TEMA Tend. Mat. Apl. Comput., **7**, No. 1 (2006), 91-100.

Uma Avaliação do Uso de um Modelo Contínuo na Análise de Dados Discretos de Sobrevivência

E.Y. NAKANO[1], Departamento de Estatística, Universidade de Brasília (UnB), 70910-900 Brasília, DF, Brasil

C.G. CARRASCO[2], Unidade Universitária de Ciências Exatas e Tecnológicas, Universidade Estadual de Goiás (UEG), 75000-000 Anápolis, GO, Brasil.

Resumo. Propomos neste trabalho uma comparação dos resultados de análises de dados de sobrevivência utilizando dois modelos equivalentes, sendo um contínuo e outro discreto. Diz-se "modelos equivalentes" pois será utilizado um modelo cuja formulação original é contínua e através deste modelo, será gerado um modelo discreto correspondente.

1. Introdução

A Análise de Sobrevivência ou Confiabilidade é um conjunto de técnicas e modelos estatísticos usados na analise de experimentos cuja variável resposta é o tempo até a ocorrência de um evento de interesse. Os indivíduos sob estudo podem ser animais, seres humanos, plantas, equipamentos, etc. Por outro lado, o evento de interesse pode ser: morte, remissão de uma doença, reação de um medicamento, quebra de um equipamento eletrônico, queima de uma lâmpada, etc. A principal característica dos dados de sobrevivência é a presença de censuras, que é a observação parcial da resposta. Essa informação, apesar de incompleta, é útil e importante para a análise.

Em muitos casos os dados de sobrevivência são obtidos (coletados) em sua forma discreta, devido a imprecisões nas mensurações ou simplesmente por ser discreta (quando o tempo é medido em meses, por exemplo). Em muitas aplicações não há justificativas teóricas para adotar, nestes casos, um modelo discreto para esses dados. O que se faz na prática é considerar que esses dados "poderiam" ser contínuos e realizar análise utilizando um modelo contínuo.

Neste contexto, o objetivo deste trabalho é verificar se há alguma perda na precisão das estimativas ao se utilizar um modelo contínuo em dados discretos. Para tanto, propomos neste trabalho uma comparação dos resultados de análises de dados de sobrevivência utilizando dois modelos equivalentes, sendo um contínuo e outro discreto. Diz-se "modelos equivalentes" pois será utilizado um modelo cuja

[1]nakano@unb.br
[2]clebercarrasco@hotmail.com

formulação original é contínua e através deste modelo contínuo, será gerado um modelo discreto correspondente.

Por se tratar de dados de sobrevivência, optaremos neste trabalho a utilização do modelo Exponencial. Esta escolha é justificada pelo fato da distribuição Exponencial ser uma das mais simples e importantes distribuições utilizadas na modelagem de dados que representam o tempo até a ocorrência de algum evento de interesse. A mesma tem sido utilizada intensivamente na literatura de sobrevivência e confiabilidade assim como a distribuição Normal é utilizada em outras áreas da estatística. Através do modelo Exponencial será formulado um modelo discreto, que deverá apresentar as mesmas características do modelo contínuo.

2. Desenvolvimento

2.1. O Modelo Exponencial

Uma importante distribuição de tempos de sobrevivência que assume independência do risco ao longo do tempo é dada pela distribuição Exponencial. A distribuição Exponencial é obtida tomando-se a função de risco constante ao longo do tempo. Desta forma, T é uma variável aleatória com distribuição Exponencial com parâmetro λ ($\lambda > 0$), se sua função densidade de probabilidades é escrita da forma:

$$f_c(t) = \lambda e^{-\lambda t}, \;\; t \geq 0. \tag{2.1}$$

A função de sobrevivência e de risco são dadas, respectivamente por:

$$S_c(t) = e^{-\lambda t}, \;\; t \geq 0, \tag{2.2}$$

$$h_c(t) = \lambda, \;\; t \geq 0.$$

Para o modelo Exponencial, a função de verossimilhança apresenta a seguinte forma:

$$L(\lambda|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i},$$

onde λ é o parâmetro a ser estimado, $\mathbf{t} = (t_1, ..., t_n)$ é o vetor dos valores observados, com seus respectivos indicadores de censuras dados por $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$.

