UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

GABRIEL OLIVEIRA DE QUEIROZ MONTEIRO

Aplicação de Métodos *Bootstrap* na Construção de Intervalos de Confiança para os parâmetros da Distribuição Gama

Recife

2022

GABRIEL OLIVEIRA DE QUEIROZ MONTEIRO

Aplicação de Métodos *Bootstrap* na Construção de Intervalos de Confiança para os parâmetros da Distribuição Gama

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador (a): Getúlio José Amorim do Amaral

Recife

2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Monteiro, Gabriel Oliveira de Queiroz.
Aplicação de Métodos Bootstrap na Construção de Intervalos de Confiança para os parâmetros da Distribuição Gama / Gabriel Oliveira de Queiroz Monteiro. - Recife, 2022.
38 : il., tab.

Orientador(a): Getúlio José Amorim do Amaral
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Estatística - Bacharelado, 2022.
Inclui referências, apêndices.

1. Bootstrap. 2. Distribuição Gama. 3. Estimadores de Máxima Verossimilhança. 4. Intervalos de Confiança. I. Amaral, Getúlio José Amorim do. (Orientação). II. Título.

310 CDD (22.ed.)

GABRIEL OLIVEIRA DE QUEIROZ MONTEIRO

Aplicação de Métodos Bootstrap na Construção de Intervalos de Confiança para os parâmetros da Distribuição Gama

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovado em: 25/10/2022

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
GETULIO JOSE AMORIM DO AMARAL
Data: 08/11/2022 17:31:24-0300
Verifique em https://verificador.iti.br

Prof. Dr. Getúlio José Amorim do Amaral (Orientador)

Universidade Federal de Pernambuco

gov.br Documento assinado digitalmente
CALITEIA SANTANA DE SOUSA
Data: 08/11/2022 16:43:15-0300
Verifique em https://verificador.iti.br

Profa. Dra. Calitéia Santana de Sousa (Examinador)

Universidade Federal de Pernambuco

Profa. Dra. Fernanda De Bastiani (Examinador)

Uni gov.br Documento assinado digitalmente
FERNANDA DE BASTIANI
Data: 08/11/2022 20:19:11-0300
Verifique em https://verificador.iti.br uco

AGRADECIMENTOS

Agradeço primeiramente a minha família que sempre me apoiou em todas as escolhas e projetos da minha vida, e sempre fez o máximo possível para investir na minha educação e formação.

Gostaria de agradecer também a todo o corpo docente da Universidade Federal de Pernambuco e ao Departamento de Estatística, já que todos os ensinamentos passados foram de extrema importância para a minha formação, para quem eu sou hoje e para a construção da minha carreira profissional. Em especial, gostaria de agradecer ao meu orientador Getúlio José Amorim do Amaral que me guiou durante não só o desenvolvimento deste trabalho, mas durante toda a minha graduação ao ter me orientado em projetos de iniciação científica e sempre ter se disponibilizado em prestar apoio quando eu precisei. Também agradeço a Renata Maria Cardoso, professora da disciplina de Estatística para Computação do Centro de Informática da UFPE, da qual fui monitor. Este momento foi o meu primeiro contato com a Estatística a nível Superior, e esta ótima experiência foi o que motivou a minha migração para o curso de Estatística.

Deixo também meus sinceros agradecimentos a todos colegas que fiz durante a graduação, que fizeram essa jornada ser mais leve, prazerosa e muito rica em troca de conhecimento e experiências.

RESUMO

As técnicas que envolvem o *bootstrap* consistem em métodos computacionais de reamostragem a partir de uma amostra de origem com o objetivo de mensurar a incerteza e variação de estimadores ao simular mais dados provenientes de uma mesma população, e consequentemente, ao gerar mais estimativas. Neste trabalho, tanto as abordagens paramétrica e a não paramétrica, na geração das réplicas *bootstrap*, foram utilizadas. Os parâmetros de interesse deste estudo são o parâmetro de forma e o de escala da Distribuição Gama, estimados através do método de Máxima Verossimilhança. Com uma amostra simulada desta distribuição de probabilidade, foram aplicados os métodos *bootstrap* percentil e o BC_a (Correção de Viés acelerado) na construção dos Intervalos de Confiança para os parâmetros de forma e de escala da distribuição gama. As diferentes abordagens citadas acima foram aplicadas a dados simulados e os resultados foram comparados e discutidos. O objetivo principal do estudo é, através de intervalos de confiança, mensurar a incerteza e a variação dos estimadores de máxima verossimilhança gerados para os parâmetros da distribuição gama. Todos procedimentos e cálculos necessários para obter os intervalos de confiança *bootstrap* foram desenvolvidos computacionalmente através do *R*.

Palavras-chaves: *Bootstrap*. Distribuição Gama. Estimadores de Máxima Verossimilhança. Intervalos de Confiança.

ABSTRACT

The techniques involving bootstrap consist of computational resampling methods based on an original sample focused on measuring the uncertainty and variation of estimators by simulating more data drawn from the same population, and thus, by generating more estimates. In this work, both parametric and non-parametric approaches in generating bootstrap replications were used. The parameters of interest in this study are the shape and scale parameter from gamma distribution, estimated through the Maximum Likelihood method. With a simulated original sample drawn from this probability distribution, the bootstrap percentile and the BC_a (Bias-Corrected accelerated) methods were applied to build Confidence Intervals for the shape and scale parameters of the gamma distribution. The different approaches mentioned above were applied to simulated data and theirs results were compared and discussed. The main objective is to measure the uncertainty and variation of the maximum likelihood estimators of the gamma distribution parameters through confidence intervals. All of the procedures and calculations necessary to build bootstrap confidence intervals were developed computationally through *R* ecosystem.

Keywords: Bootstrap. Confidence Intervals. Gamma Distribution. Maximum Likelihood Estimators.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A distribuição gama é uma distribuição de probabilidade contínua que é amplamente vista em diversos modelos que buscam analisar dados positivos assimétricos.É comumente utilizada em diferentes áreas do conhecimento, como por exemplo, finanças, atuária, meteorologia e pesca (ver Paula, 2013), por ser capaz de modelar muitos fenômenos que podem ser representados por variáveis de valor real e positivo, e com distribuição assimétrica. Devido a estrutura da sua função de densidade de probabilidade, a distribuição gama pode assumir diferentes formas, o que justifica o fato de que outras distribuições de probabilidade, como por exemplo a distribuição qui-quadrado e a exponencial, possam ser derivadas de uma Gama, dada determinada parametrização.

Ao lidar com problemas de diversas naturezas é comum trabalhar apenas com uma amostra de dados proveniente de uma determinada população. Lidar com todos os dados populacionais é quase sempre inviável por ser de difícil acesso e extremamente custoso. A inferência estatística é rica por permitir a estimação de parâmetros desconhecidos de interesse, através de dados amostrais, extraídos da população estudada. Os parâmetros de uma distribuição gama podem ser estimados através de uma amostra inicial utilizando o método de máxima verossimilhança, o que resulta em estimadores consistentes, eficientes, e se uma estimativa suficiente existe, ela também será dada pelo método, como mencionado em Thom (1958).

Entender o comportamento destes estimadores de parâmetros de interesse se faz necessário. Conhecer a distribuição amostral dos estimadores permite mensurar a incerteza e a variação deles e construir intervalos de confiança. O *bootstrap*, introduzido pro Efron (1979), consiste em métodos que utilizam da reamostragem da amostra original para gerar novas amostras, simulando a coleta de novos dados provenientes de uma mesma população. Como a abordagem *bootstrap* na geração de novas amostras pode ser paramétrica ou não paramétrica, discutiremos o uso de ambas neste estudo. Como exposto em Alves (2013), os métodos *bootstrap* permitem extrair informações acerca da distribuição de uma variável aleatória que não são triviais de serem obtidas por métodos analíticos tradicionais.

A vantagem de usar o método *bootstrap* consiste em estimar, de maneira precisa e de fácil aplicação e compreensão, características de uma distribuição. Ao gerar novas amostras a partir de uma inicial, estamos gerando mais dados acerca da população, que é satisfatório no processo da inferência estatística. A aplicação de técnicas *bootstrap* além de permitir calcular

métricas que mensuram a variação e incerteza, também oferecem uma forma de corrigir o viés de estimadores.

No nosso caso, o foco do estudo será a construção de intervalos de confiança *bootstrap* para os parâmetros de forma e escala da distribuição gama, através de uma amostra original simulada de uma população com distribuição de probabilidade gama, e de estimadores obtidos através do método de máxima verossimilhança. A construção de intervalos de confiança nos traz mais informações sobre a distribuição dos estimadores, e assim, nos ajuda a mensurar a variação e a incerteza sobre eles, já que o valor calculado de uma única estimativa pontual pode variar bastante dependendo da amostra original coletada, e é em cenários como este que as estimativas intervalares possuem grande valor. Para um número alto de amostras *bootstrap* geradas, podemos aproximar a distribuição dos estimadores através da distribuição das réplicas *bootstrap*, obtendo assim, os intervalos de confiança para os parâmetros da distribuição gama. A construção dos intervalos de confiança será feita através dos métodos do *bootstrap* percentil e do BC_a (Correção de Viés Acelerado), ambos abordados por Efron e Tibshirani (1983), e suas aplicações e resultados serão discutidos e comparados.

