Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Campus de Rio Claro

Métodos de *bootstrap* e aplicações em problemas biológicos

Edmar José Alves

Dissertação apresentada ao Programa de Pós-Graduação – Mestrado Profissional em Matemática Universitária como requisito parcial para a obtenção do grau de Mestre

Orientadora
Profa. Dra. Selene Maria Coelho Loibel

2013

TERMO DE APROVAÇÃO

Edmar José Alves

MÉTODOS DE *bootstrap* E APLICAÇÕES EM PROBLEMAS BIOLÓGICOS

Dissertação APROVADA como requisito parcial para a obtenção do grau de Mestre no Curso de Pós-Graduação Mestrado Profissional em Matemática Universitária do Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", pela seguinte banca examinadora:

Profa. Dra. Selene Maria Coelho Loibel
Orientadora

Prof. Dr. José Silvio Govone
Departamento de Estatística, Matemática Aplicada e Computação - IGCE, Rio Claro

Prof. Dra. Rita de Cássia Pavani Lamas
Departamento de Matemática - IBILCE, São José do Rio Preto

Rio Claro, 25 de outubro de 2013

Dedicado a meus pais,
Irene Faccioni Alves e Benedito Vicenti Alves.

Agradecimentos

Agradeço primeiramente a Deus, pela força que me deu durante o percurso de mais uma etapa vencida com sucesso na minha vida. Escrever os agradecimentos não foi simples como imaginei, pois foram inúmeras pessoas que fizeram parte desta etapa. Quero agradecer a todas essas pessoas que se fizeram presentes e me impulsionaram para esta grande conquista. Se não fosse minha família, não estaria escrevendo estes agradecimentos. Minha mãe, amor eterno, Irene Faccioni Alves, o meu tudo, que deu a vida por mim para me ver feliz, sorrindo e realizado. Do pé de café ao sonho do mestrado, pela suas orações, palavras de conforto e fornecendo forças para continuar sempre. Meu pai, Benedito Vicenti Alves, que mesmo exausto do trabalho, madrugava para levar ao destino de embarque. Minha irmã, Edilene, e meu cunhado, que mesmo sendo corintiano é gente boa, sempre estão dispostos a ajudar no que for necessário. Claro que não posso esquecer da turma de São José do Rio Preto, dentre eles, Thaisa, Franciélli, Wellington, Erica, Juliano, Renato, Dênis, João Paulo (filme do galango), Vinicíus, Mariana, Aline, Edilson, Débora, Amarilis, Williner, Felippe, quantas madrugadas na estrada! E apuros então nem se fala. A esses amigos, devo o meu muito obrigado de coração, pelo apoio, pelas risadas, pela ajuda, pois mesmo com as dificuldades devido a distância e também ao nosso trabalho e estudos, tornaram esses momentos mais felizes e alegres, fazendo com que o trecho de Rio Preto a Rio Claro ficasse menos cansativo do que realmente é. Também destaco aos novos amigos conquistados aqui em Rio Claro, Mariana, Leandro, Carlos e Edgar (meu parceiro de dupla sertaneja) e os amigos que vieram diretamente do nordeste, como o Olívio e o Antonio Nilson. A todos os membros do Departamento de Matemática e do Departamento de Estatística, Matemática Aplicada e Computacional (DEMAC), que de forma direta ou não, contribuíram para a implementação do curso de mestrado. Quero agradecer aos pesquisadores Célio Haddad e Eliziane Garcia de Oliveira que forneceram o dados reais a este trabalho. Deixo aqui, profunda gratidão a minha orientadora, Profa. Dra. Selene Maria Coelho Loibel, pela dedicação dada a este trabalho, serenidade, compreenção, pela atenção com que me orientou e pelo incentivo, sempre me recebendo alegre e com muito entusiasmo nas discussões.

Cada dia que amanhece
assemelha-se a uma página em branco,
no qual gravamos nossos
pensamentos, ações e atitudes.
Na essência, cada dia é a preparação
de nosso próprio amanhã.
Chico Xavier

Resumo

As técnicas de *bootstrap* são métodos computacionais intensivos que usam reamostragem para o cálculo de medidas de incerteza dos estimadores, tais como erros-padrão, viés e intervalos de confiança. Os métodos são aplicados a qualquer nível de modelagem e assim podem ser usados tanto na análise paramétrica quanto na não paramétrica. Os métodos *bootstrap* estudados são: o intervalo de confiança *bootstrap* padrão, o intervalo de confiança *bootstrap-t*, o intervalo de confiança *bootstrap* percentil, o intervalo de confiança *bootstrap* BCPB e o intervalo de confiança BC_a. Utiliza-se dois códigos computacionais a serem implementados no software Matlab. Os códigos fornecem subsídios para a aplicação das técnicas estudadas a dados simulados e reais (problemas biológicos). Os intervalos de confiança *bootstrap* foram comparados entre si e com os métodos tradicionais de estimação da incerteza de estimadores.

Palavras-chave: Estatística, Erro Padrão, Intervalo de Confiança.

