

Intervalos de Confiança para o quantil de probabilidade p , usando a metodologia PBSRLLR

João Renato Caramona Belo Sebastião

Escola Superior de Gestão do Instituto Politécnico de Castelo Branco, Largo do Município, Palácio das Palmeiras, 6060-163 Idanha-a-Nova. (jrenato@esg.ipcb.pt)

Resumo: A distribuição dos estimadores de máxima verosimilhança, em contextos de censura pelo tempo, depende de parâmetros perturbadores que impossibilitam a determinação de intervalos de confiança exactos para os parâmetros populacionais. Nestas situações, são usados resultados assintóticos, que permitem obter intervalos de confiança aproximados, com uma probabilidade de cobertura próxima da confiança do intervalo. Neste trabalho, utiliza-se a metodologia PBSRLLR⁽¹⁾ na determinação de intervalos de confiança aproximados para parâmetros populacionais, de distribuições pertencentes à família de Log-Localização e Escala, em especial Weibull e Lognormal. Pretende-se estudar a evolução da probabilidade de cobertura dos intervalos, para o quantil de probabilidade p , de acordo com vários factores:

- a dimensão da amostra (em especial amostras de pequena dimensão);
- a proporção de falhas na amostra;
- o valor de p ;
- a metodologia usada.

Foram implementadas várias funções no programa estatístico "R-PROJECT"⁽²⁾, que permitem levar a efeito o processo de inferência necessário à determinação dos intervalos de confiança. Como complemento e com a finalidade de melhorar a velocidade dos métodos, as funções que comportam um elevado número de cálculo foram escritas em "Linguagem C".

Palavras-chave: Estimação paramétrica; Bootstrap; Razão de Verosimilhanças.

Abstract: When time censored samples are used, the distribution of maximum likelihood estimator is dependent of a nuisance parameter that make impossible to calculate exact confidence intervals for the parameter of interest. In this paper we use asymptotic results to find confidence intervals for the parameter quantil- p of a Log-Location-Scale distribution. In particular, we use the method PBSRLLR to calculate these intervals and the main interest is to study the evolution of the cover probability for such intervals under some factors like sample size, proportion of sample failures, the value of p and the method used. We use the package "R-PROJECT" to implement the method PBSRLLR with same function writing in "C-Language" to decrease de time used by the calculations.

Keywords: Parametric Estimation; Bootstrap; Likelihood Ratio

⁽¹⁾Parametric Bootstrap Signed Square Root Likelihood Ratio

⁽²⁾<http://www.r-project.org>

1 Introdução

O desenvolvimento informático dos últimos anos veio permitir o aperfeiçoamento e utilização de um conjunto de técnicas que até então eram de difícil implementação devido à complexidade e volume de cálculo que comportam.

Neste trabalho é apresentado o método PBSRLLR que implementamos no programa "R-PROJECT", de modo, a determinar intervalos de confiança para o quantil de probabilidade- p de uma distribuição pertencente à família de Log-Localização e Escala.

2 Considerações iniciais

Suponhamos que o conjunto de dados disponível pode ser representado por $(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)$, em que

$$\delta_i = \begin{cases} 1 & , \text{se } t_i \text{ é observação não censurada} \\ 0 & , \text{se } t_i \text{ é observação censurada} \end{cases}$$

e os valores t_i são provenientes de uma distribuição pertencente à família de Log-Localização e Escala.

Considerando T uma variável aleatória com distribuição pertencente à família de Log-Localização e Escala, as funções de Distribuição e Densidade de Probabilidade podem ser escritas na seguinte forma, repectivamente:

$$F(t; \mu, \sigma) = \Phi \left(\frac{\log(t) - \mu}{\sigma} \right), \quad t > 0, \mu \in \mathbb{R} \text{ e } \sigma > 0 \quad (1)$$

$$f(t; \mu, \sigma) = \frac{1}{\sigma t} \phi \left(\frac{\log(t) - \mu}{\sigma} \right), \quad t > 0, \mu \in \mathbb{R} \text{ e } \sigma > 0, \quad (2)$$

onde, Φ e ϕ são funções que não dependem dos parâmetros de localização μ e escala σ . O quantil de probabilidade p é definido por:

$$t_p = \exp(\mu + \Phi^{-1}(p)\sigma). \quad (3)$$

A função de verosimilhança pode ser escrita como se apresenta a seguir:

$$\begin{aligned} L(\mu, \sigma) &= C \prod_{i=1}^n L_i \\ &= C \prod_{i=1}^n [f(t_i; \mu, \sigma)]^{\delta_i} [1 - F(t_i; \mu, \sigma)]^{1-\delta_i} \end{aligned} \quad (4)$$

onde C é uma constante que depende dos mecanismos de amostragem e de censura, sendo portanto difícil de generalizar. No entanto, supomos que não depende de parâmetros desconhecidos, podendo assim ser tomada igual a um.

