DISTRIBUIÇÃO BETA LOG-NORMAL: UMA ALTERNATIVA PARA ANÁLISE DE TEMPO DE VIDA DE DADOS DE SOBREVIVÊNCIA

Alexandre Henrique Martins[1,2], **Lourdes Coral Contreras Montenegro**[1,2]

Resumo: *Este trabalho visa apresentar a distribuição beta log-normal (BLN), bem como suas funções de densidade, acumulada e de sobrevivência. Além disso, buscamos adaptar essa função em dados de sobrevivência e compará-la com as distribuições clássicas utilizadas neste tipo de experimento como a exponencial, Weibull e log-normal. Para avaliação do modelo proposto, foram ajustadas as funções de sobrevivência dos modelos clássicos por meio do pacote do software R survival() bem como gráficos para avaliação de ajuste. Assim, comparamos com os resultados da BLN e verificamos que o ajuste dessa função é mais adequado que as distribuições anteriores quando ajustamos os testes de razão de verossimilhança clássico e generalizado. Portanto, verificamos que a beta log-normal função se adequa aos dados trabalhados e que ela pode ser utilizada em análise sobrevivência.*

Palavras-chave: *distribuição beta log-normal, análise de sobrevivência aplicada.*

1 Introdução

A distribuição beta log-normal (BLN), em termos gerais, é a união das funções beta e log-normal. Com essa junção, espera-se fazer uso em diversas áreas de pesquisa tanto em áreas específica de segurança em engenharia e em outros campos que necessitem da utilização de modelos probabilísticos para análise de confiabilidade (Castellares et al. 2009). Em análise de sobrevivência, há a utilização de modelos probabilísticos para a descrição do tempo de vida tanto de produtos industriais quanto em análise clínica de pacientes que apresentem alguma doença. Segundo Colosimo et al. (2006), para esse tipo de pesquisa, devido a melhor adequação em várias situações práticas, as distribuições frequentemente utilizadas nessa metodologia são: exponencial, Weibull e log-normal. Assim, pretendemos apresentar neste trabalho a inserção da distribuição BLN em análise de dados de sobrevivência. Para isso, adaptaremos esse modelo a problemas proposto por Colosimo et al. (2006) e realizaremos comparações dos modelos informados anteriormente por meio do Teste de Razão de Verossimilhança e por métodos gráficos.

2 Objetivo

Visamos neste trabalho apresentar a distribuição BLN no tratamento de dados de sobrevivência e verificar sua aplicabilidade neste tipo de experimento.

[1]Universidade Federal de Minas Gerais, alexandrehm@gmail.com, lourdes@gmail.com
[2]Agradecemos a FAPEMIG pelo apoio e auxílio nesta pesquisa

3 Metodologia

A distribuição beta log-normal

Segundo Castellares et al. (2009), a nova distribuição com quatro parâmetros (a, b, μ e σ^2), dita beta log-normal (BLN) foi introduzida com a expectativa de aplicação em teste de análise de sobrevivência na área da engenharia e em outros ramos de pesquisa.

Esse novo modelo é dado pela seguinte função de distribuição generalizada beta

$$F(x) = \frac{1}{B(a,b)} \int_0^{G(x)} w^{a-1}(1-w)^{b-1} dw = I_{G(x)}(a,b) \tag{1}$$

onde $a > 0$ e $b > 0$ além de serem dois parâmetros cuja função é introduzir assimetria e variar o peso da cauda representada pelo gráfico da distribuição. A função de densidade da BLN com quatro parâmetros (a, b, μ e σ^2) é definida por

$$f(x) = \frac{exp\left\{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2\right\}}{x\sigma\sqrt{2\pi}B(a,b)} \Phi\left(\frac{\log x-\mu}{\sigma}\right)^{a-1} \left\{1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right\}^{b-1} \tag{2}$$

A função de distribuição acumulada e a função de risco correspondente da BLN são expressas, respectivamente

$$F(x) = I_{\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]}(a,b) \tag{3}$$

e

$$h(x) = \frac{exp\left\{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2\right\} \Phi\left(\frac{\log x-\mu}{\sigma}\right)^{a-1} \left\{1-\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right\}^{b-1}}{x\sigma\sqrt{2\pi}B(a,b)\left[1-I_{\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]}(a,b)\right]} \tag{4}$$

A função de sobrevivência da BLN é dada por:

$$S(x) = 1 - F(x) = 1 - I_{\left[\Phi\left(\frac{\log x-\mu}{\sigma}\right)\right]}(a,b) \tag{5}$$

Teste da Razão de Verossimilhança

Para esse teste são definidas as seguintes hipóteses: H_0: O modelo de interesse é adequado vs. H_1: o modelo de interesse não é adequado