Abstract

The *bootstrap* methods are intensive computational methods that use resampling to calculate measures of uncertainty of the estimators, such as standard errors, bias and confidence intervals. The methods are applicable to any level modeling and thus can be used in both parametric analysis as the non-parametric. The *bootstrap* methods are studied: the standard *bootstrap* confidence interval, the *bootstrap-t* confidence interval, the *bootstrap* percentile confidence interval, the *bootstrap* BCBP confidence interval and the *bootstrap* BC_a confidence interval. Two computer codes are presented to be implemented in Matlab. Such codes allows the implementation of the techniques studied in simulated and real data (biological problems). The *bootstrap* confidence intervals were compared with each other and with traditional methods of estimation uncertainty estimators.

Keywords: Estatística, Standard Error, Interval Confidence.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Na inferência estatística devem ser considerados os estimadores pontuais dos parâmetros de interesse e alguma medida da incerteza destes estimadores. Pode-se citar algumas medidas de variabilidade tais como a variância, o desvio padrão e o erro médio quadrático, ou pode-se pensar na precisão (inverso da variância) e na acurácia (inverso do erro médio quadrático). Em geral, calcula-se a variância e o desvio-padrão para obter-se estimativas por intervalos para os parâmetros, fixando um coeficiente de confiança. Na maioria dos casos há a necessidade de utilizar resultados assintóticos para o cálculo destes intervalos (por exemplo, a normalidade assintótica dos estimadores de máxima verossimilhança) e muitas vezes as amostras utilizadas não são do tamanho suficiente para o uso destes resultados como por exemplo em dados biológicos . Como consequência disto pode-se obter intervalos muito amplos e, em alguns casos, em que há maior complexidade do modelo, o intervalo fora do domínio do parâmetro, fatos comuns encontrados nos problemas biológicos. Uma alternativa para esses casos é utilizar as técnicas de *bootstrap*. Pode-se utilizar também as técnicas *bootstrap* quando a distribuição da estatística de interesse não é conhecida.

As técnicas *bootstrap* introduzidas por EFRON e TIBSHIRANI [7] são métodos computacionais intensivos que usam reamostragem para calcular as medidas de incerteza dos estimadores como por exemplo: erros-padrões, viés e os intervalos de confiança. Tem como objetivo obter informações de características da distribuição de uma variável aleatória que não podem ser facilmente avaliadas por métodos analíticos tradicionais.

O método tem por base a ideia de que o pesquisador pode tratar a amostra como se ela fosse a população que deu origem aos dados e usar reamostragem com reposição da amostra original para gerar amostras *bootstrap*. A partir destas amostras *bootstrap*, é possível estimar características da população como média, variância, mediana e consequentemente encontrar o desvio padrão *bootstrap* e o viés. As vantagens da técnica *bootstrap* são inúmeras tais como não necessitar de muitas suposições para a estimação dos parâmetros das distribuições de interesse, fornecer respostas mais precisas, entendimento fácil, entre outras que veremos ao longo desse trabalho. A utilização da técnica *bootstrap* não implica que as outras técnicas devem ser deixadas de lado. Por exemplo, calculado o intervalo de confiança *bootstrap* pode-se comparar com os intervalos

de confiança assintóticos. O objetivo deste trabalho é utilizar o método *bootstrap* para determinar o erro padrão e a partir disso, construir intervalos de confiança e verificar se existe um intervalo de confiança *bootstrap* adequado a cada situação distinta. No caso dos problemas biológicos, veremos que o método *bootstrap* é uma ferramenta importante do ponto de vista prático. Por ser uma técnica computacional intensiva, há necessidade do uso de softwares específicos. Neste trabalho utliza-se o software Matlab.

O *bootstrap* e os diferentes tipos de intervalos de confiança construídos através deste método são apresentados de forma concisa. O trabalho tem por finalidade aplicar essas técnicas a dados simulados e reais (problemas biológicos), comparando-as entre si e também comparando-as com os métodos tradicionais de estimação da incerteza dos estimadores.

1.1 Descrição dos capítulos

No capítulo 2 descreve-se um pouco da história da estatística, dos primórdios a era atual, constatando sua evolução. Também é descrito o processo de reamostragem e ao final deste capítulo é caracterizado o método *bootstrap*, mencionando suas características. No capítulo 3 serão apresentadas definições básicas, importantes para o bom entendimento do trabalho, o método pioneiro de reamostragem *jackknife*, o método *bootstrap*, o estimador *bootstrap* do erro padrão juntamente com a estimativa *bootstrap* do viés e o algoritmo do mesmo. Uma breve revisão sobre intervalo de confiança assintótico é apresentada no começo do capítulo 4, em seguida é descrito os diferentes tipos de intervalos de confiança *bootstrap*, bem como as propriedades desejadas. No capítulo 5 serão mencionados exemplos de aplicação do método *bootstrap* e no capítulo 6 é descrita a aplicação do método ao problema de contagem dos indivíduos da espécie *Brachycephalus pitanga,* em atividade de vocalização.

2 Conceitos iniciais

As técnicas de *bootstrap* são procedimentos de reamostragem para o cálculo de medidas de incerteza dos estimadores. Tais técnicas foram introduzidas por EFRON e TIBSHIRANI [7] com o objetivo de obter informações de características da distribuição de uma variável aleatória, que não podem ser facilmente avaliadas por métodos analíticos tradicionais ou cuja aproximação existente tenha suposições questionáveis em alguma situação. Se comparado a outras técnicas estatísticas, o método teve seu auge um pouco tardio, devido a sua dependência do uso de computadores. Os progressos da informática, experimentados nas últimas décadas do século XX, possibilitaram a popularização do uso do computador e incrementaram o surgimento e acesso aos softwares matemáticos e estatísticos. Consequentemente, as aplicações de métodos *bootstrap* nas mais diferentes áreas da estatística se intensificaram.