Neste caso, o estimador de máxima verossimilhança do parâmetro λ é dado por

$$\widehat{\lambda}_c = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}.$$

2.2. Modelo Exponencial Discreto (Geométrica)

Todos os modelos de variáveis contínuas podem ser usados para gerar modelos discretos agrupado os tempos em intervalos unitários. A variável discreta é dada

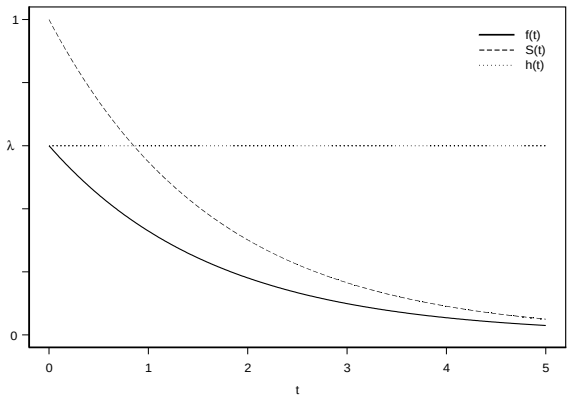

Figura 1: Funções de densidade, sobrevivência e risco da distribuição Exponencial com parâmetro λ.

por $T = [X]$, onde $[X]$ representa "a parte inteira de X" (maior inteiro menor ou igual a X). A distribuição de probabilidades de T pode ser escrita como:

$$f_d(t) = P(T = t) = P(t \leq X < t + 1), \quad t = 0, 1, 2, \ldots .$$

No caso onde X segue a distribuição dada por (2.1), temos que a função (distribuição) de probabilidades de T pode ser escrita como:

$$f_d(t) = e^{-\lambda t}\left(1 - e^{-\lambda}\right), \;\; t = 0, 1, 2, \ldots .$$

Note que T segue uma distribuição Geométrica com parâmetro $1 - e^{-\lambda}$.

A função de sobrevivência e de risco da variável aleatória T são dadas, respectivamente, por:

$$S_d(t) = P[T > t] = e^{-\lambda(t+1)}, \;\; t = 0, 1, 2, ..., \tag{2.3}$$

$$h_d(t) = 1 - e^{-\lambda}, \;\; t = 0, 1, 2, \ldots .$$

Assumindo que a contribuição para a verossimilhança do tempo censurado em t seja $S(t) = P[T > t]$ (Kalbfleisch [1], pág. 11), temos que para o modelo discreto, a função de verossimilhança apresenta a seguinte forma:

$$L(\lambda|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \left(1 - e^{-\lambda}\right)^{\delta_i} e^{-\lambda(t_i + 1 - \delta_i)},$$

onde λ é o parâmetro a ser estimado, $\mathbf{t} = (t_1, ..., t_n)$ é o vetor dos valores observados, com seus respectivos indicadores de censuras dados por $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$.

Neste caso, o estimador de máxima verossimilhança do parâmetro λ é dado por

$$\widehat{\lambda}_d = \ln\left(\frac{\sum_{i=1}^{n} t_i + n}{\sum_{i=1}^{n} t_i + n - \sum_{i=1}^{n} \delta_i}\right).$$

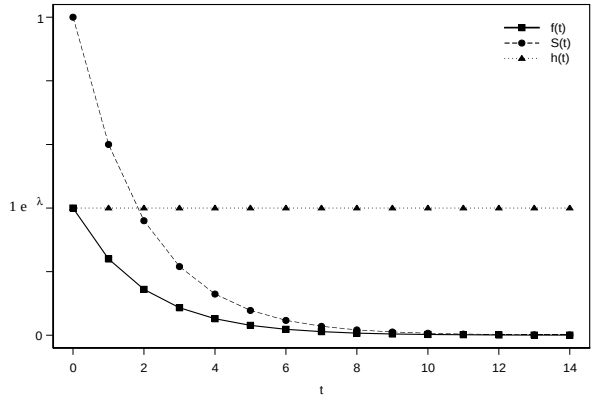

Figura 2: Funções de densidade, sobrevivência e risco da distribuição Exponencial Discreta com parâmetro λ.