Um intervalo $C_{n;1-\alpha}$ de confiança a $100(1-\alpha)\%$ para t_p diz-se de "ordem k " se $P(C_{n;1-\alpha} \supset \{t_p\}) = 1 - \alpha + O(n^{-\frac{k}{2}})$. O termo $1 - \alpha$ designa-se a probabilidade nominal do referido intervalo e $1 - \alpha + O(n^{-\frac{k}{2}})$ a sua probabilidade de cobertura. No caso de não existir o termo $O(\cdot)$, o intervalo diz-se exacto. Se a probabilidade de cobertura for maior que a probabilidade nominal, o método de determinação do intervalo diz-se conservativo. Se a probabilidade de cobertura for menor que a probabilidade nominal, o método diz-se anti-conservativo.

Os intervalos de confiança exactos são difíceis de determinar. Para censura pelo tempo não existem, em geral, procedimentos exactos de estimação. A função de verosimilhança contém uma componente discreta, relacionada com o número aleatório de falhas que dificulta a determinação de estatísticas suficientes necessárias para o procedimento de estimação exacta.

Relativamente a censura pela falha ou amostras completas, Lawless (1982) descreve métodos de obtenção de estatísticas suficientes para a família de Localização e Escala que permitem a determinação de intervalos de confiança exactos, no sentido em que $P(C_{n;1-\alpha} \supset \{t_p\}) = 1 - \alpha$.

3 Descrição dos métodos

Um método bastante usual na determinação de intervalos de confiança para um parâmetro $\theta \in \mathbb{R}$, utiliza a distribuição assintótica de

$$-2 \log \left(\frac{L(\theta)}{L(\hat{\theta})} \right), \quad (5)$$

onde $\hat{\theta}$ é o estimador de máxima verosimilhança de θ e $L(\theta)$ a respectiva função de verosimilhança. De acordo com Venzon e Moolgavkar (1988), a distribuição assintótica de (5) é um qui-quadrado com um grau de liberdade.

Como neste artigo o parâmetro de interesse é t_p , que depende dos parâmetros μ e σ , o quociente $\frac{L(\theta)}{L(\hat{\theta})}$ na equação (5) é redefinido por

$$\frac{L(\log(t_p) - \Phi^{-1}(p)\sigma, \sigma)}{L(\hat{\mu}, \hat{\sigma})}, \quad (6)$$

sendo σ considerado um parâmetro perturbador, $\hat{\mu}$ e $\hat{\sigma}$ são estimativas de máxima verosimilhança de μ e σ respectivamente. Nestas situações, maximiza-se a expressão anterior relativamente a σ para tornar o quociente dependente apenas de um parâmetro, conduzindo assim à definição da função "Profile Likelihood" para t_p

$$R(t_p) = \max_{\sigma} \left[\frac{L(\log(t_p) - \Phi^{-1}(p)\sigma, \sigma)}{L(\hat{\mu}, \hat{\sigma})} \right], \quad (7)$$

sendo L a função de verosimilhança dada por (4). De acordo com Meeker e Escobar (1998, p.627) a distribuição limite de

$$W(t_p) = -2\log(R(t_p)) \quad (8)$$

é um qui-quadrado com um grau de liberdade. Desta forma, um intervalo de $100(1 - \alpha)\%$ confiança aproximado pode ser determinado pelos valores t_p tais que

$$R(t_p) > \exp\left(\frac{-\chi_{(1-\alpha,1)}^2}{2}\right), \quad (9)$$

ou seja, os limites do intervalo de confiança são os zeros da função

$$R(t_p) - \exp\left(\frac{-\chi_{(1-\alpha,1)}^2}{2}\right), \quad (10)$$

onde, $\chi_{(1-\alpha,1)}^2$ é o quantil de probabilidade $(1 - \alpha)$ da distribuição qui-quadrado com um grau de liberdade. A metodologia usada para calcular o intervalo de confiança dado por (10) intitula-se "Razão de Verosimilhanças".