Essas técnicas permitem aproximar a distribuição de uma variável aleatória pela distribuição empírica baseada em uma amostra de tamanho finito desta variável. A amostragem é feita com reposição, da distribuição da qual os dados são obtidos, se esta é conhecida (*bootstrap* paramétrico) ou da amostra original (*bootstrap* não paramétrico).

O *bootstrap* aborda o cálculo do intervalo de confiança de parâmetros em circunstâncias em que outras técnicas não são aplicáveis, em particular no caso em que a amostra é pequena e o uso de resultados assintóticos não é possível. A técnica *bootstrap* consiste em realizar o que seria desejável na prática, se tal fosse possível: repetir a experiência. As observações são escolhidas de forma aleatória e as estimativas recalculadas. Usa-se a distribuição *bootstrap* no lugar da distribuição amostral.

Com a técnica *bootstrap* obtém-se os mesmos resultados que o processo tradicional, baseado na máxima verossimilhança. A sua grande vantagem consiste em apresentar solução para casos em que a dedução da precisão da estimativa, de seu viés e do erro médio quadrático, é complexa, além de outras vantagens tais como: não necessita de muitas suposições para a estimação de parâmetros das distribuições de interesse e fornece resposta mais precisa e de entendimento fácil.

2.1 A estatística, dos primórdios a era atual

Os dados históricos apresentados a seguir foram retirados do trabalho de José Maria Pompeu Memória (MEMÓRIA), [17]. Desde a antiguidade a estatística já começava a ser empregada para investigar as propriedades e riquezas dos povos. A utilização da estatística já remonta há quatro mil anos antes de Cristo, quando era utilizada por povos guerreiros na conquista de territórios. A própria Bíblia descreve isso: "Naqueles tempos apareceu um decreto de César Augusto, ordenando o recenseamento de toda a terra." Este recenseamento foi feito antes do governo de Quirino, na Síria. Todos iam alistar-se, cada um na sua cidade (BÍBLIA , N.T. Lucas, 2:1-3). A origem da palavra Estatística está associada à palavra latina "STATUS" (Estado). Há indícios de que 4000 anos A.C. já se faziam censos na Babilônia, China e Egito e até mesmo o 4^o livro do Velho Testamento faz referência a uma instrução dada a Moisés, para que fizesse um levantamento dos homens de Israel que estivessem prontos para guerrear. Usualmente, estas informações eram utilizadas para a taxação de impostos ou para o alistamento militar. O Imperador César Augusto, por exemplo, ordenou que se fizesse o Censo de todo o Império Romano.

A palavra "CENSO" é derivada da palavra "CENSERE", que em Latim significa "TAXAR". Em 1085, Guilherme, O Conquistador, solicitou um levantamento estatístico da Inglaterra, que deveria conter informações sobre terras, proprietários, uso da terra, empregados e animais. Os resultados deste Censo foram publicados em 1086 no livro intitulado "Domesday Book" e serviram de base para o cálculo de impostos.

Contudo, mesmo que a prática de coletar impostos,dados sobre colheitas, composição da população humana ou de animais, fosse conhecida pelos egípcios, hebreus, e gregos, e se atribuam a Aristóteles cento e oitenta descrições de Estados, apenas no século XVII a Estatística passou a ser considerada disciplina autônoma, tendo como objetivo básico a descrição dos bens do Estado.

A palavra Estatística foi implantada pelo acadêmico alemão Gottfried Achenwall (1719-1772), que foi um notável continuador dos estudos de Hermann Conrig (1606-1681). Na Enciclopédia Britânica, o verbete "STATISTICS" apareceu em 1797.

Na última metade do século XIX, os alemães Helmert (1843-1917) e Wilhelm Lexis (1837-1914), o dinamarquês Thorvald Nicolai Thiele (1838-1910) e o inglês Francis Ysidro Edgeworth (1845-1926), obtiveram resultados valiosos para o desenvolvimento da Inferência Estatística, muitos dos quais só foram completamente compreendidos mais tarde. Contudo, o impulso decisivo deve-se a Karl Pearson (1857-1936), William S. Gosset (1876-1937) e, em especial, a Ronald A. Fisher (1890-1962).

Karl Pearson (1857-1936) formou-se em 1879 pela Cambridge University e inicialmente dedicou-se ao estudo da evolução de Darwin, aplicando os métodos estatísticos aos problemas biológicos relacionados com a evolução e hereditariedade.

Entre 1893 e 1912 escreveu um conjunto de 18 artigos contribuindo extremamente para o desenvolvimento da teoria da análise de regressão e do coeficiente de correlação,

bem como do teste de hipóteses de qui-quadrado. Em sua maioria, seus trabalhos foram publicados na revista Biometrica, que fundou em parceria com Walter Frank Raphael Weldon (1860-1906) e Francis Galton (1822-1911). Além da valiosa contribuição que deu para a teoria da regressão e da correlação, Pearson fez com que a Estatística fosse reconhecida como uma disciplina autônoma.