Todo o processo computacional de estimação dos parâmetros, geração das amostras *bootstrap* e o cálculo de cada estimativa para estas novas amostras será feito através do R. Trechos dos códigos, bibliotecas utilizadas e algoritmos desenvolvidos serão disponibilizados no Apêndice para um melhor entendimento da análise.

1.1 TÓPICOS ABORDADOS

Este trabalho será estruturado em capítulos. Inicialmente, no capítulo 2, iremos dar um contexto sobre a distribuição de probabilidade contínua gama e seus parâmetros. No capítulo 3, iremos tratar sobre o método de estimação de máxima verossimilhança e o cálculo destes estimadores para os parâmetros da distribuição gama, introduzidos no capítulo anterior. No capítulo 4, descreveremos os métodos de reamostragem *bootstrap*, considerando a abordagem paramétrica e a não paramétrica. No capítulo 5, apresentaremos dois métodos distintos que podem ser utilizados na construção de intervalos de confiança através das amostras *bootstrap*, são eles os intervalos *bootstrap* percentil e o BC_a. No capítulo final, apresentaremos e discutiremos, com base em toda teoria apresentada nos capítulos anteriores, os resultados obtidos através de um estudo simulado.

2 DISTRIBUIÇÃO GAMA

A função densidade de probabilidade que define a distribuição gama pode ser escrita sob diferentes composições de parâmetros. Uma delas é em função do parâmetro de forma γ e do parâmetro de escala β. Esta será a parametrização que usaremos para a distribuição gama neste estudo. Uma outra parametrização comumente utilizada é considerando o parâmetro de forma γ e o parâmetro de taxa ou escala inversa assumindo valor igual a β^{-1}.

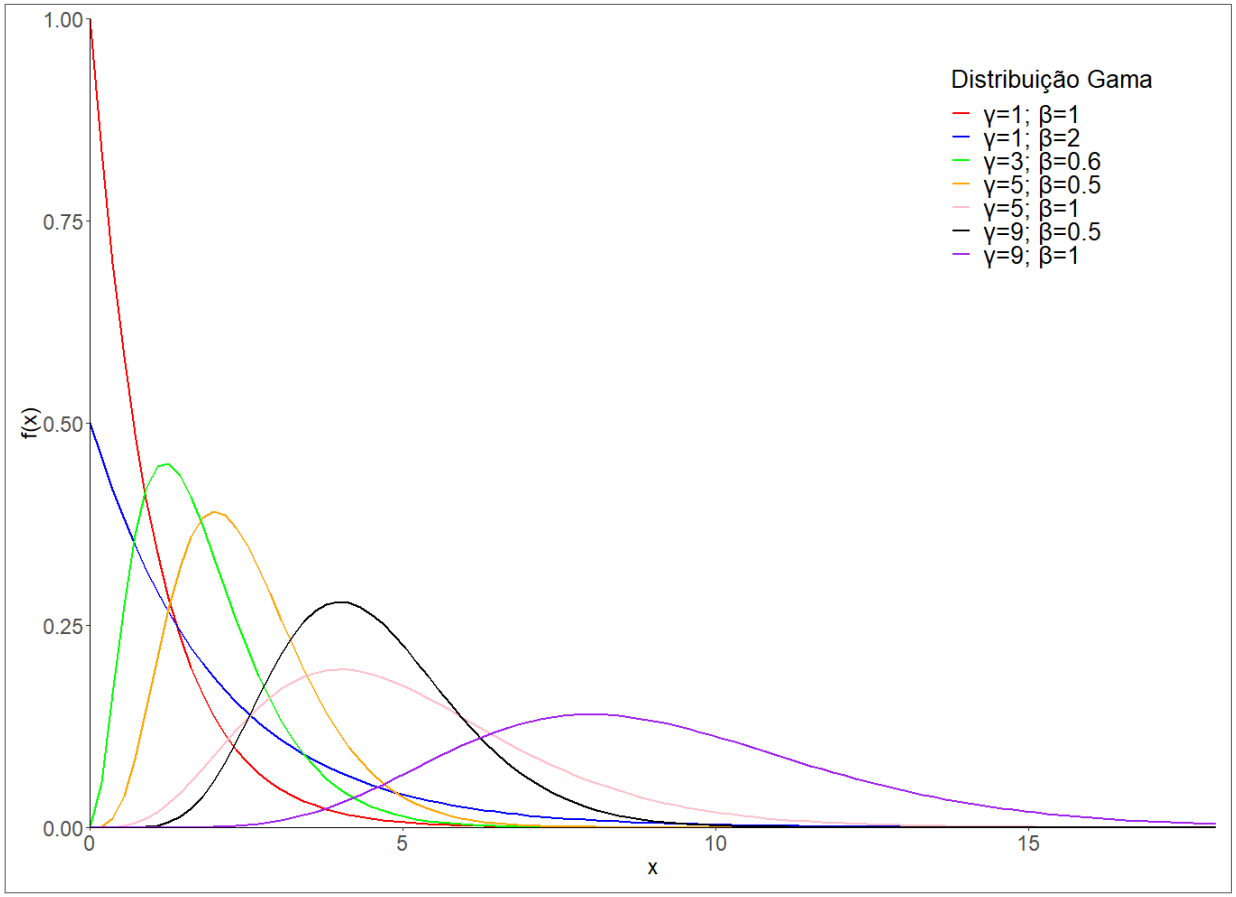
Considerando a primeira parametrização mencionada, uma variável aleatória X segue a distribuição de probabilidade contínua gama se sua função densidade de probabilidade é dada por:

$$f(x|\gamma,\beta) = \frac{1}{\beta^{\gamma}\Gamma(\gamma)} x^{\gamma-1} e^{-x/\beta}; \quad \beta > 0,\ \gamma > 0. \tag{2.1}$$

Em (2.1), Γ é a função Gama, e $f(x|\gamma,\beta) = 0$ para $x < 0$. Logo, é possível notar que a distribuição tem limite inferior em zero e é ilimitada à direita. Podemos ilustrar que uma variável aleatória segue distribuição gama como parâmetros γ e β como $X \sim Gama(\gamma,\beta)$. A distribuição é assimétrica positivamente, e a assimetria depende inversamente do parâmetro γ.

É fácil perceber que a distribuição exponencial é um caso especial da distribuição gama quando $\gamma = 1$. A distribuição gama também tem relação com a qui-quadrado, por exemplo, considerando uma $Gama(\frac{n}{2},2)$, obtemos uma ${\chi^2}_{(n)}$. Exemplos como estes mostram que a distribuição Gama pode assumir diversas formas, e consequentemente modelar variáveis de diferentes comportamentos. Na Figura 1, podemos ver a função de densidade da distribuição Gama com diferentes valores para os parâmetros de forma e escala.

Figura 1 – Função densidade de probabilidade da distribuição gama com diferentes parâmetros



Fonte: Elaborada pelo autor deste trabalho (2022)

2.1 MEDIDAS RELACIONADAS À DISTRIBUIÇÃO GAMA

Os momentos da distribuição gama podem ser encontrados através da seguinte equação escrita em função de γ e β:

$$\mu'_r = \beta^r \gamma(\gamma+1) \ \ldots \ (\gamma+r-1). \tag{2.2}$$

De (2.2), encontramos que a média, ou primeiro momento, da distribuição gama é dada por:

$$\mu'_1 = \beta\gamma. \tag{2.3}$$

A moda da distribuição é igual a $\beta(\gamma-1)$ se $\gamma > 1$. Caso contrário, a moda é zero. Em relação à mediana, não conseguimos encontrar uma forma simples fechada.

Considerando agora os momentos centrados de uma variável aleatória X qualquer, dados por: $\mu_n = \mathbf{E}[(X - \mathbf{E}[X])^n]$, onde $\mathbf{E}[X^n] = \mu'_n$. Logo, o segundo, terceiro e quarto momentos centrados da distribuição gama, que representam respectivamente a variância, a assimetria e a curtose, são:

$$\mu_2 = \sigma^2 = \beta^2\gamma. \tag{2.4}$$

$$\mu_3 = 2\beta^3\gamma. \tag{2.5}$$

$$\mu_4 = 3\beta^4\gamma(\gamma + 2). \tag{2.6}$$

Como apresentado em Magalhães(2006), o coeficiente de assimetria de Pearson para uma variável aleatória qualquer é dado por:

$$\alpha_3 = \frac{\mu_3}{\sigma^3}. \tag{2.7}$$

Avaliando (2.5) em (2,7), chegamos no coeficiente de assimetria para a distribuição gama:

$$\alpha_3 = \frac{2}{\sqrt{\gamma}}. \tag{2.8}$$

De (2.8) é trivial perceber que a assimetria tende a zero a medida que γ cresce. Isso indica que a distribuição gama se torna simétrica em torno da média para valores altos de γ. De fato, como indicado em Paula(2013), para valores altos de γ uma variável aleatória com distribuição gama se aproxima de uma distribuição normal com média $\mu = \gamma\beta$ e variância $\sigma^2 = \mu^2/\gamma = \gamma\beta^2$. Isso ilustra que além do fato da distribuição gama ser apropriada para o estudo de variáveis aleatórias com distribuição assimétrica positiva, também é interessante utilizá-la em modelagens de variáveis simétricas em que a variância depende de um componente quadrático da média.

