ISSN: 2317-0840

Distribuição Log-Normal e Log-Normal com fração de cura para dados de sobrevivência

Damião F. Santos[1†], Pablo L. R. Almeida[2], Tiago A. Oliveira[3]

[1] *UNB. Mestre em Estatística.*
[2] *UFLA. Mestre em Estatística. E-mail:* pablof_flourenco@hotmail.com.
[3] *UEPB. Docente do Departamento de Estatística. E-mail:* tiagoestatistico@gmail.com.

Resumo: *A análise de sobrevivência pode ser caracterizada por um conjunto de técnicas estatísticas que têm como objetivo principal a análise de tempos até a ocorrência de um determinado evento de interesse, onde as observações são acompanhadas ao longo de períodos de tempo. Tais técnicas são embasadas em modelos probabilísticos, como os modelos Log-Normal e de mistura com fração de cura. Entende-se como fração de cura, os indivíduos no estudo que são acompanhados por um período longo de tempo e observa-se que uma fração razoável deles não irá experimentar o evento de interesse, sendo denominados de curados ou imunes ao evento. Nesse estudo, foi utilizado técnicas de análise de sobrevivência aplicado a dados de pacientes portadores de Mieloma Múltiplo, no intuito de modelar o tempo de vida dos pacientes, através de técnicas não-paramétricas e paramétricas. Como critério de decisão, é utilizado análise descritiva, com o uso da diferença máxima entre o modelo ajustado e o modelo empírico estimado por Kaplan-Meier, além dos critérios de Akaike (AIC), Akaike corrigido (AICc), Informação Bayesiano (BIC) e Teste de Razão de Verossimilhança (TRV). Na análise descritiva observou-se que o modelo Log-Normal com fração de cura se ajusta melhor do que o modelo Log-Normal. No entanto, ao realizar os critérios de informação, assim como o TRV, permitiu selecionar o modelo mais parcimonioso, sendo esse o modelo Log-Normal, que se adequa melhor ao tempo de vida de pacientes portadores da doença Mieloma Múltiplo.*

Palavras-chave: Análise de Sobrevivência; Estimador Kaplan-Meier, Log-Normal; Log-Normal com Fração de Cura; Mieloma Múltiplo.

Abstract: *Survival analysis techniques applied to data from patients with Multiple Myeloma were used. Information criteria (AIC, AICc, BIC) and Likelihood Ratio Test (TRV) are used as the decision criterion. It was observed that the Log-Normal model was better suited to the data and for this reason was the selected model.*

Keywords: Survival analysis; Kaplan-Meier estimator; Log-Normal; Log-Normal with Cure Fraction; Multiple Myeloma.

[†]Autor correspondente: d.flaviostate@gmail.com.

Introdução

O Mieloma Múltiplo (MM) é a segunda doença onco-hematológica mais comum no mundo, perdendo apenas para os linfomas e chegando a representar 10% dos casos. É uma doença que se caracteriza pela proliferação descontrolada de células plasmáticas na medula óssea com frequente produção de imunoglobulinas anômalas monoclonais (Proteína M). No Brasil, o MM representa 1% de todos os tipos de câncer, sendo o segundo mais comum entre os hematológicos, ficando atrás dos linfomas Não-Hodgkin, em adultos.

Para estudar essas doenças, é comum utilizar algumas ferramentas de análise estatísticas, sendo uma das mais utilizadas, a Análise de Sobrevivência. Em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse, sendo esse tempo denominado tempo de falha (COLOSIMO; GIOLO, 2006). A principal característica do conjunto de dados de sobrevivência é a presença de censura, que ocorre quando algum sujeito ou objeto do estudo não experimentou o evento de interesse até o final do estudo. Outra situação em que o indivíduo não experimentará o evento de interesse é o fenômeno denominado fração de cura, ou seja, uma parcela ou fração da população estudada não irá falhar mesmo que o tempo tenda ao infinito.

O objetivo deste estudo foi o de modelar o tempo de vida de pacientes portadores da doença Mieloma Múltiplo utilizando o modelo Log-Normal com fração de cura e verificar se o mesmo apresenta melhoria em ajuste com relação ao modelo Log-Normal. Para a aplicação de tais técnicas, utilizou-se o *software* livre R

*Material e Métodos O banco de dados utilizado neste estudo foi obtido de Allison (2010), constituído de um total de 25 pacientes portadores da doença Mieloma Múltiplo (MM).