William Sealey Gosset (1876-1937) estudou Química e Matemática na New College Oxford. Em 1899 foi contratado como Químico da Cervejaria Guiness em Dublin, desenvolvendo um trabalho extremamente importante na área de Estatística. Devido à necessidade de manipular dados provenientes de pequenas amostras, extraídas para melhorar a qualidade da cerveja, Gosset obteve o teste t de Student baseado na distribuição de probabilidades.

Esses resultados foram publicados em 1908 na revista Biometrica, sob o pseudônimo de Student, dando origem a uma nova e importante fase dos estudos estatísticos. Gosset usava o pseudônimo de Student, pois a Cervejaria Guiness não desejava revelar aos concorrentes os métodos estatísticos que estava empregando no controle de qualidade da cerveja. Os estudos de Gosset podem ser encontrados em "Student Collected Papers" (Ed. por E.S.Pearson e J. Wishart, University College, Londres, 1942).

A contribuição de Ronald Aylmer Fisher (1890-1962) para a Estatística Moderna é, sem dúvidas, a mais importante e decisiva de todas. Formado em ciência natural pela Universidade de Cambridge em 1912, foi o fundador do célebre Statistical Laboratory da prestigiosa Estação Agronômica de Rothamsted, contribuindo enormemente tanto para o desenvolvimento da Estatística quanto da Genética. Ele apresentou os princípios de planejamento de experimentos, introduzindo os conceitos de aleatorização e da análise da variância, procedimentos muito usados atualmente.

No princípio dos anos 20, estabeleceu o que a maioria aceita como a estrutura da moderna Estatística Analítica, através do conceito da verossimilhança (likelihood, em inglês). O seu livro intitulado "Statistical Methods for Research Workers", publicado pela primeira vez em 1925, foi extremamente importante para familiarizar os investigadores com as aplicações práticas dos métodos estatísticos e também para criar a mentalidade estatística entre a nova geração de cientistas. Os trabalhos de Fisher encontram-se dispersos em numerosas revistas, mas suas contribuições mais importantes foram reunidas em "Contributions to Mathematical Statistics" (J. Wiley & Sons, Inc., Nova Iorque, 1950).

Fisher foi eleito membro da Royal Society em 1929 e condecorado com as medalhas Royal Medal of the Society e Darwin Medal of the Society em 1938 e 1948, respectivamente. Em 1955 foi novamente condecorado, desta vez com a medalha Copley Medal of the Royal Society.

Outra área de investigação extremamente importante para o desenvolvimento da Estatística é a Teoria das Probabilidades. Usualmente, costuma-se atribuir a origem do Cálculo de Probabilidades às questões relacionadas aos jogos de azar que o célebre

cavaleiro Méré (1607-1684) encaminhou à Blaise Pascal (1623-1662).

No entanto, outros autores sustentam que o Cálculo de Probabilidades teve a sua origem na Itália, com especial referência para Luca Pacioli (1445-1517), Girolamo Cardano (1501-1576), Nicolo Fontana Tartaglia (1500-1557) e Galileo Galilei (1564-1642).

Três anos depois de Pascal ter previsto que a "aliança do rigor geométrico" com a "incerteza do azar" daria lugar a uma nova ciência, Christiaan Huygens (1629-1695) publicou o trabalho denominado "De Raciociciis in Ludo Aleae", que é considerado o primeiro livro sobre o Cálculo de Probabilidades. Além disso, ainda teve a notável particularidade de introduzir o conceito de esperança matemática.

Gottfried Wilhelm von Leibniz (1646-1716) também dedicou-se ao estudo do Cálculo de Probabilidades, publicando um trabalho sobre a "arte combinatória" e outro sobre aplicações às questões financeiras. Leibniz também estimulou Jacques Bernoulli (1654-1705) ao estudo do Cálculo de Probabilidades, cuja grande obra, denominada "Ars Conjectandi", foi publicada oito anos após a sua morte.

Em "Ars Conjectandi de Jacques Bernoulli", foi publicada e rigorosamente provada a Lei dos Grandes Números de Bernoulli, considerada o primeiro teorema limite. Pode-se dizer que graças às contribuições de Bernoulli o Cálculo de Probabilidades adquiriu o status de ciência.

Além da obra póstuma de Bernoulli, o início do século XVIII foi marcado pelos livros de Pierre Rémond de Montmort (1678-1719), denominado "Essai d'Analyse sur les Jeux de Hazard", e de Abraham De Moivre (1667-1754), intitulado "The Doctrine of Chances".

De Moivre era francês de nascimento, mas desde a sua infância refugiou-se na Inglaterra devido às guerras religiosas, fazendo aplicações ao cálculo de anuidades e estabelecendo uma equação simples para a lei da mortalidade entre 22 anos e o limite da longevidade que fixou em 86 anos. Mais tarde, na "Miscellanea Analytica", apresentou resultados aos quais Laplace deu uma forma mais geral e que constituem o segundo teorema limite.

Os estudos dos astrônomos Pierre-Simon Laplace (1749-1827), Johann Carl Friedrich Gauss (1777-1855) e Lambert Adolphe Jacques Quetelet (1796-1874) foram fundamentais para o desenvolvimento do Cálculo de Probabilidades. Devido aos novos métodos e ideias, o trabalho de Laplace de 1812, intitulado "Théorie Analytique des Probabilités", até o presente é considerado um dos mais importantes trabalhos sobre a matéria.