Considere-se agora a função definida por

$$\begin{aligned} V(t_p) &= \text{sinal}(t_p - \hat{t}_p) \sqrt{W(t_p)} \\ &= \text{sinal}(t_p - \hat{t}_p) \sqrt{-2\log(R(t_p))}. \end{aligned} \quad (11)$$

em que \hat{t}_p é o estimador de máxima verosimilhança de t_p , $R(t_p)$ é a função "Profile Likelihood" definida por (7) e a função sinal

$$\text{sinal}(x) = \frac{x}{|x|}. \quad (12)$$

Para o caso de grandes amostras, a distribuição de $V(t_p)$ pode ser aproximada pela distribuição Normal com valor médio igual a zero e variância igual a um (Schmee e Doganaksoy (1993)). Contudo, no caso de pequenas amostras e dependendo dos processos de censura, esta aproximação tende a piorar. Na metodologia PBSRLLR é usada simulação para aproximar a distribuição de $V(t_p)$ por uma distribuição $V^*(t_p)$, chamada distribuição empírica.

De acordo com Davinson e Hinkley (1997), uma estimativa para a função de distribuição de $V(t_p)$ pode ser definida por

$$F_{V^*(t_p)}(t) = \frac{1}{B} \sum_{i=1}^B I\{v_i^* \leq t\} \quad (13)$$

onde I denota a função característica, B o número de amostras usadas na simulação e os v_i^* são calculados como se indica a seguir:

- Geram-se B amostras da distribuição em estudo;
- Para $i = 1, \dots, B$, seja L_i^* a função de verosimilhança da i -ésima amostra simulada e $\hat{\mu}_i^*$, $\hat{\sigma}_i^*$ e \hat{t}_{pi}^* as estimativas de máxima verosimilhança de μ , σ e t_p respectivamente.
- Os valores v_i^* são definidos por

$$\begin{aligned} v_i^* &= \text{sinal}(\hat{t}_p - \hat{t}_{pi}^*) \sqrt{-2 \log(R^*(t_p))} \\ &= \text{sinal}(\hat{t}_p - \hat{t}_{pi}^*) \sqrt{-2 \log \left(\max_{\sigma} \left[\frac{L_i^*(\log(\hat{t}_p) - \Phi^{-1}(p)\sigma, \sigma)}{L_i^*(\hat{\mu}_i^*, \hat{\sigma}_i^*)} \right] \right)}, \\ & \quad i = 1, \dots, B \end{aligned} \quad (14)$$

Seja $v^*(\alpha)$ o "quantil $-\alpha$ " da distribuição empírica $V^*(t_p)$. Este valor é definido como sendo o $(B+1)\alpha$ -ésimo valor ordenado de $V^*(t_p)$, admitindo que $(B+1)\alpha$ é um inteiro. No caso de $(B+1)\alpha$ não ser inteiro, pode ser utilizada uma interpolação linear na escala de quantis Normal, sendo $v^*(\alpha)$ definido por

$$v^*(\alpha) = v_{(k)}^* + \frac{\Phi_{NOR}^{-1}(\alpha) - \Phi_{NOR}^{-1}\left(\frac{k}{B+1}\right)}{\Phi_{NOR}^{-1}\left(\frac{k+1}{B+1}\right) - \Phi_{NOR}^{-1}\left(\frac{k}{B+1}\right)} (v_{(k+1)}^* - v_{(k)}^*) \quad (15)$$

onde $k = [(B+1)\alpha]$, notando por $[x]$ a parte inteira de x e Φ_{NOR} a função de distribuição Normal Standartizada.

Admitindo que $V^*(t_p)$ produz uma boa aproximação à distribuição $V(t_p)$, o intervalo de $100(1-\alpha)\%$ confiança aproximado é dado pelos valores t_p tais que:

$$P\left(v^*\left(\frac{\alpha}{2}\right) \leq V(t_p) \leq v^*\left(1 - \frac{\alpha}{2}\right)\right) \approx 1 - \alpha. \quad (16)$$

Os limites superior e inferior do intervalo de confiança são respectivamente as soluções das equações:

$$V(t_p) - v^*\left(1 - \frac{\alpha}{2}\right) = 0 \quad \text{e} \quad V(t_p) - v^*\left(\frac{\alpha}{2}\right) = 0. \quad (17)$$