Com a apresentação e contextualização da distribuição gama e suas características, iremos avançar para o próximo tópico, onde discutiremos sobre os estimadores de máxima verossimilhança para os parâmetros γ e β da distribuição gama.

3 ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA

Ao lidar com distribuições cujos parâmetros são desconhecidos, é comum fazer inferências acerca destes parâmetros através de dados observados de uma amostra coletada da população que segue uma determinada distribuição.

Uma das formas mais antigas de proceder com a estimação dos parâmetros de uma distribuição de probabilidade é através da máxima verossimilhança. Apresentado por Fisher (1922), o método consiste em escolher uma estimativa de um parâmetro ou vetor de parâmetros desconhecidos, que denotaremos por θ, através de uma amostra particular que, para variáveis aleatórias discretas, maximize a probabilidade de obter esta amostra particular. Para o caso das variáveis contínuas, o método é baseado em encontrar as estimativas de θ que maximizem a função densidade de probabilidade dado a amostra inicial observada. Algumas das propriedades do estimador de máxima verossimilhança(EMV) mais importantes e que são garantidas pelo método são a eficiência e consistência. Para mais detalhes, consultar Hoel et al. (1971). Como consequência destas qualidades, o EMV é amplamente utilizado.

3.1 DEFINIÇÃO MATEMÁTICA

Dado um conjunto de valores de uma amostra aleatória denotado por $X_1 = x_1, \ldots, X_n = x_n$ coletada de uma população com função de distribuição $f(x|\theta)$, a função de verossimilhança é:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{3.1}$$

O estimador de máxima verossimilhança é dado pelo valor de θ, denotado por $\hat{\theta}$, que maximiza a função $L(\theta)$, isto é, o valor do parâmetro θ que torna a amostra aleatória utilizada no processo do cálculo do estimador $\hat{\theta}$ como sendo a "mais provável" dentre todas amostras possíveis de serem extraídas da população de referência, como discutido em Bussab e Morettin(2013). Maximizar $L(\theta)$ é equivalente a maximizar o $logL(\theta)$, que por sua vez é mais simples de ser calculado. A função $\ell(\theta)$ é conhecida como log-verossimilhança e é definida por:

$$\ell(\theta) = logL(\theta) = \sum_{i=1}^{n} logf(x_i|\theta). \tag{3.2}$$

Logo, o EMV $\hat{\theta}$ de θ é tal que:

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta). \tag{3.3}$$

De (3.3), podemos encontrar o EMV de θ, isto é $\hat{\theta}$, resolvendo a seguinte equação:

$$\frac{\partial}{\partial\theta}\ell(\theta) = 0. \tag{3.4}$$

A matriz de variância-covariância dos EMV pode ser calculada através de:

$$Var(\hat{\theta}) = I^{-1}(\hat{\theta}). \tag{3.5}$$

Em que $I(\hat{\theta})$ é a matriz de informação de Fisher, detalhada em Ly et al.(2017).

Através da matriz indicada em (3.5), é possível obter as estimativas para o erro padrão dos estimadores de máxima verossimilhança, calculando a raiz quadrada dos elementos da diagonal da matriz.

3.2 ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA DOS PARÂMETROS DA DISTRIBUIÇÃO GAMA

Aplicando a função densidade de probabilidade da distribuição gama (2.1) na função de log-verossimilhança (3.2), obtemos a log-verossimilhança da distribuição gama:

$$\ell(\gamma,\beta) = -n\gamma\log\beta - n\log\Gamma(\gamma) + (\gamma-1)\sum_{i=1}^{n}\log x_i - \frac{1}{\beta}\sum_{i=1}^{n}x_i. \tag{3.6}$$

Avaliando a log-verossimilhança da distribuição gama (3.6) e calculando suas derivadas parciais de primeira ordem em relação a β e γ e igualando-as a zero, como na equação (3.4), chegamos em:

$$\frac{\partial}{\partial\beta}\ell(\gamma,\beta) = \frac{\bar{x}}{\hat{\beta}} - \hat{\gamma} = 0. \tag{3.7}$$

$$\frac{\partial}{\partial\gamma}\ell(\gamma,\beta) = \log\hat{\beta} + \psi(\hat{\gamma}) - \frac{1}{n}\sum_{i=1}^{n}\log x_i = 0. \tag{3.8}$$

Onde $\psi(\hat{\gamma})$ é a função digamma, definida como $\psi(\hat{\gamma}) = \frac{\partial}{\partial\hat{\gamma}}\log\Gamma(\hat{\gamma})$.

Da equação (3.7) é fácil perceber que:

$$\hat{\beta} = \frac{\bar{x}}{\hat{\gamma}}. \tag{3.9}$$

Substituindo (3.9) em (3.8), resulta na equação:

$$\log\hat{\gamma} - \psi(\hat{\gamma}) - \log\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\log x_i = 0. \tag{3.10}$$

Encontrar os estimadores de máxima verossimilhança para γ e β é equivalente a resolver as equações (3.9) e (3.10), em que a primeira depende da solução da segunda. Porém, a solução para equação (3.10) não é trivial. Por não ser de forma fechada, a solução desta equação pode ser encontrada por métodos numéricos. Um deles, e que foi o utilizado neste trabalho, é o método de Newton-Raphson.

3.2.1 Método de Newton-Raphson

O método de Newton-Raphson é um método numérico utilizado para encontrar uma aproximação das raízes de uma função real. O método é conhecido por sua rápida convergência, já que para cada iteração é necessário apenas calcular a função e a derivada de primeira ordem da função, avaliadas em um determinado ponto.

Considerando $f(x)$ como uma função contínua no intervalo $[a, b]$ e seja r uma raiz desta função real, com $r \in (a, b)$, $f(r) = 0$ e com a derivada de primeira ordem de $f(x)$ diferente de zero. Como apresentado em Lobão(2017), a ideia por trás do método consiste em primeiramente definir um "chute inicial" x_0 para a raiz da função $f(x)$. A partir deste ponto de partida, calcula-se uma nova observação referente a uma possível raiz de $f(x)$ da seguinte forma:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}. \tag{3.11}$$

Para um determinado valor inteiro e positivo de n e um valor pré-determinado do $erro$, o qual representa uma componente do critério de parada do algoritmo de Newton-Raphson, ao indicar o valor máximo de diferença entre os valores de duas raízes em 2 iterações quaisquer consecutivas, se $|x_n - x_{n-1}| <= erro$, então x_n é a raiz desejada, caso contrário, devemos calcular x_{n+1} através de (3.11) e avaliar novamente $|x_{n+1} - x_n|$.

Definir uma raiz qualquer e distante dos valores finais aproximados pelo método pode resultar em uma execução do algoritmo mais custosa, já que levará mais iterações para convergir para a raiz final. Voltando para a análise dos parâmetros da distribuição gama, uma solução é utilizar a estimação através do método dos momentos para encontrar os "chutes iniciais" de γ e β, referentes a x_0, de modo que seja necessário menos iterações do algoritmo de Newton-Raphson para encontrar o valor final de $\hat{\gamma}$ e $\hat{\beta}$ resultante do método.

3.2.1.1 Estimativa do ponto de partida através do Método dos Momentos

Iremos brevemente apresentar como calcular o ponto de partida de maneira mais eficiente para a execução do método de Newton-Raphson. Shang(2021) apresenta os estimadores através do método dos momentos para os parâmetros γ e β da distribuição gama como:

$$\hat{\gamma} = \frac{\bar{\mu}^2}{\hat{\sigma}^2}. \tag{3.12}$$

$$\hat{\beta} = \frac{\hat{\sigma}^2}{\bar{\mu}}. \tag{3.13}$$

Logo, com a amostra inicial que utilizamos para obter os EMV, também podemos calcular os estimadores através do método dos momentos, como indicado em (3.12) e (3.13), para fornecer um melhor ponto de partida para o algoritmo Newton-Raphson no processo de achar a solução para (3.4).

Agora que já apresentamos a distribuição gama e definimos a abordagem utilizada para estimar seus parâmetros, podemos introduzir o método *bootstrap* no próximo capítulo.

4 AMOSTRAS BOOTSTRAP

O processo de reamostragem consiste em gerar novas amostras a partir de uma amostra inicial. Em muitas situações é comum ter apenas uma amostra extraída de uma população com distribuição F disponível. O processo de reamostragem permite obter novas amostras, através da reamostragem da amostra inicial observada, e calcular novas estimativas acerca de um conjunto de parâmetros de interesse da distribuição F.

Introduzido por Efron(1979), o método *bootstrap* é uma técnica estatística de reamostragem. Consideremos uma amostra aleatória original $x = (x_1, x_2, \ldots, x_n)$, observada de uma população com distribuição desconhecida F, e que desejemos estimar, através de x, um parâmetro (ou vetor de parâmetros) θ de tal forma que $\theta = t(F)$. Calculando a estimativa, dada por $\hat{\theta} = s(x)$, Efron e Tibshirani (1993) questionam o quão preciso é o estimador $\hat{\theta}$ obtido a partir da amostra original x. Este questionamento é válido pois não conhecemos a distribuição amostral do estimador $\hat{\theta}$, e por isso, é neste cenário que método *bootstrap* tem grande impacto.