Johann Carl Friedrich Gauss, professor de astronomia e diretor do Observatório de Gottingen, em 1809 apresentou o estudo intitulado "Theoria combinationis Observatorium Erroribus Minimis Obnoxia", explanando uma teoria sobre a análise de observações aplicável a qualquer ramo da ciência, alargando o campo de aplicação do Cálculo de Probabilidades.

Com Lambert, Adolphe Jacques Quetelet, por sua vez, inicia-se a aplicação aos fenômenos sociais. O seu escrito "Sur l'homme et le développement de ses facultés" foi

publicado em segunda edição com o título "Physique sociale ou Essai sur le développement des facultés de l'homme", que incluía pormenorizada análise da teoria da probabilidade. Quetelet introduziu também o conceito de "homem médio" e chamou particular atenção para a notável consistência dos fenômenos sociais. Por exemplo, mostrou que fatores como a criminalidade apresentam permanências em relação a diferentes países e classes sociais.

Antoine Augustin Cournot (1801-1877) percebeu a importância da Teoria das probabilidades na análise estatística, tendo sido o pioneiro no tratamento matemático dos fenômenos econômicos. Suas ideias foram publicadas em "Exposition de la théorie des chances et des probabilités".

Na segunda metade do século XIX a Teoria das Probabilidades atingiu um dos pontos mais altos com os trabalhos da escola russa fundada por Pafnuty Lvovich Chebyshev (1821-1894), que contou com representantes como Andrei Andreyevich Markov (1856-1922) e Aleksandr Mikhailovich Lyapunov (1857-1918).

Contudo, o seu maior expoente foi Andrey Nikolayevich Kolmogorov (1903-1987), a quem se deve um estudo indispensável sobre os fundamentos da Teoria das Probabilidades, denominado "Grundbegrife der Warscheinlichkeitrechnung", publicado em 1933. Em 1950 foi traduzido para o Inglês sob o título "Foundations of Probability".

Ainda seguindo os dados históricos, retirados do trabalho de MEMÓRIA [17], na era atual houve um aumento crescente do uso dos computadores. Segundo Cox (1997) essa época é considerada a época áurea do pensamento estatístico. Neste período apresentou-se grandes trabalhos sobre inferência com Fisher, Neyman, Egon, Pearson e Wald, foi um período de consolidação dos trabalhos anteriores. Para Cox(1997) a primeira revolução da estatística surgiu com a introdução das máquinas de calcular.

Os "cérebros eletrônicos" como foram chamados inicialmente os computadores tem feito verdadeiras maravilhas, a ponto dos entusiastas da Inteligência Artificial acreditarem que com o tempo, será possível duplicar qualquer atividade da mente humana, já que esta é também uma máquina. A Estatística moderna é uma tecnologia quantitativa para a ciência experimental e observacional que permite avaliar e estudar as incertezas e os seus efeitos no planejamento e interpretação de experiências e de observações de fenômenos da natureza e da sociedade.

Pode-se afirmar que os métodos estatísticos foram desenvolvidos como uma mistura de ciência, tecnologia e lógica para a solução e investigação de problemas em várias áreas do conhecimento humano. Para RAMOS [21], a Estatística é uma ciência multidisciplinar que abrange praticamente todas as áreas do conhecimento humano. Podem fazer análises e utilizar de resultados estatísticos um economista, agrônomo, químico, geólogo, matemático, biólogo, sociólogo, psicólogo, cientista, políticos entre outros. Neste sentido, gráficos e tabelas são apresentados na exposição de resultados de uma empresa. Dados numéricos são usados para aprimorar e aumentar a produção. Censos demográficos ajudam o Governo a entender melhor as necessidades de sua população

e a organizar seus gastos com saúde e assistência social.

A chegada de computadores cada vez mais avançados foi decisiva e fez com que a estatística se tornasse mais acessível aos pesquisadores nos diferentes campos de atuação.

Atualmente, os softwares permitem a manipulação de grande quantidade de dados, o que veio a facilitar o emprego dos métodos estatísticos. A utilização de técnicas estatísticas vem crescendo cada vez mais, devido principalmente a sua utilidade na comparação de serviços, análise para desenvolvimento de produtos, verificação de qualidade entre outras. Por exemplo, censos demográficos auxiliam o governo a entender melhor sua população e a organizar seus gastos com saúde, educação, saneamento básico e infraestrutura. Com a velocidade da informação, a estatística passou a ser uma ferramenta essencial na produção do conhecimento. Segundo RAO E WU [22], a estatística é uma ciência que investiga o levantamento de dados, com a máxima quantidade de informação possível para um dado custo; o processamento de dados para a quantificação da quantidade de incerteza sob o menor risco possível. De fato a estatística tem sido utilizada na pesquisa científica para a otimização de recursos econômicos, para o aumento da qualidade e produtividade, na otimização em análise de decisões, em questões judiciais, previsões e em muitas outras áreas. A estatística tem ampliado a sua participação na linguagem, atividades profissionais da atualidade, considerando que os números e respectivos significados traduzem as questões do cotidiano, possibilitando a análise com base em fatos e dados.