Uma das técnicas mais utilizadas para estimar a função de sobrevivência é o estimador de Kaplan-Meier, sendo conhecido também por limite-produto. Kaplan e Meier (1958) mostraram que $\hat{S}(t)$ é um estimador de máxima verossimilhança não-paramétrico de $S(t)$ e desde então, este estimador vem sendo amplamente utilizado em estudos clínicos e de confiabilidade. Desta forma, suponha que existam n indivíduos no estudo e tem-se $k(\leq n)$ falhas distintas nos tempos $t_1 < t_2 < \ldots < t_k$. O estimador de Kaplan-Meier é expresso por:

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

em que, n_j é o número de indivíduos sob risco em t_j, ou seja, aqueles que não falharam e nem foram censurados até o instante imediatamente anterior a t_j, e d_j é o número de falhas em t_j, $j = 1, \ldots, k$.

De acordo com Colosimo e Giolo (2006), assim como a distribuição Weibull, a distribuição Log-Normal é muito utilizada para caracterizar tempos de vida de produtos e indivíduos, além de ser bastante utilizada para descrever situações clínicas, como o tempo de vida de pacientes com leucemia. Sendo assim, a função de densidade da distribuição Log-Normal é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right\}, t > 0,$$

em que $\mu \in \Re$, é a média do logaritmo do tempo de falha, assim como $\sigma > 0$ é o desvio-padrão.

As funções de sobrevivência e de risco da referida distribuição são dadas, respectivamente, por:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right),$$

e

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

em que $\Phi(.)$ é a função de distribuição acumulada de uma normal padrão. A característica da função taxa de falha do modelo Log-Normal é que inicialmente tem forma crescente e depois decrescente.

De acordo com Santos (2017), definido como fração de cura, esse fenômeno ocorre quando, mesmo após um longo período de acompanhamento, o indivíduo não falhará. Considere, por exemplo, que o evento de interesse em um estudo clínico seja a morte de pacientes diagnosticados com câncer e a variável resposta seja o tempo de sobrevivência após o tratamento do indivíduo. Nesses estudos é comum observar indivíduos que se curam da doença após o tratamento e, sendo assim, não experimentará o evento de interesse pelas razões estudadas.

Um modelo com fração de cura foi proposto por Berkson e Gage (1952). Ainda de acordo com Santos (2017), esse modelo foi definido como um modelo de mistura em que há uma proporção de indivíduos curados ou imunes na população e uma proporção de indivíduos suscetíveis. Para tanto, a população em estudo é dividida em duas subpopulações, de tal forma que uma seja formada pelos indivíduos que estão suscetíveis à falha (S) e a outra, pelos indivíduos não suscetíveis à falha (NS) ou curados. Dessa forma, é considerada a variável aleatória C_i com distribuição Bernoulli associada a cada indivíduo i para indicar se o i-ésimo indivíduo é suscetível ou não suscetível, isto é,

$$C_i = \begin{cases} 1, & \text{se o i-ésimo indivíduo é suscetível} \\ 0, & \text{se o i-ésimo indivíduo é não suscetível.} \end{cases}$$

Considere agora que $P(C_i = 0) = 1 - \phi$ é a probabilidade da i-ésima observação não ser suscetível ou ser curado, e $P(C_i = 1) = \phi$, a probabilidade da i-ésima observação ser suscetível. Ao considerar que $P(\text{NS}) = 1 - \phi$ com função de sobrevivência $S_{NS}(t)$ e $P(\text{S}) = \phi$ com função de sobrevivência $S_S(t)$. Assim, a função de sobrevivência de forma mista é definida da seguinte forma:

$$\begin{aligned} S_{FC}(t) &= P(NS)P(T > t|NS) + P(S)P(T > t|S) \\ &= (1 - \phi)S_{NS}(t) + \phi S_S(t) \\ &= (1 - \phi) + \phi S_S(t) \end{aligned} \quad .$$

Desta maneira, a função de densidade e de sobrevivência para a distribuição Log-Normal com fração de cura é definida por:

$$f_{FC}(t) = \phi\left[\frac{1}{\sqrt{2\pi}t\sigma}\exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right\}\right], t > 0, \mu \in \Re; \sigma > 0,$$

e

$$S_{FC}(t) = (1-\phi) + \phi\left[\Phi\left(\frac{-\log(t)+\mu}{\sigma}\right)\right], t > 0, \mu \in \Re; \sigma > 0.$$

Destaca-se que sendo $P(NS) = 1 - \phi = 0$ ou $\phi = 1$, tem-se que $S_{FC}(t) = S_S(t)$ e $\lim_{t\to\infty} S_{FC}(t) = 1-\phi$, que é a proporção de indivíduos não suscetíveis ao evento de interesse e $\phi \in [0, 1]$.