Considerando a amostra original x de F já apresentada, podemos gerar uma nova amostra bootstrap a partir de x definida por:

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*). \tag{4.1}$$

Em que $x_1^*, x_2^*, \ldots, x_n^*$ são selecionados aleatoriamente e com reposição a partir dos elementos de x. A partir da amostra *bootstrap* x^*, podemos calcular a réplica *bootstrap*, isto é, $\hat{\theta^*} = s(x^*)$. Replicado este processo de reamostragem e cálculo do estimador $\hat{\theta^*}$ B vezes, é possível aproximar a distribuição de $\hat{\theta}$ através da distribuição empírica F^* das réplicas *bootstrap* $\hat{\theta_b^*}$ obtidas para $b = 1, 2, \ldots, B$, onde B é o número total de amostras *bootstrap* reproduzidas.

Desta forma, conseguimos ter maior conhecimento sobre $\hat{\theta}$ e consequentemente fazer inferências mais assertivas sobre os parâmetros θ da população F. Informações sobre viés e erro padrão de $\hat{\theta}$ podem ser estimadas através das réplicas *bootstrap*. Podemos também construir intervalos de confiança para θ através de $\hat{\theta_b^*}$, que será o foco deste trabalho e será o tema abordado no próximo capítulo.

Apesar de ser computacionalmente custoso, este método é de fácil entendimento, traz ótimos resultados e é facilmente aplicado nas mais diversas situações, por ser livre de suposições acerca dos dados trabalhados, que são comumente assumidas em outras abordagens

estatísticas.

A forma de gerar uma amostra *bootstrap* x^* a partir de x pode ser feita pelo método paramétrico e não paramétrico. O *bootstrap* paramétrico assume uma distribuição de probabilidade que originou os dados. Já o *bootstrap* não paramétrico não assume uma distribuição para os dados e utiliza o próprio conjunto original x para gerar novas amostras. A seguir, apresentaremos em detalhes como executar as duas abordagens.

4.1 BOOTSTRAP NÃO PARAMÉTRICO

Uma das vantagens do método não paramétrico é que não precisamos assumir nenhuma distribuição para a população da qual a amostra original foi extraída. Desta forma, temos uma abordagem genérica e que traz resultados satisfatórios para as mais diversas distribuições de probabilidades.

Consideremos novamente $x = (x_1, x_2, \ldots, x_n)$, uma amostra aleatória de uma população com distribuição F, com parâmetros dado por θ. Como reproduzido em Dore et al.(2016), as réplicas *bootstrap* podem ser calculadas a partir da reamostragem não paramétrica de acordo com as seguintes etapas:

1. Selecione B amostras aleatórias de mesmo tamanho n com reposição a partir de x.
2. Para cada amostra bootstrap x^{*b} gerada, calcule o estimador $\hat{\theta}_b^*$ utilizando os dados obtidos na amostra, em que $b = 1, 2, \ldots, B$.

4.2 BOOTSTRAP PARAMÉTRICO

Reproduzindo os mesmos dados considerados na Seção 4.1, as réplicas *bootstrap* podem ser obtidas a partir da reamostragem paramétrica de acordo com o seguinte algoritmo:

1. Calcule o estimador $\hat{\theta}$ de θ de forma que $\hat{\theta} = s(x)$, em que $s(x)$ é uma função da amostra original x.
2. Obtenha B amostras independentes de mesmo tamanho n a partir F($\hat{\theta}$).
3. Para cada amostra bootstrap x^{*b} gerada, calcule o estimador $\hat{\theta}_b^*$ utilizando os dados obtidos na amostra, em que $b = 1, 2, \ldots, B$.

Introduzido o método *bootstrap*, vamos avançar para o próximo capítulo, onde discutiremos sobre a construção de intervalos de confiança para os determinados parâmetros θ de uma distribuição qualquer F.

5 INTERVALOS DE CONFIANÇA *BOOTSTRAP*

Até este capítulo, apresentamos noções sobre estimadores pontuais, isto é, $\hat{\theta} = s(x)$, em que $s(x)$ é uma função qualquer de uma amostra x da distribuição F com parâmetros θ. Para diferentes amostras retiradas de F, a estimativa obtida pelo estimador pontual pode ser diferente. Logo, apenas pela avaliação de um estimador pontual, não podemos determinar qual a dimensão do erro que estamos cometendo ao estimar θ. Intervalos de confiança auxiliam nesta tarefa, ao gerar estimativas intervalares para um determinado parâmetro, através da distribuição amostral do estimador pontual.

Supondo que sejam coletadas aleatoriamente n amostras de mesmo tamanho de F. Para cada amostra, construímos um intervalo de confiança para θ com nível de significância $\alpha = 0.05$, isto é, nível de confiança equivalente a 95%. Para intervalos de confiança exatos, seria esperado que 95% destes intervalos englobassem o valor real do parâmetro θ. Esta é a interpretação para um intervalo de confiança com nível de significância α.

Abordaremos nesse capítulo a teoria por trás da construção de intervalos de confiança *bootstrap* através de dois métodos. A partir de agora, consideraremos $\hat{\theta}$ como sendo o estimador de máxima verossimilhança de θ definido no Capítulo 3.

Para calcular um intervalo de confiança é necessário conhecer a distribuição de $\hat{\theta}$. Contudo, alguns estimadores possuem a distribuição assintótica conhecida: é o caso do estimador de máxima verossimilhança. o EMV é assintoticamente normal desde que: a função de log-verossimilhança $\ell(\theta)$ seja três vezes continuamente diferenciável, o valor esperado de todas as derivadas de primeira e segunda ordem existam e as derivadas de terceira ordem sejam limitadas por uma função com valor esperado finito (ver Dalitz, 2017). Hoel et al.(1971) também mostra que para grandes amostras, isto é, para altos valores de n, a distribuição de $\hat{\theta}$ se aproxima de uma distribuição normal.

Logo, é possível construir um intervalo de confiança para um parâmetro θ a partir de $\hat{\theta}$ baseado na distribuição normal:

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}. \tag{5.1}$$

Desta forma, podemos pensar em duas alternativas para estimar a variância do estimador de máxima verossimilhança $\hat{\theta}$. Uma delas é pela equação (3.5) apresentada no Capítulo 3. Porém, quando as condições já mencionadas para o EMV ter distribuição assintótica normal

não forem atendidas, a matriz hessiana não pode ser calculada por conta das derivadas de segunda ordem. A outra opção é pelo estimador *jackknife*, que será brevemente apresentado a seguir.

5.1 ESTIMADOR *JACKKNIFE*

Como apresentado em Efron e Tibshirani(1993), o método *jackknife* consiste na ideia de calcular n vezes um estimador $\hat{\lambda}$ a partir de uma amostra dada por $x_1, \ldots, x_n$, porém em cada cálculo da réplica *jackknife* devemos desconsiderar um valor x_i da amostra. Desta forma, obtemos n amostras *jackknife* de tamanho $n-1$. Considerando $\hat{\lambda}_{(i)}$ como sendo o estimador calculado sem o ponto x_i, então o estimador *jackknife* para o erro padrão de $\hat{\lambda}$ é dado por:

$$\hat{se}_{jack} = \Big[\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\lambda}_{(i)} - \hat{\lambda}_{(\cdot)})^2\Big]^{1/2}, \tag{5.2}$$

onde na equação (5.2), $\hat{\lambda}_{(\cdot)} = \sum_{i=1}^{n} \hat{\lambda}_{(i)}/n$.

Apesar da variância do estimador $\hat{\theta}$ poder facilmente ser computada e não requisitar o cálculo de nenhuma derivada, Dalitz(2017) menciona que não temos garantia que o estimador $\hat{\theta}$ é normalmente distribuído e nem que o estimador *jackknife* para o erro padrão de $\hat{\theta}$ é um bom estimador. Logo, a construção do intervalo de confiança para θ através da equação (5.1) nem sempre terá resultados satisfatórios. Podemos optar então por construir intervalos de confiança *bootstrap*. Efron e Tibshirani(1993) afirmam que para construir intervalos de confiança *bootstrap*, com boas estimativas dos limites de confiança, são necessárias mais de 500 réplicas *bootstrap*. Assim, para a teoria que será apresentada a seguir e para os estudos dos dados simulados, consideraremos um valor grande, $B = 1000$.

5.2 INTERVALO *BOOTSTRAP* PERCENTIL

O primeiro intervalo de confiança *bootstrap* apresentado será o percentil. A ideia por trás deste intervalo de confiança é muito simples por apenas se basear nos percentis das réplicas *bootstrap* parar definir os limites inferior e superior do intervalo, como veremos a seguir.

Para B amostras *bootstrap* $x^{*1}, \ldots, x^{*B}$, onde B é um valor finito, e calculadas as réplicas *bootstrap* $\hat{\theta}_b^*$ para cada amostra, o intervalo de confiança *bootstrap* percentil aproximado para θ com nível de confiança $1-2\alpha$ é dado por:

$$[\hat{\theta}_{\%,inf}, \hat{\theta}_{\%,sup}] \approx [\hat{\theta}_B^{*(\alpha)}, \hat{\theta}_B^{*(1-\alpha)}]. \tag{5.3}$$

Onde $\hat{\theta}_B^{*(\alpha)}$ é o (100α)º percentil empírico dos valores de $\hat{\theta}_b^*$. Analogamente, $\hat{\theta}_B^{*(1-\alpha)}$ é o o $(100(1-\alpha))$º percentil empírico dos valores de $\hat{\theta}_b^*$.