Com relação ao método *bootstrap*, não se pode estudá-lo sem antes mencionar o método *jackknife*, que é o primeiro método de reamostragem. A ideia de reamostragem surgiu em 1935 e consiste em um conjunto de técnicas ou métodos que se baseiam em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra. Não usa a distribuição de probabilidades, no entanto calcula-se uma distribuição rotineira de estatísticas estimadas, ou seja, tendo a amostra original criamos várias amostras e com a ajuda do computador, estimamos um valor de uma estatística para cada amostra.

A metodologia *jackknife* é um complemento aos trabalhos pioneiros de QUENOUILLE [20], 1949, que introduziu um método de estimação do viés de um estimador baseado na divisão da amostra em duas subamostras. Esse método se popularizou em 1958 com TUKEY [27], que propôs a sua utilização na construção de estimadores da variância, nomeando de *jackknife*, também chamado de "leave-one-out".

Em 1972 SCHUCANY, GRAY e OWEN [25] em seus trabalhos generalizaram a técnica de *Jakknife* na redução do viés de um estimador.

BABU e SINGH [1], em 1983, obtiveram a consistência do estimador *bootstrap* da variância para qualquer amostra.

EFRON e TIBSHIRANI [7] em 1983, estenderam a técnica *bootstrap* para a construção de intervalos de confiança.

HALL [11], em 1986, sugeriu o intervalo de confiança *bootstrap*-t. Este caso, como

veremos, funciona bem quando sabemos que a distribuição da estatística na distribuição *bootstrap* é aproximadamente normal e a estatística apresenta viés pequeno.

O *bootstrap*, introduzido fortemente por EFRON e TIBSHIRANI [7] em 1979, vem historicamente na linha do *jackknife* e pode-se dizer que é uma técnica não paramétrica que procura substituir a análise estatística teórica pelo uso da computação, cada vez mais acessível nos dias de hoje. É considerada uma estratégia mais abrangente que o *jackknife*, pois permite um maior número de replicações, como veremos ao longo do trabalho.

2.2 O processo de reamostragem

A reamostragem é o nome que se dá a um conjunto de técnicas ou métodos que se baseiam em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra (única). Ou seja, consiste em sortear com reposição dados pertencentes a uma amostra retirada anteriormente, de modo a formar uma nova amostra. Surgiu em meados de 1935, entretanto a aplicação de tais técnicas se desenvolveu mais nos últimos anos, pois com o avanço tecnológico os softwares estão mais práticos, rápidos e acessíveis, uma vez que os procedimentos de reamostragem utilizam o computador de forma intensiva. Segundo DAVISON E HINKLEY [5], repetir um procedimento de análise original com muitas réplicas de dados é denominado método computacional intensivo. Criando várias amostras da amostra original, a reamostragem precisa apenas do poder computacional para estimar um valor de uma estatística para cada amostra.

Uma diferença chave entre os vários métodos de reamostragem é se as amostras são extraídas com ou sem reposição. A amostragem com reposição obtém uma observação a partir da amostra e então a coloca de volta na amostra para possivelmente ser usada novamente. A amostragem sem reposição obtém observações da amostra, mas uma vez obtidas as observações não estão mais disponíveis. O verdadeiro poder da reamostragem vem de amostragem com reposição. Pesquisas têm mostrado que esse método fornece estimativas diretas dos intervalos de confiança.

Existem diversas técnicas de reamostragem que visam estimar parâmetros de uma distribuição de interesse. Nesta dissertação, apresenta-se a técnica de reamostragem *bootstrap*, descrevendo também a técnica *jackknife*.

2.3 Características do *bootstrap*

O *bootstrap* é uma técnica estatística, computacional intensiva de reamostragem, introduzida por EFRON e TIBSHIRANI [7] em 1979, e tem como finalidade obter informações de características da distribuição de alguma variável aleatória. Para isto, aproxima-se uma distribuição de probabilidade através de uma função empírica obtida de uma amostra finita. Normalmente, esta técnica é empregada quando a distribuição

de interesse é de difícil, ou até impossível, avaliação analítica ou quando só a teoria assintótica está disponível. O termo *bootstrap* vem do inglês "to pull oneself up by one's *bootstrap*" e é oriundo da ideia de que é possível emergir de um afogamento puxando pelo cadarço do próprio sapato. Em um contexto estatístico, essa frase transmite a ideia de que é possível obter propriedades de grandes amostras a partir de poucas observações. Esta frase pode ser reescrita ainda com outras palavras: em situações de dificuldade tentar realizar o impossível. Na estatística as "situações de dificuldades" podem ser vistas como os problemas complexos e o "impossível" é a utilização de uma metodologia que pode precisar de uma grande quantidade de esforço e cálculos, mas fornece solução. A ideia básica deste método é estimar características desejadas reproduzindo-se o mecanismo probabilístico gerador dos dados originais, com a distribuição de probabilidade desconhecida destes dados sendo substituída por uma outra conhecida que possa aproximá-la. Em geral a distribuição de probabilidade da estatística de interesse não é conhecida. O *bootstrap* vem contribuir neste caso pois: não exige diferentes fórmulas para cada probabilidade, pode ser utilizado em casos gerais que não dependem da distribuição original da estatística do parâmetro. É claro que não se deve deixar as outras técnicas de lado, elas podem ser usadas na argumentação das conclusões obtidas. Segundo EFRON e TIBSHIRANI [7], o *bootstrap* é um procedimento robusto de simulação estatística para atribuir medidas de precisão a estimativas de parâmetros estatísticos. Um dos atrativos deste procedimento é encontrar medidas de precisão sem termos que lançar mãos de fórmulas matemáticas complexas. A técnica *bootstrap* envolve soluções de problemas complexos, é útil pois não necessita de muitas suposições para estimar parâmetros de interesse das distribuições, geralmente fornece respostas mais precisas e de fácil entendimento e implementação. Através dela, os parâmetros como a média, a variância e até mesmo os parâmetros como o máximo e mínimo que são menos utilizados de uma população, podem ser estimados pontualmente e por intervalo. Segundo MANTEIGA, SÁNCHES e ROMO [15], as vantagens da técnica *bootstrap* são inúmeras, tais como: obtenção de bons intervalos de confiança, entendimento fácil, não necessita de muitas suposições para estimação de parâmetros das distribuições de interesse.