4 Exemplo prático

Como aplicação prática desta metodologia foi levado a efeito um estudo de simulação com o objectivo de estudar a evolução da probabilidade de cobertura de intervalos de confiança para o quantil de probabilidade p em função da dimensão da amostra, da proporção de falhas e do valor de p . Gerámos aleatoriamente amostras de uma distribuição LogNormal com parâmetros de localização $\mu = 0$ e escala $\sigma = 1$, de dimensão (n) igual a 10, 15, 20, 25, 30

e 50 sendo cada uma destas amostras censuradas artificialmente nos valores correspondentes a 20%, 30%, 50%, 70% e 90% de falhas (p_f). Para cada uma destas situações foram calculados intervalos de 95% de confiança para os quantis de probabilidade 0.1, 0.3, 0.5, 0.7 e 0.9. Este processo foi repetido 500 vezes a fim de se registar o número de intervalos que continham o parâmetro de interesse t_p , sendo a probabilidade de cobertura aproximada pelo quociente entre o número de intervalos de confiança que contêm o parâmetro t_p e o número total de intervalos determinados.

O valor utilizado para simular a distribuição de $V(t_p)$ foi $B = 1000$.

Além da metodologia PBSRLLR também foram calculados, para as mesmas amostras, intervalos de confiança usando as metodologias LLR (Razão de Verossimilhanças) e NORM (Aproximação à distribuição Normal).

Para cada valor de n e p foi elaborado o respectivo gráfico representando a probabilidade de cobertura dos intervalos de confiança aproximados em função da percentagem de falhas (Figuras (1) e (2)).

5 Conclusões

Observando as Figuras (1) e (2) podemos constatar que a metodologia PBSRLLR se destaca consideravelmente das outras duas metodologias utilizadas. Nas diferentes situações estudadas, PBSRLLR conduz a probabilidades de cobertura próximas da probabilidade nominal, ainda que para amostras de pequena dimensão.

Para amostras com dimensão inferior a 30, a probabilidade de cobertura produzida pelas metodologias NORM e LLR é consideravelmente inferior à probabilidade nominal. No entanto, se a dimensão da amostra for superior a 30, LLR parece conduzir a intervalos de confiança com probabilidade de cobertura mais próxima da probabilidade nominal. Constata-se que esta situação ocorre também ao aplicarmos NORM, mas somente para amostras com dimensão superior a 50.

O trabalho informático requerido por PBSRLLR é significativamente superior ao requerido pelos outros dois métodos e depende fortemente da dimensão da amostra. Aconselhamos a usar PBSRLLR nas situações em que dispomos de amostras de dimensão moderada ($n \leq 50$). Nas situações em que dispomos de grandes amostras, PBSRLLR poderá tornar-se não utilizável devido às limitações informáticas existentes actualmente em termos do tempo necessário para efectuar todas as operações requeridas na implementação do método. No entanto, como nestas situações parece existir uma aproximação dos resultados das três metodologias, sugerimos a utilização de LLR.