Para estimação dos parâmetros do modelo Log-Normal e modelo Log-Normal com fração de cura, será utilizado o método da máxima verossimilhança. O logaritmo da função de verossimilhança do modelo Log-Normal com fração de cura é dado por:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i \log(\phi) + \sum_{i=1}^{n} \delta_i \log\left\{\frac{1}{\sqrt{2\pi}t\sigma}\exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}\right\}$$
$$+ \sum_{i=1}^{n}(1-\delta_i)\log\left\{(1-\phi) + \phi\left[\Phi\left(\frac{-\log(t)+\mu}{\sigma}\right)\right]\right\} + C,$$

sendo, $\boldsymbol{\theta} = (\phi, \mu, \sigma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$.

Após a estimação dos parâmetros do modelo, de acordo com Nakano e Carrasco (2006), ao utilizar o erro máximo cometido na estimação, pode-se definir como um teste estatístico de ajuste de modelos, utilizando-se as diferenças entre as estimativas do modelo estudado e as estimativas pelo método de Kaplan-Meier, sendo expresso por:

$$\epsilon = \max|\hat{S}(t) - \hat{S}_{km}(t)|$$

Por meio dos critérios de Akaike (AIC), Akaike corrigido (AICc) e de Informação Bayesiano (BIC) é possível verificar qual modelo é mais adequado aos dados em estudo. A ideia básica dos critérios de informação é selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros. Além disso, utiliza-se o Teste de Razão de Verossimilhança (TRV), que as hipóteses associadas ao teste são dadas por:

$$TRV = \begin{cases} H_0, & \text{Os dados seguem uma distribuição LN } (\phi = 1) \\ H_1, & \text{Os dados seguem uma distribuição LNFC } (\phi \neq 1) \end{cases}$$

Resultados e discussão

Inicialmente é realizada a análise do tempo de sobrevivência dos indivíduos pertencentes ao estudo e, por meio da Figura 1, que apresenta a função de sobrevivência estimada pelo método de Kaplan e Meier (1958), percebe-se que a probabilidade de sobrevivência diminui com o passar do tempo e também é observado que após o tempo $t = 1.296$ a probabilidade de sobrevivência se estabilizou, sendo assim, há indícios da presença de fração de curados dentre os indivíduos em estudo.

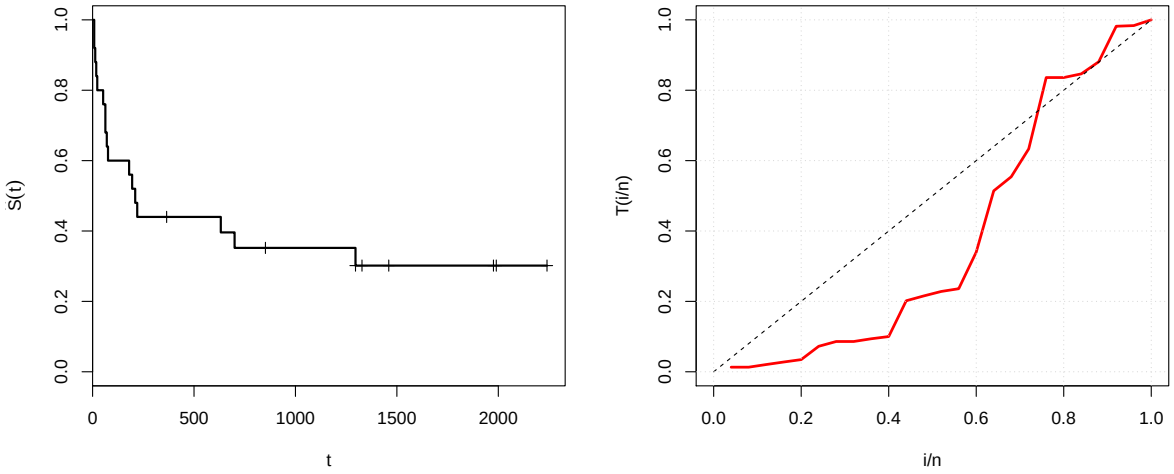

Figura 1: Função de sobrevivência estimada por Kaplan-Meier e de risco acumulado.