Dada uma estimativa de um determinado parâmetro calculado a partir de uma amostra de dados, dois dos objetivos principais do *bootstrap* são: estimar o erro padrão da referida estimativa e estimar o intervalo de confiança apropriado.

Há diversos métodos distintos para o cálculo de intervalo de confiança *bootstrap*. Os métodos são: intervalo de confiança *bootstrap* padrão, intervalo de confiança *bootstrap*-t, intervalo de confiança *bootstrap* Percentil, intervalo de confiança Percentil Corrigido em relação ao Viés (BCPB) e o intervalo de confiança de Correção do Viés Acelerado (Biased Corrected Accelerated) BC_a.

Segundo NAVIDI [19], o interessante é estimar o viés das estimativas dos parâmetros e efetuar as correções necessárias.

O método de simulação *bootstrap* foi originalmente proposto por Bradley Efron em um influente artigo publicado no Annals of Statistics, em 1979. Este método de simulação se baseia na construção de distribuições amostrais por reamostragem, e é muito utilizado para estimar intervalos de confiança e estimar o viés.

Vários esquemas diferentes de simulação *bootstrap* têm sido propostos na literatura e muitos deles apresenta-se bom desempenho em uma ampla variedade de situações.

Por outro lado, se o interesse for fazer inferência para algum outro parâmetro, como por exemplo o coeficiente de correlação, não há nenhuma fórmula analítica simples que permita calcular o seu erro padrão. O método de *bootstrap* foi construído para fazer simulações para este tipo de problema. Dado o custo alto e a escassez consequente de dados em muitas aplicações, combinadas com o custo reduzido e abundância do poder da computação, o método de *bootstrap* se torna uma técnica muito atraente por extrair informações de dados empíricos DIACONIS e EFRON [6].

A técnica consiste em realizar amostragens de tamanho igual ao da amostra original com reposição da mesma. Em outras palavras, são realizados n sorteios, sendo n o número de observações disponíveis na amostra, com reposição da amostra inicial, o que origina uma amostra *bootstrap*.

Este procedimento deve ser repetido B vezes para que se obtenham B amostras *bootstrap*. Em posse destas B amostras, é possível construir uma distribuição *bootstrap* da variável de interesse. Essa distribuição estimada é que será utilizada para realizar inferências e tirar informações sobre o parâmetro em estudo, ou seja, em posse desta distribuição *bootstrap* é possível obter informações e até testar hipóteses. EFRON e TIBSHIRANI [7] provam que a distribuição *bootstrap* converge para a distribuição verdadeira quando B (número de amostras *bootstrap*) tende ao infinito. Cada aplicação deve levar em conta a peculiaridade da amostra em estudo, além, é claro, das características da variável de interesse.

O termo *bootstrap* paramétrico é utilizado quando se tem alguma suposição da distribuição, já o termo *bootstrap* não paramétrico é utilizado quando não se tem nenhuma suposição da distribuição do conjunto de dados, ou seja, a distribuição da qual os dados vieram é desconhecida. Na forma não paramétrica não se pode empregar a distribuição normal de probabilidade para fazer estimativas de parâmetro e de intervalo de confiança. No caso do *bootstrap* paramétrico, o método consiste em gerar amostras baseadas na distribuição de probabilidade conhecida utilizando como parâmetros desta distribuição a estimativa dos mesmos, obtida através da amostra geral.

Neste capítulo foi apresentado o procedimento geral do *bootstrap* de forma bem intuitiva. Nos próximos capítulos, abordar-se-á de forma detalhada o emprego do *bootstrap* para estimar o erro padrão, bem como a estimativa *bootstrap* do viés e os intervalos de confiança *bootstrap*.

3 O *bootstrap*

Neste capítulo é descrito o método *bootstrap*, tanto na sua forma não paramétrica como na sua forma paramétrica, como obter a estimativa *bootstrap* do erro padrão e a estimativa do viés. Também é mencionado como obter a amostra *jackknife* e utilizá-la para estimar o erro padrão. Primeiramente, são apresentadas algumas definições e resultados importantes para o bom entendimento do trabalho.

3.1 Definições básicas

Definição 3.1. *Experimento aleatório é qualquer experimento cujos resultados são aleatórios. A repetição do experimento nas mesmas condições não leva necessariamente ao mesmo resultado.*

Definição 3.2. *Espaço amostral é o conjunto de todos os possíveis resultados de um experimento aleatório. Notação:* Ω.