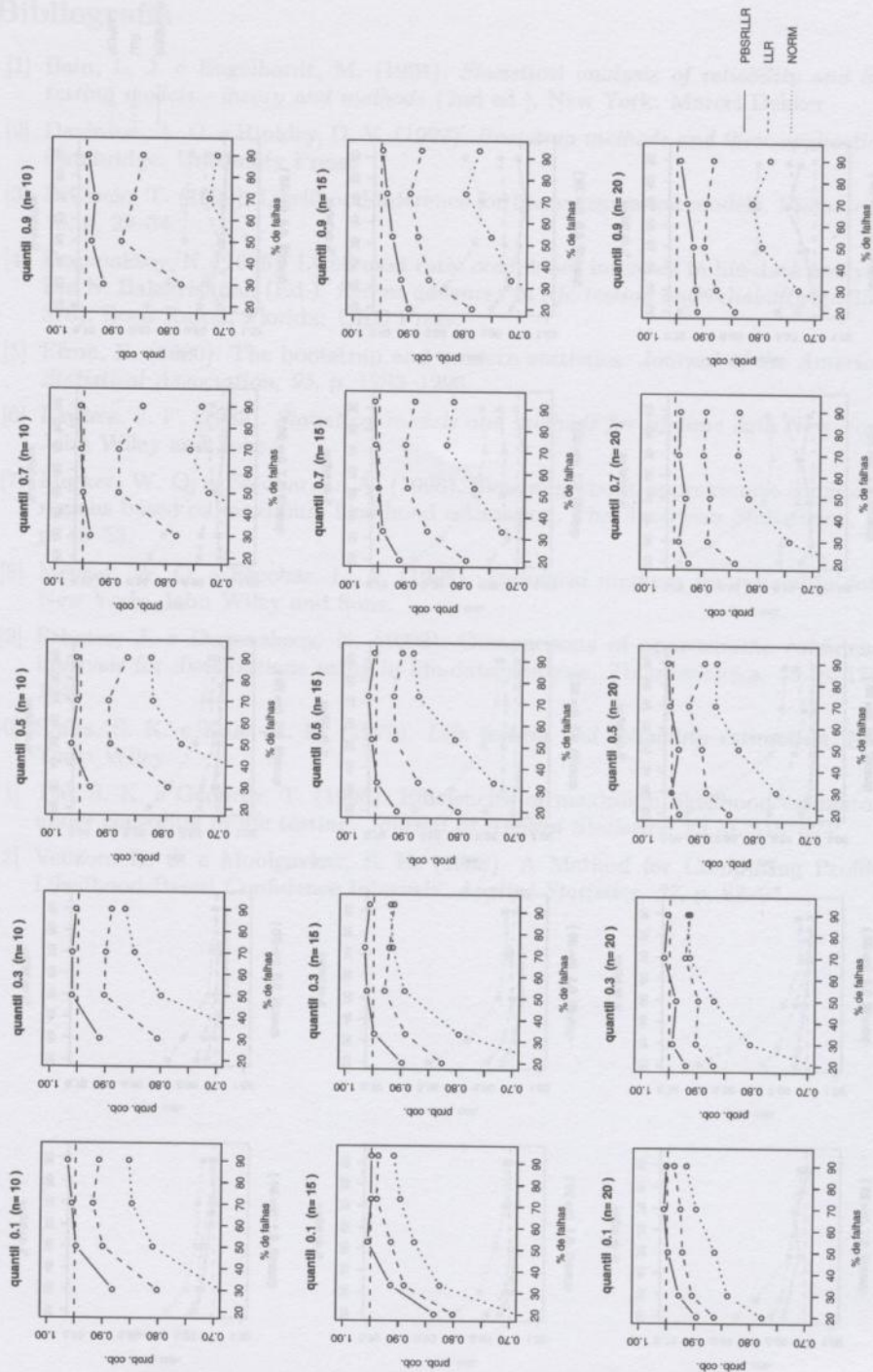


Figura 1: Probabilidade de cobertura (n=10, 15 e 20)