A estatística de teste para o TRV é dada por:

$$TRV = -2\log\left[\frac{L(\hat{\theta}_M)}{L(\hat{\theta}_G)} = 2\log\left[\log L(\hat{\theta}_G) - \log L(\hat{\theta}_M)\right]\right] \tag{6}$$

onde, $\log L(\hat{\theta}_G)$ é o logaritmo da função de verossimilhança do modelo generalizado e $\log L(\hat{\theta}_M)$ é o logaritmo da função de verossimilhança do modelo de interesse. Temos que, sob H_0, a estatística de teste tem, aproximadamente, uma distribuição qui-quadrado com graus de liberdade igual à diferença do número de parâmetros de cada um dos modelos que são comparados.

Teste da Razão de Verossimilhança Generalizado (TRVg)

As hipóteses são definidas da seguinte maneira: H_0: O modelo G_γ é mais adequado para o conjunto de dados do que F_θ vs. H_1: O modelo F_θ é mais adequado para o conjunto de dados do que G_γ

Para tal teste usa-se a seguinte estatística:

$$TLR, NN = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{f(y_i|x_i, \hat{\theta})}{g(y_i|x_i, \hat{\gamma})} \right\} \times$$
$$\times \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(y_i|x_i, \hat{\theta})}{g(y_i|x_i, \hat{\gamma})} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i|x_i, \hat{\theta})}{g(y_i|x_i, \hat{\gamma})} \right)^2 \right\}^{-1} \quad (7)$$

onde, $f(y_i|x_i, \hat{\theta})$ e $g(y_i|x_i, \hat{\gamma})$ são as funções densidade de F_θ e G_γ, respectivamente e n é o tamanho da amostra dos dados que serão analisados. Além disso, sob H_0, essa estatística de teste tem distribuição aproximadamente normal padrão.

4 Resultados e Discussões

Utilizaremos como base para nossas análises o problema 4, do capítulo 3, do livro "Análise de sobrevivência aplicada" e cuja referência é Colosimo et al. (2006). O exercício aborda a aplicação das técnicas paramétricas de análise de sobrevivência para estimar o tempo médio e mediano de vida de um tipo de isolador elétrico funcionando a uma temperatura de 200°C. Para isso, foram coletados 60 isoladores, dos quais 45 haviam falhado e 15 ainda estavam funcionando após o tempo de 2729 horas (censura). No entanto, abordaremos nesta pesquisa somente a função de distribuição que melhor se adequa aos dados. As funções de densidade de probabilidade utilizadas serão a exponencial, Weibull e log-normal:

$$\text{exponencial: } f(t) = \frac{1}{\alpha} exp\left\{ -\frac{1}{\alpha} \right\}, t \geq 0 \quad (8)$$

$$\text{weibull: } f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} exp\left\{ -\frac{t}{\alpha}^\gamma \right\}, t \geq 0 \quad (9)$$

$$\text{log-normal: } f(t) = \frac{1}{\sqrt{2\pi} t \sigma} exp\left\{ -\frac{1}{2} \left(\frac{\log(t) - \mu}{\sigma} \right)^2 \right\}, t > 0 \quad (10)$$

O software que nos serviu de plataforma foi o R, pois o mesmo obtém o pacote *survival()*, base para este tipo de análise.

A função de sobrevivência é gerada pela seguinte relação:

$$S(x) = 1 - F(x) \text{ onde } F(t) \text{ é a função acumulada de probabilidade} \quad (11)$$

Assim, com base nos dados do problema, foram gerados os modelos de sobrevivência para cada uma das distribuições, conforme a seguir:

exponencial: $\ddot{S}(t)_e = exp\{-t/2017,756\}$, onde $\alpha = 2017,756$

weibull: $\ddot{S}(t)_w = exp\{-(t/1993,215)^{1,28131}\}$, onde $\alpha = 1993,215$ e $\gamma = 1,28131$

log-normal: $\ddot{S}(t)_{ln} = exp\{-(\log(t) - 7,224766)/0,9505452\}$, onde $\mu = 7,224766$ e $\sigma = 0,9505452$

beta log-normal: $1 - I_{\left[\Phi\left(\frac{\log t - 8,783272}{0,9227727}\right)\right]}(0,5600049; 14,56351)$, onde $\mu = 8,783272, \sigma = 0,9227727$, $a = 0,5600049$ e $b = 14,56351$

De posse dessas informações, geramos os seguintes gráficos de sobrevivências estimadas em comparação ao modelo não paramétrico de Kaplan-Meier:

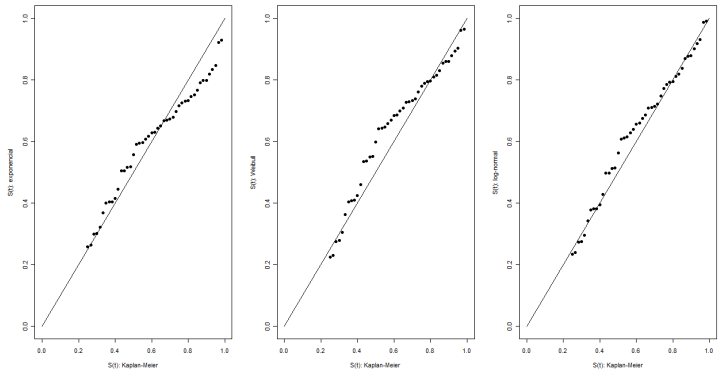

Figura 1: gráficos das sobrevivências estimadas por Kaplan-Meier versus as sobrevivências estimadas pelos modelos exponencial, Weibull e log-normal.

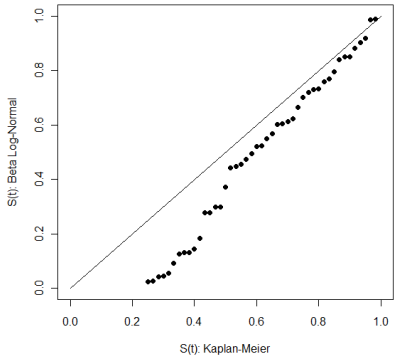

Figura 2: gráfico de sobrevivência estimada por Kaplan-Meier versus BLN.

Pelas figuras 1 e 2, notamos que a distribuição que apresenta o desvio menos significativo em relação ao método de Kaplan-Meier é a distribuição Log-Normal. Contudo, necessitamos de outro padrão de análise via testes de hipóteses para ratificar a conclusão gráfica e verificar se a BLN realmente não se adapta melhor aos dados trabalhados.

Apesar de não existir uma função generalizada que abranja as quatro funções estudadas até então, usaremos a BLN como função de comparação em dois momentos: (1º) utilizaremos o TRVg para verificar se a BLN é mais adequada que os modelos de Weibull e exponencial e (2º) como a log-normal é um caso particular da beta

Os resultados seguem na tabela abaixo:

Tabela 1: Resultado dos testes TRV e TRV generalizado

Comparação	Teste utilizado	Valor do teste	P-valor
BLN×exponencial	TRV generalizado	$3,99727$	$< 0,0001$
BLN×Weibull	TRV generalizado	$3,857019$	$< 0,0001$
BLN×log-normal	TRV	$62,1694$	$< 0,0001$

Pelos dados da tabela 1, podemos verificar que, em comparação aos modelos propostos, a inferência realizada pelos TRV e TRVg demonstrou que há evidências que a BLN é um modelo mais adequado para explicar o tempo de vida dos isoladores elétricos.

Tabela 2: Critérios de seleção

Distribuição	AIC	BIC	HQ
BLN	711, 1812	719, 5586	708, 8196
exponencial	776, 8767	778, 971	776, 2863
Weibull	775, 394	779, 5827	774, 2132
log-normal	769, 3506	773, 5393	768, 1698

5 Conclusão

Com base nas informações tratadas acima, concluímos que, apesar de graficamente a BLN não apresentar um ajuste adequado quando comparada à distribuição não paramétrica Kaplan-Meier, os testes de hipóteses elaborados confirmaram que o modelo beta log-normal apresentou melhor ajuste aos dados do que as funções Weibull, exponencial e log-normal. Além disso, verificamos pelos critérios de seleção de modelos AIC, BIC e HQ (Hannan - Quinn) que o modelo BLN foi o que melhor ajustou o tempo de funcionamento dos isoladores elétricos.

Referências

[1] Bolfarine, H., Sandoval, M. C., Introdução à inferência estatística, 1ª edição, Rio de Janeiro: SBM, 2001.

[2] Castellares, F., Montenegro, L. C., Cordeiro, G. M., The Beta Log-normal Distribution, Belo Horizonte, 2009.

[3] Colossimo, E. A., Giolo, S. R., Análise de sobrevivência aplicada, São Paulo: E, Blücher: 2006.

[4] Johnson, N. L., Kotz, S., Balakrishnan, N., Continous Univariate Distributions, 2ª edição, Vol 1, Wiley Séries in Probability and Mathematical Statistics, Nova York: John Wiley & Sans, 1994.

[5] Ross, S., A first course of probability, 5ª edição, New Jersey: Pratice Hall, 1998.

[6] Vuong, Q. H., Likelihood Ratio Testes for Model Selection and Non-Nested Hypotheses, *Econometrica*, **57(2)**, 307-333, 1989.