978), é obtida por intermédio da técnica conhecida como Estimação de Máxima Verossimilhança.

Supondo que a população tenha função densidade que contenha um parâmetro populacional θ a ser estimado por meio de determinada estatística. A função de densidade, então, pode ser denotada por $f(x,\theta)$. Admitindo uma amostra aleatória de tamanho n: $x_1, x_2, ..., x_n$, a função de densidade conjunta dessa amostra será:

$$L = f(x_1,\theta) f(x_2,\theta) ... f(x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

que é chamada Função de Verossimilhança. O estimador de máxima verossimilhança pode ser obtido maximizando o logaritmo da função de verossimilhança. Uma forma de maximizar e achar os pontos críticos é derivando em relação a θ e igualando-a a zero. Assim, tem-se que:

$$\frac{1}{f(x_1,\theta)}\frac{\partial f(x_1,\theta)}{\partial\theta}+\ldots+\frac{1}{f(x_n,\theta)}\frac{\partial f(x_n,\theta)}{\partial\theta}=0$$

Dessa equação, pode-se obter θ em termos da amostra aleatória.

O método é possível de generalização. No caso de vários parâmetros, fazem-se as derivadas parciais em relação a cada um deles e, posteriormente, igualam-se a zero.

Na maioria das vezes, o cálculo para a obtenção dos estimadores de máxima verossimilhança não é possível analiticamente; então, métodos iterativos de aproximação numérica, como o Método de Newton-Raphson, são utilizados.

Simulação das Amostras

Tornou-se cada vez mais freqüente em estudos científicos o uso dos métodos de simulação para estudar novos procedimentos estatísticos ou para comparar resultados de diferentes técnicas estatísticas já existentes. Os processos de simulação que envolvem componentes aleatórios são pertencentes ao método de Monte Carlo, o qual, de uma maneira bastante simplificada , é um algoritmo que consiste em simular dados a partir de uma seqüência pseudo-aleatória, baseada na distribuição uniforme (0,1). Esse método é aplicado neste estudo considerando o Teorema da Transformação Integral da Probabilidade, que é apresentado, conforme Mood et al. (1974), da seguinte forma:

Teorema 1: Teorema da Transformação Integral da Probablidade

1º) Se a variável aleatória contínua $X \sim F_X(\,.\,)$, então:

$U = F_X(X) \sim U(0,1)$;

2º) Se $U \sim U(0,1)$ e $X = F_X^{-1}(U)$, então $X \sim F_X(\,.\,)$.

O Teorema 1, associado ao método de Monte Carlo, permitiu simular 5.000 amostras de tempos de vida (longevidade) de machos e fêmeas, de tamanho 40 e 31, respectivamente, de acordo com o modelo probabilístico determinado segundo a metodologia explicada em (3.2.1).

Teste Log-rank

Para comparar dois grupos de indivíduos quanto à sobrevivência, pode-se aplicar o teste não-paramétrico log-rank. A estatística desse teste é obtida de forma similar ao teste de Mantel & Haenszel (1959), para combinar tabelas de contingência.

Para testar as hipóteses:

$$H_0: S_1(t) = S_2(t)$$

$$H_1: S_1(t) \neq S_2(t)$$

considere $t_1 < t_2 < \ldots < t_k$, os tempos de falhas distintos da amostra formada pela combinação das duas amostras correspondentes aos dois grupos de indivíduos.

Suponha que no tempo t_j ocorram d_j falhas, e n_j indivíduos estão sob risco em um tempo imediatamente anterior à t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra i, $i = 1, 2$ e $j = 1, \ldots, k$. Os dados podem ser dispostos em uma tabela de contingência, conforme a seguir:

TABELA 1 – Tabela de contingência gerada no tempo t_j.