3. Ilustração Numérica

Tietze [6] apresentou dados do tempo para a gravidez para casais que desejam ter uma criança. As mulheres em estudo pararam de usar qualquer tipo de contraceptivo a partir do dia de início do experimento. Neste exemplo, a variável T representa o número de meses até que a mulher tenha engravidado. Note que, neste caso, $t = 0$ indica que a mulher engravidou no primeiro mês de estudo.

Tabela 1: Dados de fertilidade de mulheres.

t (meses)	nº de mulheres expostas ao risco de engravidar no início do mês	nº de mulheres grávidas durante o mês	nº de censuras durante o mês
0	611	199	0
1	412	103	0
2	309	64	0
3	245	36	12
4	197	33	7
5	157	30	9
6	118	18	5
7	95	13	5
8	77	9	5
9	63	10	5
10	48	3	2
11	43	0	5

Fonte: Dados de Tietze [6]

Analisando os resultados da Tabela 2, pode-se notar que o modelo discreto apresentou resultados melhores que o modelo contínuo. Estes resultados podem ser verificados observando a Figura 2.

Tabela 2: Estimativa da Função de Sobrevivência para os dados da Tabela 1.

t		Modelo Discreto (2.3)	Modelo Contínuo (2.2)
(meses)	Kaplan-Meier	$\widehat{\lambda}_d = 0.31426$	$\widehat{\lambda}_c = 0.38484$
0	.6527	.7301 (.0774)	1 (.3473)
1	.4729	.5330 (.0600)	.6806 (.2076)
2	.3613	.3891 (.0279)	.4632 (.1019)
3	.2984	.2841 (.0143)	.3152 (.0168)
4	.2365	.2074 (.0291)	.2145 (.0220)
5	.1769	.1514 (.0255)	.1460 (.0309)
6	.1371	.1105 (.0265)	.0994 (.0377)
7	.1058	.0807 (.0251)	.0676 (.0382)
8	.0814	.0589 (.0225)	.0460 (.0354)
9	.0488	.0430 (.0058)	.0313 (.0175)
10	.0342	.0314 (.0028)	.0213 (.0129)
11	.0342	.0229 (.0113)	.0145 (.0197)

Neste trabalho, usaremos como critério para julgar os modelos, a noção de "distância" entre a estimativa do modelo com a estimativa empírica (Kaplan & Meier [2]). Muitos testes estatísticos de ajuste de modelos são baseados nestas distâncias (veja por exemplo, Kendall & Stuart [3] Cap. 30 ou Stephens [5]). Definiremos aqui essa distância como o erro cometido na estimação. Desta forma,

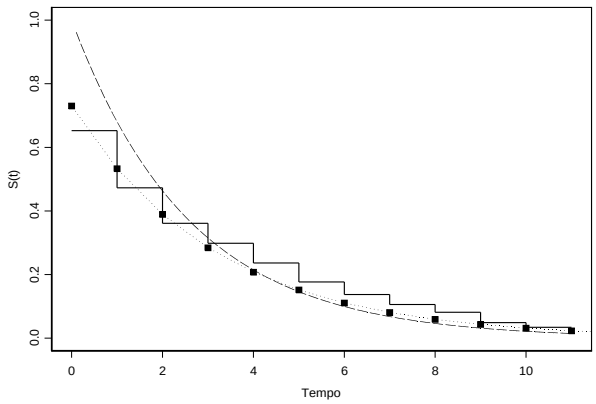

Figura 3: Funções de Sobrevivência estimadas a partir dos dados da Tabela 1. A função contínua é a estimativa dada pelo estimador de Kaplan-Meier. A função tracejada é a estimativa obtida pelo modelo contínuo, enquanto que os quadrados sólidos são as estimativas apresentadas pelo modelo discreto.