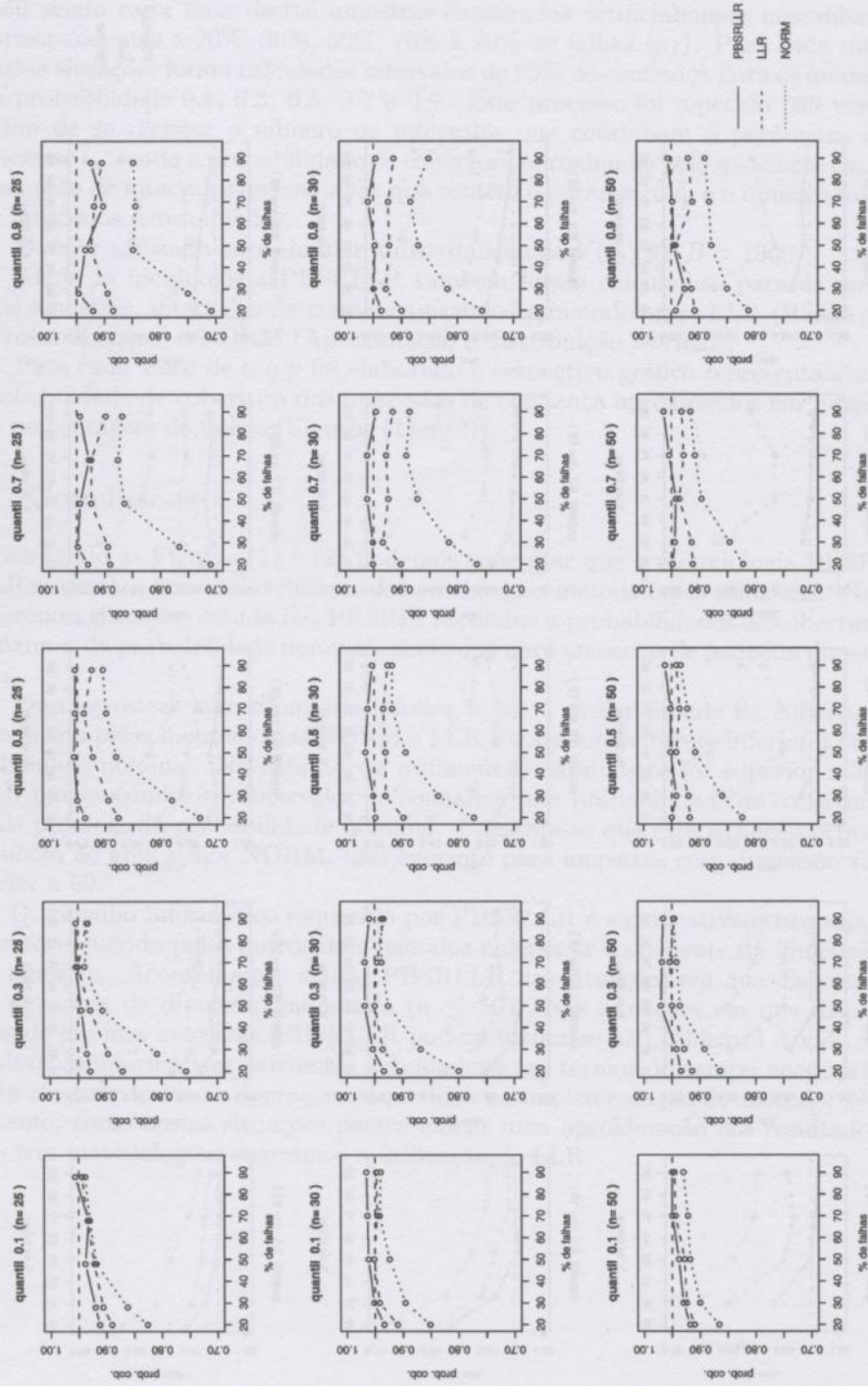


Figura 2: Probabilidade de cobertura (n=25, 30 e 50)

Bibliografia

- [1] Bain, L. J. e Engelhardt, M. (1991). *Statistical analysis of reliability and life-testing models - theory and methods* (2nd ed.). New York: Marcel Dekker.
- [2] Davinson, A. C. e Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: University Press.
- [3] DiCiccio, T. (1990). Likelihood inference for linear regression models. *Biometrika*, 75, p. 29-34.
- [4] Doganaksoy, N. (1995). Likelihood ratio confidence intervals in life-data analysis. Em N. Balakrishnan (Ed.), *Recent advances in life-testing and reliability* (p. 359-376). Boca Raton, Florida: CRC Press.
- [5] Efron, B. (2000). The bootstrap and modern statistics. *Journal of the American Statistical Association*, 95, p. 1293-1296.
- [6] Lawless, J. F. (1982). *Statistical models and methods for lifetime data* New York: John Wiley and Sons.
- [7] Meeker, W. Q. e Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, 49, p. 48-53.
- [8] Meeker, W. Q. e Escobar, L. A. (1998). *Statistical methods for reliability data*. New York: John Wiley and Sons.
- [9] Schmee, J. e Doganaksoy, N. (1993). Comparisons of approximate confidence intervals for distributions using in life-data analysis. *Technometrics*, 35, p. 175-184.
- [10] Sinha, S. K. e Kale, B. K. (1979). *Life testing and reliability estimation*. New York: Wiley.
- [11] Tse, S. K. e Geoffrey, T. (1996). Efficiencies of maximum likelihood estimators under censoring in life testing. *Journal of Applied Statistics*, 23, p. 515-524.
- [12] Venzon, D. J. e Moolgavkar, S. H. (1988). A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Applied Statistics*, 37, p. 87-94.

1. Introdução

Este artigo apresenta a derivação de intervalos de confiança baseados em máxima verossimilhança para parâmetros de uma distribuição de Weibull com um único parâmetro de escala. Os métodos de derivação são comparados com os métodos de derivação baseados em máxima verossimilhança para parâmetros de uma distribuição de Weibull com dois parâmetros. Os resultados são comparados com os resultados obtidos por outros autores. O método para os parâmetros de escala é o método de máxima verossimilhança.

Palavras-chave: Intervalos de confiança, máxima verossimilhança, distribuição de Weibull.