	Grupo1	**Grupo 2**	**Total**
Falha	d_{1j}	d_{2j}	d_j
Não falha	$n_{1j}-d_{1j}$	$n_{2j}-d_{2j}$	n_j-d_j
Total	n_{1j}	n_{2j}	n_j

Pode-se observar que a distribuição de d_{1j} é hipergeométrica, já que se tem um grupo de indivíduos no tempo t_j e é feita uma amostragem sem reposição, sendo d_{1j} a variável aleatória que indica o número de falhas nesse grupo no tempo t_j, então:

$$P(d_{1j}) = \frac{C_{n_{1j},d_{1j}} \times C_{n_{2j},d_{2j}}}{C_{n_j,d_j}} \; ; \; d_{1j} = 0, 1, 2, \ldots, n_{1j}$$

A média de d_{1j} é $e_{1j} = n_{1j}d_j n_j^{-1}$; isso quer dizer que, se não houver diferença entre as populações no tempo t_j, d_j pode ser dividido entre as duas amostras de acordo com a razão entre o número de indivíduos sob risco em cada amostra e o número total sob risco.

A variância de d_{1j} é, então:

$$(V_j)_1 = n_{1j}(n_j-n_{1j})d_j(n_j-d_j)n_j^{-2}(n_j-1)^{-1}$$

Portanto, a estatística $d_{1j} - e_{1j}$ tem média 0 e variância $(V_j)_1$.

Se as k tabelas de contingência forem independentes, um teste aproximado para a igualdade das duas funções de sobrevivência pode ser baseado na estatística

$$T = \frac{\left[\sum_{j=1}^{k}(d_{1j}-e_{1j})\right]^2}{\sum_{j=1}^{k}(V_j)_1}$$

que segue uma distribuição assintóticamente χ^2 com 1 grau de liberdade, para grandes amostras.

RESULTADOS E DISCUSSÃO

Sendo a distribuição Weibull que melhor se ajusta aos dados de longevidade da *Ceraeochrysa cubana* determinados por Silva (1991), com os dados reais estimaram-se os parâmetros (α e β) dessa distribuição. Para o conjunto de machos, têm-se $\hat{\alpha}_M = 133,487$ e $\hat{\beta}_M = 2,389$, ao passo que para fêmeas, $\hat{\alpha}_F = 94,786$ e $\hat{\beta}_F = 2,788$. Assumindo os estimadores de α e β como parâmetros da distribuição Weibull, geraram-se a partir dessa distribuição 5000 amostras para machos e 5000 para fêmeas, aplicando, posteriormente, em cada amostra, os mecanismos de censura aleatória e tipo I.

Teste Log-rank

Na Tabela 1 são apresentados os resultados do teste não-paramétrico log-rank para comparar as curvas de sobrevivência de machos e fêmeas da *Ceraeochrysa cubana*.

TABELA 1 – Estatísticas do log-rank para comparação de curvas de sobrevivência: machos *versus* fêmeas, dados originais.

	Aleatória	**Tipo I**
Completos	11,555*	11,555*
10% censurados	9,7865*	8,8841*
20% censurados	9,7606*	6,4702*
30% censurados	10,3477*	3,7653 n.s.
40% censurados	9,1188*	2,9979 n.s.
50% censurados	6,8356*	2,3559 n.s.

*** Existe diferença significativa**

n.s.: Não existe diferença significativa

As estatísticas apresentadas na Tabela 1 mostram que mesmo censurando em até 50% as observações de machos e fêmeas, no caso da censura aleatória, o teste permanece significativo. Já no caso da censura tipo I, a partir da diminuição de 30% do tempo do experimento, as funções de sobrevivência começam a não apresentar diferenças significativas.

Os resultados da Tabela 1 induzem a uma incerteza em relação à eficiência do teste log-rank para comparar curvas de sobrevivência. Tal fato baseia-se no valor da estatística, que diminui à medida que a porcentagem de censura aumenta, isto é, segundo o log-rank, duas curvas de sobrevivência diferentes são cada vez mais parecidas quando a porcentagem de observações censuradas aumenta. A eficiência do teste log-rank foi medida aplicando-o em cada uma das 5.000 amostras simuladas.

Aplicando o log-rank para cada experimento simulado, foi possível determinar a proporção de testes significativos na comparação das funções de sobrevivência de machos *versus* fêmeas. Os resultados são mostrados na Tabela 2.