definimos o erro máximo cometido na estimação pelo modelo contínuo e discreto por

$$\varepsilon_c = \max \left| \widehat{S}_c(t) - \widehat{S}_{KM}(t) \right|, \tag{3.1}$$

$$\varepsilon_d = \max \left| \widehat{S}_d(t) - \widehat{S}_{KM}(t) \right|. \tag{3.2}$$

Neste exemplo temos que $0.0774 = \varepsilon_d < \varepsilon_c = 0.3473$. Indicando um melhor ajuste do modelo discreto para este conjunto de dados.

4. Simulações

As simulações têm como objetivo comparar a eficiência dos modelos discreto e contínuo na estimação da função de sobrevivência. Para tanto, foram gerados dados de tempos de vida discretos e estimativas foram feitas através dos dois modelos. A geração e análise dos dados foram realizados pelo *software* R (*http://r-project.org*).

Os dados discretos foram gerados a partir de uma distribuição Geométrica. O estudo de simulação foi realizado em três etapas distintas de forma verificar a influência da variabilidade dos dados, do tamanho da amostra e da quantidade de censuras. Os modelos foram avaliados através dos erros definidos em (3.1) e (3.2).

4.1. Influência da variabilidade dos dados

Foram consideradas 1000 simulações de amostras de tamanho 50 com 20% de censura. Os dados foram gerados através da distribuição Geométrica com diversos valores para o parâmetro p: 0.05, 0.1, 0.2 e 0.5.

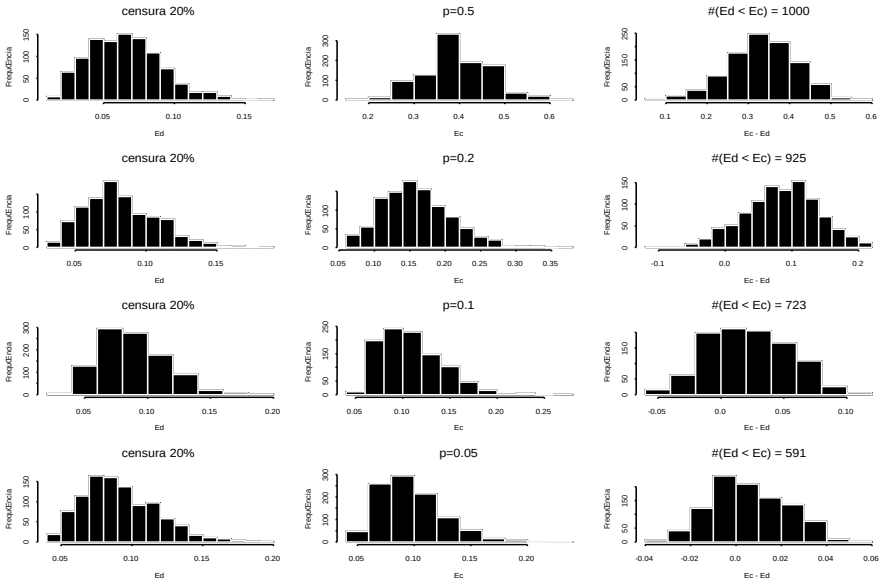

Figura 4: Erros cometidos pelos modelos de acordo com a variabilidade dos dados.

Visto que a variância da distribuição Geométrica com parâmetro p é dada por $(1-p)/p^2$, tem-se que para $p = 0.5$, 0.2, 0.1 e 0.05 a variância esperada dos dados são, respectivamente, 2, 20, 90 e 380.