Além do *bootstrap* percentil ser de fácil compreensão, Efron e Tibshirani(1993) mostraram que este intervalo possui algumas propriedades interessantes de serem observadas:

Propriedade 1. Invariância a Transformações Monótonas:

Um intervalo de confiança para um parâmetro qualquer θ é dito ser invariante a transformações monótonas se um intervalo para qualquer transformação monótona do parâmetro, dado por $\phi = m(\theta)$, é o intervalo de confiança de θ mapeado por $m(\theta)$. Pensando no intervalo *bootstrap* percentil, temos:

$$[\hat{\phi}_{\%,inf}, \hat{\phi}_{\%,sup}] \approx [m(\hat{\theta}_{\%,inf}), m(\hat{\theta}_{\%,sup})]. \tag{5.4}$$

Efron e Tibshirani(1993) exemplificaram a importância desta propriedade para um intervalo de confiança, ao comparar um intervalo *bootstrap* percentil com um intervalo normal padrão semelhante ao apresentado na equação (5.1), que por sua vez não possui essa propriedade, como também apontado por Efron e Tibshirani(1993).

Propriedade 2. Preservação da Amplitude:

Para alguns parâmetros de algumas distribuições, existem restrições acerca dos valores que estes parâmetros podem assumir. Os parâmetros da distribuição gama, por exemplo, devem ser maiores que zero. Logo, esperamos que os intervalos de confiança "respeitem"essas restrições e produzam apenas intervalos de valores que possam, de fato, ser assumidos pelos parâmetros. Estes é a propriedade da Preservação da Amplitude, e intervalos com essa característica tendem a ser mais precisos e confiáveis.

5.3 INTERVALO *BOOTSTRAP* BC$_a$

O intervalo *bootstrap* percentil visto na Seção 5.2 é um dos intervalos *bootstrap* mais simples de ser entendido e aplicado, porém nem sempre é o método *bootstrap* mais indicado para construção de intervalos de confiança.

O método de Correção do Viés Acelerado, ou BC_a, pode ser considerado como uma versão melhorada do *bootstrap* percentil tanto na teoria quanto na prática, onde esperamos

dos intervalos *bootstrap* BC_a uma cobertura maior em relação ao verdadeiro valor de um parâmetro θ.

Assim como o *bootstrap* percentil, os limites inferiores e superiores do intervalo de confiança *bootstrap* BC_a também são dados através dos percentis empíricos da distribuição das réplicas *bootstrap*, mas não necessariamente os mesmos percentis definidos em (5.3).

Os percentis referente aos limites do intervalo BC_a irão depender de dois valores que deveremos calcular: $\hat{a}$ e $\hat{z}_0$, que são, respectivamente, um fator de aceleração e um de correção do viés. Como já evidenciado pelo próprio nome do método, os intervalos BC_a propõem corrigir o viés das estimativas obtidas por $\hat{\theta}$.

Considerando novamente B amostras *bootstrap* $x^{*1}, \ldots, x^{*B}$ e calculadas as réplicas *bootstrap* $\hat{\theta}_b^*$ para cada amostra, o intervalo de confiança *bootstrap* BC_a para θ, com nível de confiança $1 - 2\alpha$, é dado por:

$$[\hat{\theta}_{inf}, \hat{\theta}_{sup}] = [\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}]. \tag{5.5}$$

onde

$$\alpha_1 = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\Big) \;\; ; \;\; \alpha_2 = \Phi\Big(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\Big). \tag{5.6}$$

Em (5.6), Φ representa a função de distribuição de probabilidade acumulada de uma normal padrão. Já $z^{(\alpha)}$ representa o ponto referente ao (100α)º percentil da distribuição normal padrão. É fácil perceber que para $\hat{a} = z^{(\alpha)} = 0$, o intervalo BC_a é equivalente ao obtido pelo *bootstrap* percentil. Valores de $\hat{a}, z^{(\alpha)} \neq 0$ alteram os percentis usados no cálculo dos limites do intervalo, e são essas alterações responsáveis por corrigir algumas deficiências no *bootstrap* percentil, como explicado em Efron e Tibshirani(1993).

Voltando para as constantes $\hat{a}$ e $\hat{z}_0$, podemos calcular $\hat{z}_0$ através da proporção de réplicas *bootstrap* com valor menor do que a estimativa $\hat{\theta}$ da amostra original:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B}\right), \tag{5.7}$$

onde $\Phi^{-1}(\cdot)$ representa a inversa da função de distribuição de probabilidade acumulada de uma normal padrão. De maneira prática, $\hat{z}_0$ avalia a diferença entre a mediana de $\hat{\theta}^*$ e $\hat{\theta}$.

Já o fator de aceleração $\hat{a}$, representa a taxa da mudança do erro padrão de $\hat{\theta}$ com relação ao verdadeiro valor do parâmetro θ. Uma das formas mais simples de calculá-lo é através do princípio *jackknife*, que já foi apresentado na Seção 5.1. Daí, temos:

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}}. \tag{5.8}$$

Assim como o método *bootstrap* percentil, os intervalos de confiança BC_a possuem as propriedades 1 e 2 indicadas na Seção 5.2. Além disso, uma outra vantagem do método BC_a sobre o percentil é em relação a acurácia dos intervalos. Os intervalos BC_a são acurados de segunda ordem, enquanto os intervalos gerados pelo método *bootstrap* percentil são acurados de primeira ordem. Para mais detalhes, ver Efron e Tibshirani(1993).

Avançaremos para a última seção deste Capítulo, onde descreveremos algumas formas de avaliar um determinado intervalo de confiança.

5.4 AVALIAÇÃO DE INTERVALOS DE CONFIANÇA

Antes de avançar para o próximo Capítulo, é válido apresentar alguns conceitos que utilizaremos para descrever intervalos de confiança no estudo de simulação. Para isto, consideremos um intervalo de confiança qualquer representado por $[\hat{\theta}_{inf}, \hat{\theta}_{sup}]$.

5.4.1 Amplitude do Intervalo de Confiança

Este conceito é bem simples e auto-explicativo. A amplitude de um intervalo de confiança pode ser definida por:

$$amplitude = \hat{\theta}_{sup} - \hat{\theta}_{inf}. \tag{5.9}$$

5.4.2 Forma do Intervalo de Confiança

Como o próprio nome já diz, essa métrica mensura a assimetria do intervalo em relação à estimativa pontual $\hat{\theta}$ e é calculada da seguinte maneira:

$$forma = \frac{\hat{\theta}_{sup} - \hat{\theta}}{\hat{\theta} - \hat{\theta}_{inf}}. \tag{5.10}$$

De (5.10) é fácil perceber que para valores de $forma > 1$, temos uma maior distância de $\hat{\theta}_{sup}$ de $\hat{\theta}$ do que de $\hat{\theta}$ até $\hat{\theta}_{inf}$. Podemos partir desta mesma lógica para tirar conclusões da assimetria quando a $forma < 1$. Para valor de $forma = 1$, temos que o intervalo é simétrico em relação a $\hat{\theta}$.

5.4.3 Cobertura do Intervalo de Confiança

Para obter este valor de cobertura podemos simular, através de um método escolhido, M intervalos de confiança(onde M é um valor inteiro grande) para um parâmetro θ conhecido através de M amostras diferentes extraídas de uma mesma população. A partir destes intervalos, calculamos a proporção referente a quantos deles englobam o valor verdadeiro do parâmetro θ.

Com todos esses conceitos definidos, podemos avançar para o próximo capítulo, onde simularemos dados pertencentes a uma distribuição gama com determinados parâmetros γ e β e estimaremos seus parâmetros através do método de máxima verossimilhança já apresentado no capítulo 3. Por fim, construíremos intervalos de confiança *bootstrap* para os dois parâmetros e discutiremos os resultados obtidos.

6 ESTUDO DE SIMULAÇÃO

Agora, utilizaremos toda teoria vista nos capítulos anteriores para construir intervalos de confiança *bootstrap* para o parâmetro de forma γ e O parâmetro de escala β, da distribuição gama, a partir de dados simulados. Os estimadores $\hat{\gamma}$ e $\hat{\beta}$ são os estimadores de máxima verossimilhança de γ e β, respectivamente. Para este experimento, todos os processos computacionais serão realizados através da linguagem de programação *R*. O cálculo dos estimadores de máxima verossimilhança $\hat{\gamma}$ e $\hat{\beta}$ serão realizados através do mesmo procedimento descrito no Capítulo 3, porém com auxilio da biblioteca *"maxLik"* disponível no *R*. Todas as demais funções utilizadas neste estudo foram desenvolvidas no *R* e podem ser visualizadas no Apêndice A.

Consideremos uma amostra aleatória de tamanho moderado,$n = 40$, originada de uma distribuição gama, com parâmetro de forma $\gamma = 9$ e parâmetro de escala $\beta = 0.5$. Os estimadores de máxima verossimilhança calculados são $\hat{\gamma} = 10.2$ e $\hat{\beta} = 0.454$. O número de réplicas *bootstrap* geradas será de $B = 1000$. Por sua vez, as amostras *bootstrap* serão obtidas através de ambas abordagens paramétrica, considerando a distribuição gama com parâmetros $\hat{\gamma}$ e $\hat{\beta}$, e não paramétrica apresentadas no capítulo 4. Uma vez obtidas, é importante avaliar o comportamento da distribuição *bootstrap* das réplicas $\hat{\gamma}_b^*$ e $\hat{\beta}_b^*$.