TABELA 2 – Proporção de testes log-rank significativos (α = 0,05). Curvas de Sobrevivência: machos *versus* fêmeas.

	Aleatória	**Tipo I**
10% censurados	0,9418	0,9112
20% censurados	0,9324	0,8544
30% censurados	0,9046	0,7600
40% censurados	0,8746	0,6056
50% censurados	0,8436	0,4082

Pode-se observar que à medida que foram introduzindo porcentagens de censura, aleatória e tipo I, a proporção de testes significativos foi diminuindo, e nos casos da censura tipo I, essa diminuição começa a ser mais acentuada a partir de 30%, porcentagem que iniciou a não significância entre as funções de sobrevivência na análise dos dados reais. Isso mostra que o teste vai perdendo a eficiência com a presença de porcentagens maiores de censura.

CONCLUSÕES

O teste log-rank, aplicado à longevidade com informação completa sobre os indivíduos, mostrou que as curvas de sobrevivência de machos e fêmeas são diferentes.

À medida que se introduz uma porcentagem maior de censura ao conjunto de dados, percebeu-se, via simulação, que a possibilidade de o teste log-rank apresentar resultados incorretos aumenta.

É possível aplicar os mecanismos de censura com uma certa restrição, já que a partir da introdução de 30% de censura, tanto na aleatória quanto na tipo I, o principal teste estatístico para a Análise de Sobrevivência passa a não ser confiável.

REFERÊNCIAS BIBLIOGRÁFICAS

GONÇALVES, L. **Biologia e capacidade predatória de Podisus nigrolimbatus Spínola, 1832 e Podisus connexivus Bergroth, 1891 (Hemiptera: Pentatomidae: asopinae) em condições de laboratório**. 1990. Dissertação (Mestrado em Entomologia) – Universidade Federal de Lavras, Lavras.

LAWLESS, J. F. **Statisticals models and methods for lifetime data**. New York: John Wiley, 1982. 580 p.

MANTEL, N.; HAENSZEL, W. Statistical aspects of the analysis of data from retrospective studies of disease. **Journal of the National Cancer Institute**, Washington, v. 22, p. 719-748, 1959.

MICHELETTI, S. M. F. B. **Estudos sobre a *Cerconota anonella* (SEEO., 1830) (Lep.: Oecophoridae) em gravioleira (*Annona muricata L.*) no estado de Alagoas**. 1999. 87 p. Tese (Doutorado em Entomologia) - Escola Superior de Agricultura de Luiz de Queiroz, Piracicaba.

MIRAMONTES, O.; SOUZA, O. The nonlinear dynamics of survival and social facilitation in termites. **Journal of Theoretical Biology**, London, v.181, n. 4, p. 373-380, Aug. 1996.

MOOD, A. M.; GRAYBILL, F. A; BOES, D. C. **Introduction to theory of statistics**. 3. ed. New York: Wiley & Sons, 1974. 842 p.

PEREIRA, P. J.; VIVANCO, M. J. F. Estudo da função de sobrevivência de um conjunto de dados entomológicos utilizando a distribuição Weibull, com a introdução de uma censura aleatória. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 46.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGROPECUÁRIA, 9., 2001. Piracicaba. **Anais...** Piracicaba: ESALQ/USP, 2001. p. 178 –180.

SILVA, R. L. **Aspectos bioecológicos e determinação das exigências térmicas de *Ceraecochrysa cubana* (Hagen, 1861) (Neuroptera, Chrysopidae) em laboratório**. 1991.160 p. Dissertação (Mestrado em Entomologia) - Universidade Federal de Lavras, Lavras.

SPIEGEL, M. R. **Probabilidade e Estatística**. São Paulo: McGraw-Hill do Brasil, 1978. 519 p.

VIVANCO, M. J. F. **Análise de valores extremos no tratamento estatístico da corrosão de equipamentos**. 1994. 107 p. Dissertação (Mestrado em Estatística) - Universidade Estadual de Campinas, Campinas**.**