Pode-se notar, através da Figura 4, que o aumento da variabilidade dos dados causa uma diminuição do erro cometido na estimação do modelo contínuo. Fato já previsível visto que um aumento da variabilidade dos dados causaria uma maior amplitude dos dados, fazendo com que ocorra um menor erro na aproximação. Ou seja, o uso de um modelo contínuo não se faz adequado quando os dados (discretos) apresentam uma baixa variabilidade.

4.2. Influência do tamanho da amostra

Neste caso considerou-se 1000 simulações de amostras geradas a partir da distribuição Geométrica com parâmetro p = 0.25 e 20% de censura. Foram utilizados diversos tamanhos de amostras: 10, 20, 50 e 100.

Observando a Figura 5 nota-se que, quanto maior o tamanho da amostra, melhor é a precisão nas estimativas tanto para o modelo discreto como para o modelo contínuo. No entanto, apesar das estimativas apresentadas para os dois modelos

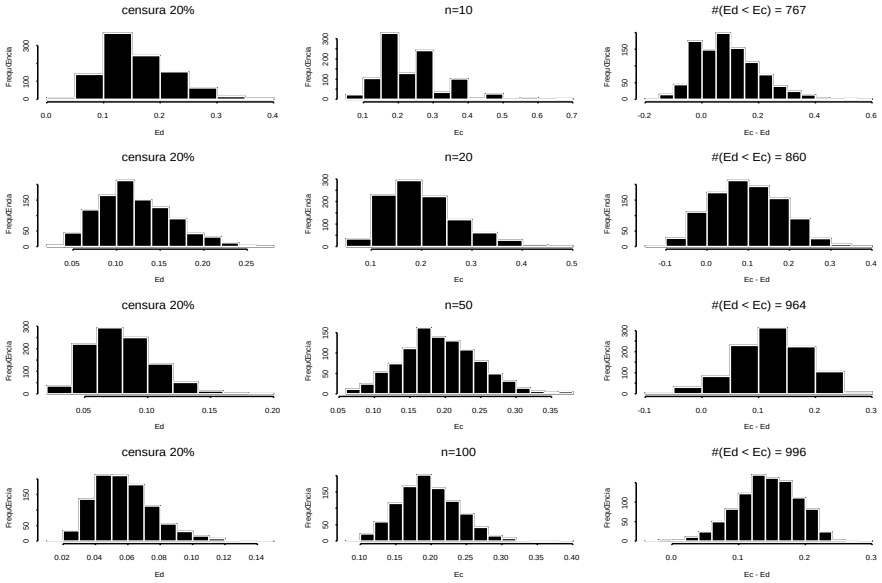

Figura 5: Erros cometidos pelos modelos de acordo com o tamanho da amostra.

melhorarem quando a amostra cresce, observamos que essa melhora é mais acentuada no modelo discreto. Isso pode ser justificado pelo fato de, ao aumentar o tamanho da amostra, os dados ficam mais representativos, ou seja, aumentam as evidências dos dados serem realmente discretos.

4.3. Influência da quantidade de censuras

Realizaram-se 1000 simulações de amostras de tamanho 50 de uma distribuição Geométrica com parâmetro p = 0.25. As amostras foram geradas com diferentes quantidades de censuras: 0%, 5%, 30% e 50%.

Os resultados apresentados pela Figura 6 sugerem que a presença de censura nos dados causa um aumento no erro cometido pelo modelo discreto e, em contrapartida, uma diminuição desse erro no caso do modelo contínuo, revelando uma maior aceitação do modelo contínuo para conjuntos de dados com grande percentual de censura. Uma explicação para isso pode ser dada pelo fato da censura comprometer a informação contida nos dados. Quanto maior a quantidade de censura, menos representativo fica o conjunto de dados. Ou seja, a o aumento da censura causa um efeito contrário do aumento do tamanho da amostra.