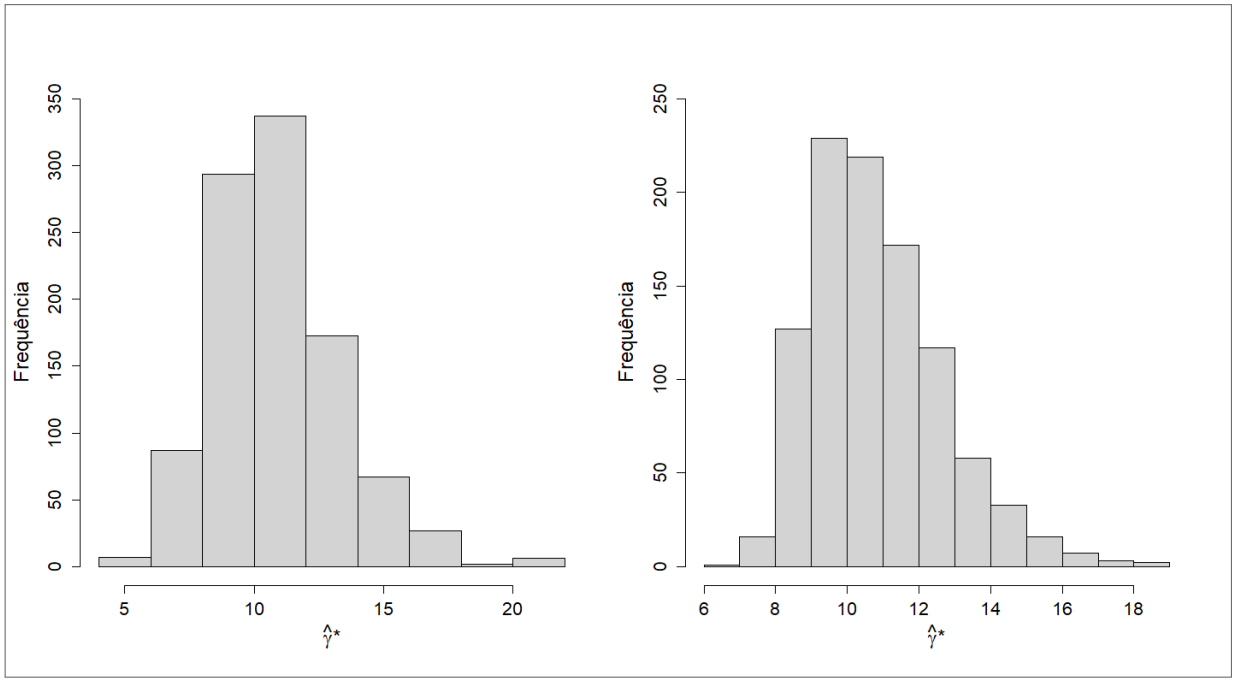
Da Figura 2, percebemos que tanto a amostragem feita pelo método *bootstrap* paramétrico quanto pelo não paramétrico resultaram em distribuições assimétricas de $\hat{\gamma}_b^*$, com as caudas da direita mais pesadas.

Pela Figura 3, observamos que a amostragem *bootstrap* paramétrica resultou em uma distribuição assimétrica de $\hat{\beta}_b^*$, porém para o método não paramétrico, a distribuição de $\hat{\beta}_b^*$ está mais próxima de uma normal.

A partir desta análise inicial, observamos que intervalos de confiança tradicionais que são baseados na distribuição normal padrão não seriam tão adequados nesse cenário, já que a distribuição dos dados não é normal, com exceção da distribuição de $\hat{\beta}_b^*$ obtida pela amostragem não paramétrica.

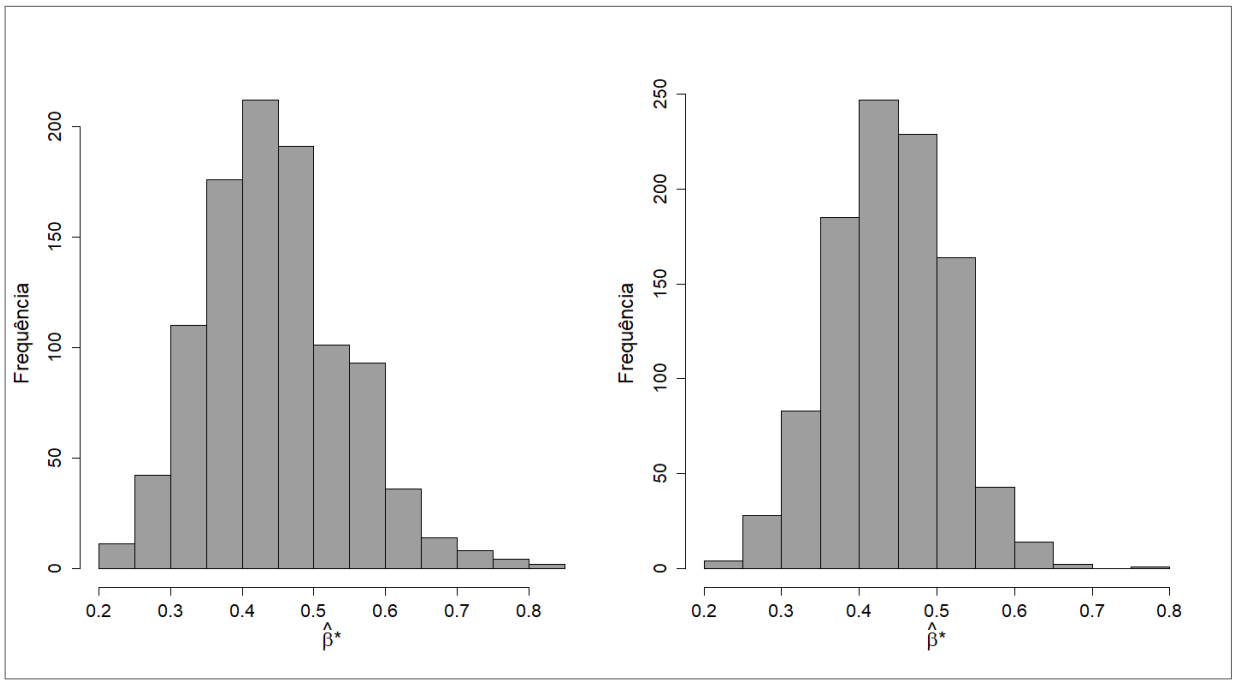
Para a construção dos intervalos de confiança para γ e β, serão aplicados os métodos *bootstrap* percentil e BC_a, considerando a distribuição empírica de $\hat{\gamma}_b^*$ e $\hat{\beta}_b^*$ para $b = 1, \ldots, B$, resultante de cada uma das abordagens de amostragem *bootstrap* executadas. Tomaremos $\alpha = 0.05$, isto é, o nível de confiança dos intervalos será de 95%. Calcularemos as métricas

Figura 2 – Gráfico da esquerda: Distribuição de $\hat{\gamma}_b^*$ obtidos pela amostragem *bootstrap* paramétrica. Gráfico da direita: Distribuição $\hat{\gamma}_b^*$ obtidos pela amostragem *bootstrap* não paramétrica.



Fonte: Elaborado pelo autor (2022)

Figura 3 – Gráfico da esquerda: Distribuição de $\hat{\beta}_b^*$ obtidos pela amostragem *bootstrap* paramétrica. Gráfico da direita: Distribuição $\hat{\beta}_b^*$ obtidos pela amostragem *bootstrap* não paramétrica.



Fonte: Elaborado pelo autor (2022)

que foram apresentadas no Capítulo 5 e que nos ajudarão a entender e avaliar as estimativas intervalares. Exclusivamente para o cálculo da cobertura, simularemos $M = 200$ ensaios de

Monte Carlo para reproduzir 200 intervalos diferentes, e assim estimaremos a cobertura dos intervalos de confiança obtidos através de cada método aplicado neste experimento.

Para a leitura das tabelas com as informações e resultados dos intervalos de confiança gerados, assumimos que LI e LS representam os limites inferiores e superiores de um intervalo de confiança (denotado por IC), respectivamente. Já os termos (P) e (NP) representam se o IC referente a cada método aplicado foi construído realizando a amostragem *bootstrap* paramétrica ou não paramétrica.

Através da Tabela 1, podemos visualizar os intervalos de 95% de confiança construídos para γ de acordo com diferentes métodos, e alguns valores que nos ajudam a avaliar o comportamento de cada estimativa intervalar.