Além disso, a partir da Figura 1, supõe-se um modelo que contenha a função de risco incialmente crescente e depois apresenta uma característica decrescente. Conforme descrito anteriormente, a distribuição Log-Normal apresenta essas características e portanto, será testado os modelos Log-Normal (LN) e também Log-Normal com fração de cura (LNFC).

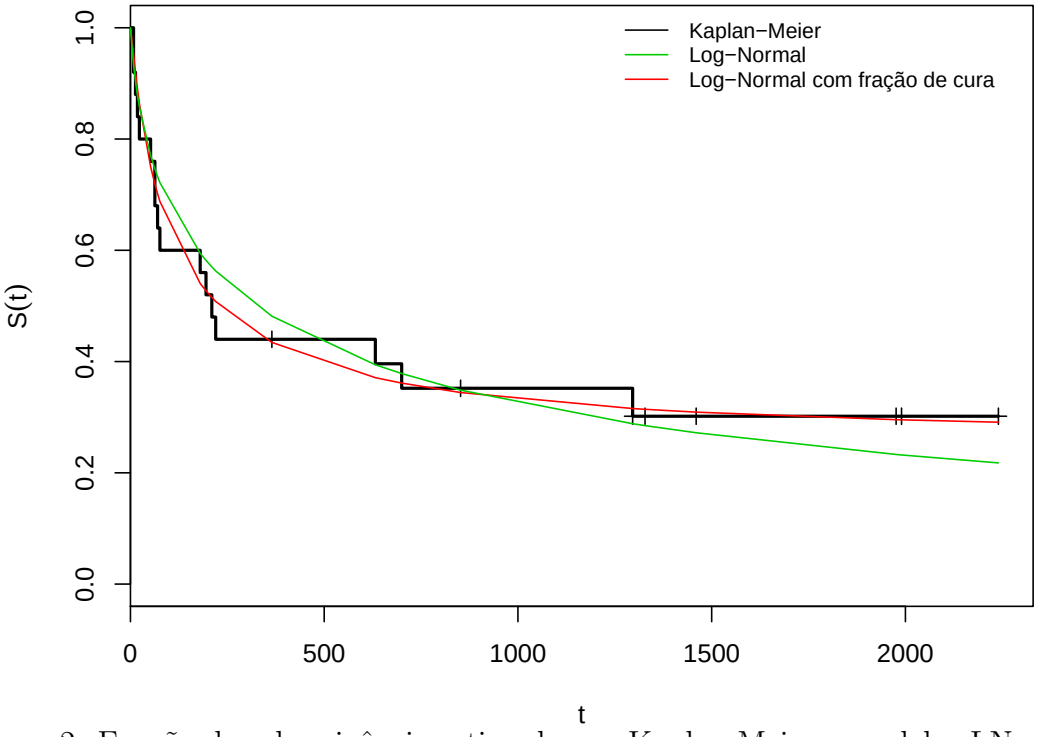

Figura 2: Função de sobrevivência estimada por Kaplan-Meier e modelos LN e LNFC.

Pode-se observar por meio da Figura 2, que o ajuste do modelo Log-Normal com fração de cura, aparentemente está melhor do que o modelo Log-Normal, devido a aproximação com a distribuição de sobrevivência empírica de Kaplan-Meier. A Tabela 1 apresenta as estimativas dos parâmetros dos modelos LN e LNFC, com seus respectivos erros-padrão. Nota-se que os os erros-padrão das estimativas dos parâmetros de ambos os modelos são de pequena magnitude, que indica que os parâmetros foram bem estimados e significativos, de acordo com o valor-p$< 0,05$.

Tabela 1: Estimativas dos modelos Log-Normal e Log-Normal com Fração de Cura.