5. Conclusões Finais

Face aos resultados obtidos pode-se concluir que:

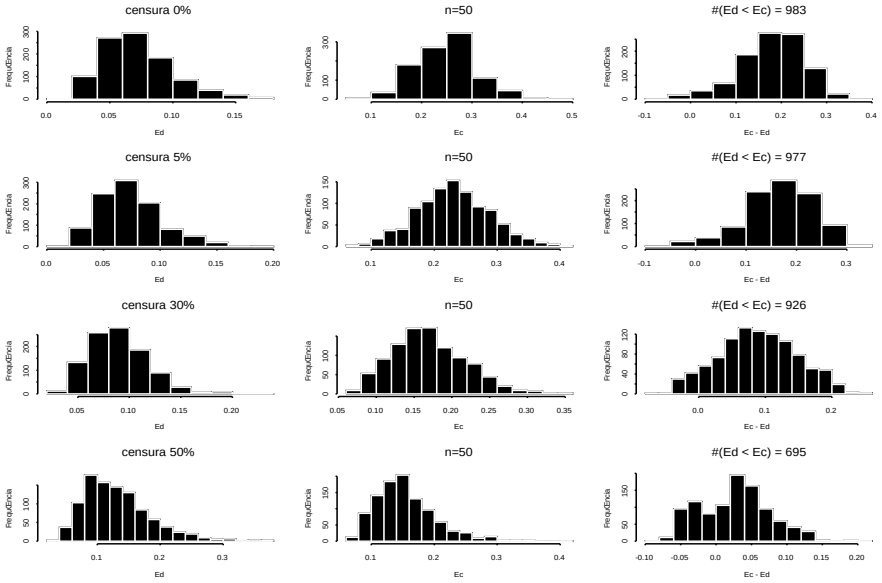

Figura 6: Erros cometidos pelos modelos de acordo com a quantidade de censuras.

• o uso de um modelo contínuo em dados discretos pode ser adequado quando a variabilidade dos dados é alta;

• como esperado, as estimativas de ambos modelos melhoram quando o tamanho da amostra cresce;

• o aumento do tamanho da amostra acentua a necessidade de se utilizar um modelo discreto para a análise dos dados;

• mesmo apresentando melhores resultados, as estimativas do modelo discreto parece não ser robusta em relação às censuras, enquanto que o desempenho do modelo contínuo parece não se alterar mesmo com uma grande quantidade de censuras. Desta forma o uso do modelo discreto se mostra mais adequado em conjunto de dados com baixa proporção de censuras.

Com base nas conclusões obtidas, verificou-se que nem sempre é aceitável a utilização de um modelo contínuo para a análise de dados discretos, pois em alguns casos pode-se observar um resultado pouco satisfatório.

Sendo assim o pesquisador deve estar atento aos seus dados e não utilizar um determinado modelo indiscriminadamente.

Neste trabalho os efeitos que poderiam influenciar o desempenho dos modelos em questão (variabilidade, tamanho da amostra e censura) foram trabalhados isoladamente. Ou seja, para estudar um determinado efeito, analises foram feitas variando o efeito de interesse e mantendo os demais fixos. Novos estudos de simulação podem ser feitos para avaliar a influência conjunta destes efeitos e de outros que poderiam também influenciar a análise dos dados.

Abstract. In this work we present a comparison of the results of analyses of survival data using two equivalent models, a continuous and another discrete one. We say equivalent models because, it will be used a model that the original formulation is continuous, and based on this continuous model we will generated a corresponding discrete model.

Referências

[1] J.D. Kalbfleisch e R.L. Prentice, "The Statistical Analysis of Failure Time Data", John Wiley & Sons, New York, 1980.

[2] E.L. Kaplan e P. Meier, Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457-481.

[3] M.G. Kendall e A. Stuart, "The Advanced Theory of Statistics", Griffin, London, 2 ed., 1968.

[4] J.F. Lawless, "Statistical Models and Methods for Lifetime Data", John Wiley & Sons, New York, 1982.

[5] M.A. Stephens, EDF statistics for goodness of fit and some comparisons, *J. Am. Stat. Assoc.*, **69** (1974), 730-737.

[6] C. Tietze, Fertility after discontinuation of intrauterine and oral contraception, *International Journal of Fertility*, **31** (1968), 385-389.