Tabela 1 – Intervalos de 95% de Confiança para γ

IC Bootstrap	LI	LS	Amplitude	Forma	Cobertura
$BC_a(P)$	5.98	14.9	8.92	1.11	95.5%
$BC_a(NP)$	7.21	16.7	9.49	2.18	91.5%
$Percentil(P)$	7.36	13.4	6.03	1.12	93%
$Percentil(NP)$	8.14	15.1	6.96	2.38	89.5%

Fonte: Elaborada pelo autor (2022)

Tabela 2 – Intervalos de 95% de Confiança para β

IC Bootstrap	LI	LS	Amplitude	Forma	Cobertura
$BC_a(P)$	0.309	0.746	0.437	2.016	96.5%
$BC_a(NP)$	0.279	0.657	0.378	1.157	91%
$Percentil(P)$	0.327	0.628	0.300	1.373	93%
$Percentil(NP)$	0.295	0.583	0.288	0.809	88%

Fonte: Elaborada pelo autor (2022)

Ao observar as Tabelas 1 e 2, percebemos que os intervalos de confiança obtidos pelo método BC_a possuem maior amplitude se comparados com o *bootstrap* percentil. Lembrando que a amostra original considerada neste experimento era de tamanho $n = 40$, que é uma amostra de tamanho moderado. Para valores maiores de n, esperaríamos intervalos com amplitude menor. Contudo, a cobertura dos intervalos BC_a foi superior, e por isso, confere uma maior confiabilidade para estas estimativas intervalares. Como já discutido no Capítulo 5, este resultado prático em específico já era esperado para o intervalo BC_a. Pela própria definição,

o intervalo BC_a é mais acurado do que o *bootstrap* percentil, principalmente para dados assimétricos.

Em relação a assimetria, o valor de Forma obtido para quase todos intervalos indicam uma assimetria positiva do intervalo em relação às estimativas de máxima verossimilhança $\hat{\gamma}$ e $\hat{\beta}$. Apenas o intervalo percentil para β, construído através da amostragem *bootstrap* não paramétrica, apresentou uma leve assimetria negativa, isto é, o limite inferior do intervalo está mais próximo de $\hat{\beta}$ do que o superior.

Um ponto interessante que vale ressaltar é sobre a propriedade da preservação de amplitude destes intervalos, já que tanto o *bootstrap* percentil como o BC_a possuem essa qualidade. Para todos os resultados indicados nas Tabelas 1 e 2, observamos que, de fato, os intervalos estão incluindo apenas valores possíveis de serem assumidos por γ e β, isto é, $\gamma, \beta > 0$.

Já quando avaliamos os intervalos de confiança em relação ao método que a reamostragem *bootstrap* foi executada, constatamos que a abordagem paramétrica resultou em uma maior taxa de cobertura. Este resultado também já era esperado, uma vez que neste estudo simulado conhecíamos a distribuição de probabilidade de origem da amostra original e que assumimos esta mesma distribuição, com os parâmetros estimados pelo método da máxima verossimilhança, no procedimento gerador das amostras *bootstrap*. Porém, nem sempre é possível determinar a priori uma distribuição de probabilidade que rege uma população de interesse, neste caso a abordagem não paramétrica é uma ótima alternativa por ser livre de suposições e, ainda assim, entregar intervalos de confiança satisfatórios para um parâmetro qualquer θ.

7 CONSIDERAÇÕES FINAIS

Através do experimento realizado com dados simulados, percebemos que os métodos *bootstrap* possibilitam mensurar a incerteza de um estimador de maneira simples e eficaz. Através de um processo computacional intensivo, simulamos novas amostras que nos agregam informação sobre a distribuição daquele estimador.

De uma forma geral, as distribuições *bootstrap* se aproximam das distribuições amostrais dos estimadores, porém elas estão centradas nas estimativas obtidas através da amostra original, enquanto a distribuição amostral está centrada no verdadeiro valor do parâmetro.

Apesar de ser um algoritmo custoso, por necessitar simular novos dados e calcular novas estimativas um número excessivo de vezes, os intervalos de confiança *bootstrap* trazem resultados satisfatórios e são bastante confiáveis. A abordagem de reamostragem paramétrica levou vantagem neste experimento sobre a não paramétrica, mas isto aconteceu pois conhecíamos a distribuição que originou os dados iniciais. Em muitos problemas reais não conseguimos assumir uma distribuição a priori, logo a amostragem *bootstrap* não paramétrica é tão importante por gerar resultados ainda muito bons para qualquer distribuição original dos dados.

O método BC_a se mostrou uma melhor opção do que o *bootstrap* percentil devido a sua forma de construção, que inclui fatores de correção do viés e de aceleração. Isto confere uma maior confiabilidade, principalmente quando lidamos com dados assimétricos. A escolha do tamanho moderado da amostra original ($n = 40$) foi motivado pelo desejo de validar os resultados dos métodos *bootstrap* com uma quantidade inicial não muito grande de dados disponíveis, e de fato, mostramos o valor das técnicas *bootstrap*.

7.1 TRABALHOS FUTUROS

Para trabalhos futuros envolvendo as técnicas *bootstrap*, trabalhar com dados reais seria interessante. Além disso, o estudo sobre estimadores pontuais corrigidos através das amostras *bootstrap*, definidos em Efron e Tibshirani (1993), é extremamente útil por possibilitar a redução do viés de estimadores originalmente viesados.

Uma outra iniciativa seria construir outros tipos de intervalo de confiança *bootstrap* e, junto com o que já foi desenvolvido, disponibilizar todo o código através de uma biblioteca no *R*. Com isto, teríamos uma biblioteca capaz de gerar diversos intervalos de confiança utilizando

o método *bootstrap* e, para cada intervalo, calcular as métricas de avaliação já implementadas e apresentadas neste trabalho. O grande desafio seria aprimorar os algoritmos desenvolvidos para reduzir o custo computacional das execuções.

REFERÊNCIAS

ALVES, E. J. *Métodos de bootstrap e aplicações em problemas biológicos*. Dissertação (Mestrado) — Universidade Estadual Paulista "Júlio de Mesquita Filho", 2013. Disponível em: <https://repositorio.unesp.br/bitstream/handle/11449/94336/alves_ej_me_rcla.pdf?sequence=1>.

BUSSAB, W. de O.; MORETTIN, P. A. *Estatística Básica*. 8. ed. São Paulo: Editora Saraiva, 2013.

DALITZ, C. *Construction of Confidence Intervals*. 2017. Disponível em: <https://lionel.kr.hsnr.de/~dalitz/data/publications/fb03-tb-2017-01-en.pdf>. Acesso em: 30 ago. 2022.

DORE, L. H. G.; AMARAL, G. J. A.; CRUZ, J. T. M.; WOOD, A. T. A. Bias-corrected maximum likelihood estimation of the parameters of the complex bingham distribution. *Brazilian Journal of Probability and Statistics*, Brazilian Statistical Association, v. 30, n. 3, p. 385–400, 2016.

EFRON, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, 1979.

EFRON, B.; TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. Nova Iorque: Chapman Hall, Inc, 1993.

FISHER, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, The Royal Society, v. 222, p. 309–368, 1922.

HOEL, P. G.; PORT, S. C.; STONE, C. J. *Introduction to Statistical Theory*. Boston: Houghton Mifflin Company, 1971.

LOBÃO, D. C. *Introdução aos Métodos Numéricos*. 2017. Disponível em: <https://www.professores.uff.br/diomarcesarlobao/wp-content/uploads/sites/85/2017/09/note6.pdf>.

LY, A.; MARSMAN, M.; VERHAGEN, J.; GRASMAN, R.; WAGENMAKERS, E.-J. *A Tutorial on Fisher Information*. 2017. 1-59 p. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0022249617301396>.

MAGALHÃES, M. N. *Probabilidade e Variáveis Aleatórias*. 2. ed. São Paulo: Editora da Universidade de São Paulo, 2006.

PAULA, G. A. *MODELOS DE REGRESSÃO: com apoio computacional*. 2013. Disponível em: <https://www.ime.usp.br/~giapaula/texto_2013.pdf>. Acesso em: 10 ago. 2022.

SHANG, X. *ESTIMATION OF PARAMETERS OF GAMMA AND GENERALIZED GAMMA DISTRIBUTIONS BASED ON CENSORED EXPERIMENTAL DATA*. Tese (Doutorado) — Southern Methodist University, 2021. Disponível em: <https://scholar.smu.edu/hum_sci_statisticalscience_etds/24/?utm_source=scholar.smu.edu\%2Fhum_sci_statisticalscience_etds%2F24&utm_medium=PDF&utm_campaign=PDFCoverPages>.

THOM, H. C. S. A note on the gamma distribution. *Monthly Weather Review*, v. 86, n. 4, p. 117–122, 1958.