Modelo	Parâmetro	Estimativa	Erro-Padrão	valor-p
Log-Normal	μ	5,7849	0,5316	$< 2 \times 10^{-16}$
	σ	2,4745	0,4597	$7,3313 \times 10^{-08}$
Log-Normal com fração de cura	μ	4,6512	0,5140	$< 2 \times 10^{-16}$
	σ	1,6923	0,3975	$2,0663 \times 10^{-05}$
	ϕ	0,7348	0,1135	$9,6358 \times 10^{-11}$

Tabela 2: Estimativas da função de sobrevivência pelo método de Kaplan-Meier, pelo modelo Log-Normal e Log-Normal com fração de cura.

t	Kaplan-Meier	LN	LNFC
8	0,920	0,933	0,953
13	0,880	0,903	0,920
18	0,840	0,879	0,890
23	0,800	0,858	0,864
52	0,760	0,771	0,750
63	0,680	0,746	0,719
70	0,640	0,733	0,702
76	0,600	0,722	0,688
180	0,560	0,595	0,540
195	0,520	0,582	0,527
210	0,480	0,570	0,515
220	0,440	0,563	0,508
365	0,440	0,481	0,434
632	0,396	0,394	0,371
700	0,352	0,378	0,361
852	0,352	0,349	0,344
1.296	0,302	0,288	0,316
1.328	0,302	0,285	0,314
1.460	0,302	0,272	0,309
1.976	0,302	0,233	0,296
1.990	0,302	0,232	0,295
2.240	0,302	0,218	0,291
	ϵ	0,123	**0,088**

De acordo com a Tabela 2 percebe-se que as estimativas dos dois modelos estão próximas das estimativas da função empírica pelo método de Kaplan-Meier. No entanto,

percebe-se que nos últimos tempos o modelo LNFC se aproxima mais do que o LN. Assim, ao utilizar o valor de $\epsilon = 0,088$ do modelo Log-Normal com fração de cura, tem-se mais indícios que o modelo mais adequado é o Log-Normal.

Mesmo nas análises descritivas anteriormente mostrando que o modelo Log-Normal com fração de cura apresentou melhor ajuste aos dados, através dos critérios de informação observou-se que os dois modelos são bem próximos em termos de ajustes, pois os valores do AIC para o modelo LNFC é menor do que para o modelo LN, no entanto, para o AICc e BIC a situação inverte.

Tabela 3: AIC, AICc, BIC e TRV para os modelos Log-Normal e Log-Normal com fração de cura.

Modelos	AIC	AICc	BIC	TRV	valor-p
Log-Normal	247,4740	**248,0195**	**249,9118**	2,2038	0,1377
Log-Normal com fração de cura	**247,2702**	248,4131	250,9268		

Além disso, ao realizar o TRV ao nível de significância de $\alpha = 0,05$, não rejeita-se a hipótese H_0, ou seja, os dados não seguem uma distribuição Log-Normal com fração de cura, e sim, o melhor ajuste é com a distribuição Log-Normal.

Conclusões

Observou-se a partir das técnicas de análise de sobrevivência e dos testes, que o modelo Log-Normal foi o modelo que mais se adequou aos dados dos tempos de sobrevivência pacientes portadores da doença Mieloma Múltiplo (MM). Apesar de ter sido identificado por meio de análise descritiva com ajuste gráfico e pelo erro máximo das estimativas dos modelos LN e LNFC em relação as estimativas empíricas, que o melhor modelo para o ajuste dos dados seria o LNFC, ao utilizar o Teste de Razão de Verossimilhança (TRV) e os critérios de informação AIC, AICc e BIC, observou-se que o modelo mais adequado aos dados é de fato o Log-Normal, utilizando-se o nível de significância $\alpha = 0,05$.

O Teste de Razão de Verossimilhança permitiu selecionar o modelo mais parcimonioso, em que mesmo com o acréscimo de um parâmetro e o ajuste sendo descritivamente melhor, não significa que o mesmo é mais eficiente, e com isso, é evitado conclusões equivocadas.

Referências Bibliográficas

ALLISSON, P.D. *Survival Analysis Using SAS: A Pratical Guide, Second Edition.* 2. ed. SAS Institute Inc., Cary, NC, USA, 2010.

BERKSON, J.; GAGE, R.P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v.47, n.259, p.501-515, 1952.

COLOSIMO, E.A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada..* São Paulo: Edgard Blücher, 2006.

SANTOS, D.F. *Modelo de Regressão Log-Logístico discreto com fração de cura para dados de sobrevivência*. Dissertação (Mestrado) - Universidade de Brasília, 2017.

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

NAKANO, E.Y.; CARRASCO, C.G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends in Applied and Computational Mathematics*, v.7, n.1, p.91-100, 2006.

R CORE TEAM. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, URL http://www.R-project.org/.