APÊNDICE A – ALGORITMOS DESENVOLVIDOS NO *R*

```
## Configuração do experimento simulado:
library(maxLik)
library(latex2exp)

set.seed(5) #para reprodução do experimento com mesmos resultados
B = 1000
m = 200
n = 40
shape = 9
scale = 0.5
alpha = 0.025

## log-verossimilhança da distribuição gama:
gammall = function(theta,data){
  value = sum(dgamma(data,shape=theta[1],scale=theta[2],log=T))
  return(value)
}

## Cálculo dos estimadores de máxima verossimilhança da gama:
mlegamma = function(x) {
  momgamma = mean(x)^2/var(x)
  mombeta = var(x)/mean(x)
  gamma = maxLik(logLik=gammall,start=c(momgamma,mombeta),data=x)
  shape_est = as.numeric(gamma$estimate[1])
  scale_est = as.numeric(gamma$estimate[2])
  result = c(shape_est, scale_est)
  return(result)
}

## Amostragem bootstrap paramétrica:
boot_p = function(shape_hat,scale_hat,B){
  shape_pboot = c()
  scale_pboot = c()
  for (i in 1:B){
    x = rgamma(n=n, shape=shape_hat, scale=scale_hat)
    result = mlegamma(x)
    shape_pboot = c(shape_pboot,result[1])
    scale_pboot = c(scale_pboot,result[2])
  }
  return (matrix(c(shape_pboot,scale_pboot),nrow=B,ncol=2))
}

## Amostragem bootstrap não paramétrica:
boot_np = function(data,B){
  shape_npboot = c()
  scale_npboot = c()
  for (i in 1:B){
    x = sample(data,n,replace=TRUE)
    result = mlegamma(x)

    shape_npboot = c(shape_npboot,result[1])
    scale_npboot = c(scale_npboot,result[2])
  }
  return (matrix(c(shape_npboot,scale_npboot),nrow=B,ncol=2))
}

## Cálculo do fator hat(a):
a_hat_calc = function(original_sample) {
  data_aux = original_sample
  shp = c()
  scle = c()
  #principio do método jackknife:
  for (i in 1:length(original_sample)){
    data_aux = data_aux[-i]
    fit = mlegamma(data_aux)
    shape_jk_i = fit[1]
    scale_jk_i = fit[2]
```

```
    shp = c(shp,shape_jk_i)
    scle = c(scle,scale_jk_i)

    data_aux = original_sample
  }

  shp_dot = mean(shp)
  scle_dot = mean(scle)

  a_hat_shp = sum( (shp_dot - shp)^3 )/ (6*( (sum( (shp_dot - shp)^2 ))^(3/2) ))
  a_hat_scle = sum( (scle_dot - scle)^3 )/ (6*( (sum( (scle_dot - scle)^2 ))^(3/2) ))
  return (c(a_hat_shp,a_hat_scle))
}

##Cálculo do fator hat(z_0):
z0_hat_calc = function(shape_boot,shape_orig,scale_boot,scale_orig) {
  phi_shape = qnorm(sum((shape_boot < shape_orig)*1)/B)
  phi_scale = qnorm(sum((scale_boot < scale_orig)*1)/B)
  return (c(phi_shape,phi_scale))
}

##Intervalo de Confiança BCa:
ci_bca = function(original_sample,shape_boot,shape_hat,scale_boot,scale_hat,alpha) {
  a = a_hat_calc(original_sample)
  a_shape = a[1]
  a_scale = a[2]

  z0 = z0_hat_calc(shape_boot,shape_hat,scale_boot,scale_hat)
  z0_shape = z0[1]
  z0_scale = z0[2]


  z_alpha = qnorm(alpha)
  z_alphacomp = qnorm(1-alpha)

  alpha1_shape = pnorm(z0_shape + ((z0_shape+z_alpha)/(1 - a_shape*(z0_shape + z_alpha))))
  alpha2_shape = pnorm(z0_shape + ((z0_shape+z_alphacomp)/(1 - a_shape*(z0_shape + z_alphacomp))))

  alpha1_scale = pnorm(z0_scale + ((z0_scale+z_alpha)/(1 - a_scale*(z0_scale + z_alpha))))
  alpha2_scale = pnorm(z0_scale + ((z0_scale+z_alphacomp)/(1 - a_scale*(z0_scale + z_alphacomp))))


  shape_bca_lo = quantile(shape_boot, probs=c(alpha1_shape), names=FALSE)
  shape_bca_up = quantile(shape_boot, probs=c(alpha2_shape), names=FALSE)

  scale_bca_lo = quantile(scale_boot, probs=c(alpha1_scale), names=FALSE)
  scale_bca_up = quantile(scale_boot, probs=c(alpha2_scale), names=FALSE)

  return(c(shape_bca_lo,shape_bca_up,scale_bca_lo,scale_bca_up))
}

##Intervalo de Confiança bootstrap percentil:
ci_bperc = function(shape_boot,scale_boot,alpha) {

  shape_bperc = quantile(shape_boot, probs=c(alpha,(1-alpha)), names=FALSE)
  shape_bperc_up = shape_bperc[2]
  shape_bperc_lo = shape_bperc[1]

  scale_bperc = quantile(scale_boot, probs=c(alpha,(1-alpha)), names=FALSE)
  scale_bperc_up = scale_bperc[2]
  scale_bperc_lo = scale_bperc[1]

  return(c(shape_bperc_lo,shape_bperc_up,scale_bperc_lo,scale_bperc_up))
}
```

```
##Variáveis auxiliares para construção de tabelas e gráficos avaliando o experimento:
cobertura_shape = c(0,0,0,0)
cobertura_scale = c(0,0,0,0)
se_param = c()
se_nonparam = c()
bias_bootparam = c()
bias_bootnonparam = c()

shp_hat = 0
scle_hat = 0

hist_shape_param = c()
hist_shape_nonparam = c()
hist_scale_param = c()
hist_scale_nonparam = c()

shape_df_fst = data.frame(LO=rep(NA,4),UP=rep(NA,4))
scale_df_fst = data.frame(LO=rep(NA,4),UP=rep(NA,4))

## m intervalos de confiança sendo gerados para estimar a cobertura:
for(t in 1:m) {

myData = rgamma(n=n, shape=shape, scale=scale)
fit = mlegamma(myData)
shape_hat = fit[1]
scale_hat = fit[2]

shape_df = data.frame(LO=rep(NA,4),UP=rep(NA,4))
scale_df = data.frame(LO=rep(NA,4),UP=rep(NA,4))

# Fazendo intervalo BCA paramétrico:
boot_estimators = boot_p(shape_hat,scale_hat,B)
ci_limits = ci_bca(myData,boot_estimators[,1],shape_hat,boot_estimators[,2],scale_hat,alpha)
shape_df[1,] = ci_limits[1:2]
scale_df[1,] = ci_limits[3:4]

# Fazendo intervalo percentil paramétrico:
ci_limits = ci_bperc(boot_estimators[,1],boot_estimators[,2],alpha)
shape_df[2,] = ci_limits[1:2]
scale_df[2,] = ci_limits[3:4]

# Fazendo intervalo BCA NP:
boot_estimators_np = boot_np(myData,B)
ci_limits = ci_bca(myData,boot_estimators_np[,1],shape_hat,boot_estimators_np[,2],scale_hat,alpha)
shape_df[3,] = ci_limits[1:2]
scale_df[3,] = ci_limits[3:4]

# Fazendo intervalo percentil NP:
ci_limits = ci_bperc(boot_estimators_np[,1],boot_estimators_np[,2],alpha)
shape_df[4,] = ci_limits[1:2]
scale_df[4,] = ci_limits[3:4]

# Gerando métricas para a primeira iteração:
if(t == 1) {
  se_param = sd(boot_estimators)
  se_nonparam = sd(boot_estimators_np)
  bias_bootparam = c(mean(boot_estimators[,1]),mean(boot_estimators[,2])) - c(shape_hat,scale_hat)
  bias_bootnonparam = c(mean(boot_estimators_np[,1]),mean(boot_estimators_np[,2])) -
 c(shape_hat,scale_hat)
  hist_shape_param = boot_estimators[,1]
  hist_shape_nonparam = boot_estimators_np[,1]
  hist_scale_param = boot_estimators[,2]
  hist_scale_nonparam = boot_estimators_np[,2]

  shape_df_fst = shape_df
  scale_df_fst = scale_df
```

```
  shp_hat = shape_hat
  scle_hat = scale_hat
}

# Possibilitando cálculo cobertura através de:
cobertura_shape = cobertura_shape + (shape_df$LO<=shape & shape_df$UP>=shape)*1
cobertura_scale = cobertura_scale + (scale_df$LO<=scale & scale_df$UP>=scale)*1
}


## Histogramas ##
par(mfrow=c(1,2))
hist(hist_shape_param,xlab=TeX(r'(\hat{\gamma}*)'),ylab="Frequência",main="",ylim=c(0,350),cex.axis =
 1.3, cex.lab = 1.5)
hist(hist_shape_nonparam,xlab=TeX(r'(\hat{\gamma}*)'),ylab="Frequência",main="",ylim=c(0,250),cex.axis =
 1.3, cex.lab = 1.5)

par(mfrow=c(1,2))
hist(hist_scale_param,xlab=TeX(r'(\hat{\beta}*)'),ylab="Frequência",main="",cex.axis = 1.3, cex.lab =
 1.5,col="404040")
hist(hist_scale_nonparam,xlab=TeX(r'(\hat{\beta}*)'),ylab="Frequência",main="",cex.axis = 1.3, cex.lab =
 1.5,col="404040")

## Cálculos tabelas IC##

shape_df_fst_c = shape_df_fst
shape_df_fst$amp = shape_df_fst$UP - shape_df_fst$LO
shape_df_fst$forma = (shape_df_fst$UP - shp_hat)/(shp_hat - shape_df_fst$LO)
options(digits=3)
cobertura_shape/m
shape_df_fst
#
scale_df_fst_c = scale_df_fst
scale_df_fst$amp = scale_df_fst$UP - scale_df_fst$LO
scale_df_fst$forma = (scale_df_fst$UP - scle_hat)/(scle_hat - scale_df_fst$LO)
options(digits=3)
scale_df_fst
cobertura_scale/m
```