Definição 3.3. *Considere um experimento aleatório e* Ω *o espaço amostral associado a esse experimento. Uma função* X*, que associa a cada elemento* $\omega \in \Omega$ *um número real,* $X(\omega)$*,é denominada variável aleatória (v.a.). Notação: X,Y,Z, etc.*

As variáveis aleatórias são fundamentais para as aplicações, pois elas representam as características de interesse em uma população.

Definição 3.4. *População é o conjunto de todos os elementos ou resultado sob investigação (universo do estudo).*

Definição 3.5. *Uma amostra aleatória (a.a.) é um subconjunto da população, selecionado aleatoriamente.*

Exemplo 3.1. Em um estudo deseja-se estudar o conteúdo estomacal dos peixes do Rio São Francisco. Para isso, foram selecionados 100 peixes aleatoriamente. A população, neste caso, são os peixes do Rio São Francisco e a amostra aleatória são os 100 peixes selecionados.

Definição 3.6. *Uma amostra casual simples de tamanho n de uma variável aleatória X, com uma dada distribuição, é o conjunto de n variáveis aleatórias independentes* $X_1, X_2, ..., X_n$ *cada uma com a mesma distribuição de X. Ou seja, a amostra será a n-upla ordenada* $x = (x_1, x_2, ..., x_n)$*, onde* x_i *indica a observação do* i*-ésimo elemento sorteado. Todas as amostras tem a mesma chance de serem selecionadas.*

Neste trabalho, considere-se amostra casual simples com reposição.

Definição 3.7. *Uma estatística é uma característica da amostra, ou seja uma estatística T é uma função de* $(x_1, x_2, ..., x_n)$*,* $T = f(x_1, x_2, ..., x_n)$*.*

Definição 3.8. *Parâmetros são medidas de interesse da população. Em geral são desconhecidos e precisam ser estimados. Notação:* $\mu, \theta, \lambda, \alpha$*.*

Definição 3.9. *Estimadores são funções dos valores da variável aleatória na amostra, construídas de forma adequada para estimar um parâmetro de interesse. Notação:* $\widehat{\mu}, \widehat{\theta}, \widehat{\lambda}, \widehat{\alpha}$*.*

3.1.1 Variáveis aleatórias discretas

Definição 3.10. *Seja X uma variável aleatória. Se o conjunto de valores possíveis de X for enumerável (finito ou infinito), denomina-se X de variável aleatória discreta. Isto é, os possíveis valores de X podem ser postos em lista como* $x_1, x_2, x_3, ...$

Definição 3.11. *Seja X uma variável aleatória discreta. A cada possível resultado* x_i *associamos um valor* $p(x_i) = P(X = x_i)$*, denominado função de probabilidade para os valores* x_i*. As funções de probabilidade devem satisfazer as seguintes condições:*

1. $0 \leq P(X = x_i) \leq 1, i = 1, 2, 3, ...$ *e*
2. $\sum_{i=1}^{\infty} P(X = x_i) = 1.$

Existem duas medidas importantes das variáveis aleatórias, tanto para o caso discreto como no caso contínuo. A esperança como medida de posição (locação) e a variância como medida de variabilidade dos valores das variáveis aleatórias.

Definição 3.12. *Seja X uma v.a. discreta, com valores possíveis* $x_1, x_2, x_3, ..., x_n$ *e seja* $p(x_i) = P(x = x_i), i = 1, ..., n$*. A esperança de X, também chamada de valor esperado de X, denotada por* $E(X)$ *ou* μ*, é definida como,*

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i). \tag{3.1}$$

Este valor também é denominado valor médio ou expectância de X.

Definição 3.13. *Seja X uma variável aleatória discreta. Define-se a variância de X, denotada por $Var(X)$ ou σ^2 por,*

$$Var(X) = E\left(X^2\right) - \left[E\left(X\right)\right]^2 = \sum_{i=1}^{\infty} x_i^2 P(X = x_i) - \left[\sum_{i=1}^{\infty} x_i P(X = x_i)\right]^2. \qquad (3.2)$$

3.1.2 Variáveis aleatórias contínuas

Definição 3.14. *Seja X uma variável aleatória. Suponha que o contradomínio de X seja um intervalo ou uma coleção de intervalos do conjunto dos reais. Então diz-se que X é uma variável aleatória contínua. Ou seja as variáveis aleatórias assumem valores num conjunto não enumerável.*

Definição 3.15. *Considere uma v.a contínua X que pode assumir qualquer valor em um intervalo dos reais. Definimos uma função $f_X(x)$, denominada função densidade de probabilidade tal que,*

$$P\left(a < X < b\right) = \int_b^a f_X\left(x\right) dx.$$

As funções densidades de probabilidade devem satisfazer as seguintes condições:

1. $f_X(x) \geq 0$ *e*
2. $\int_{-\infty}^{\infty} f_X\left(x\right) dx = 1.$

Definição 3.16. *Seja X uma variável aleatória contínua com função densidade de probabilidade f. Definimos o valor esperado ou esperança de X por,*

$$E\left(X\right) = \int_{-\infty}^{\infty} x f_X\left(x\right) dx. \qquad (3.3)$$

A interpretação de $E(X)$ para o caso contínuo é similar ao mencionado para variáveis aleatórias discretas.

Definição 3.17. *Seja X uma variável aleatória contínua. Definimos a variância de X, denotada por $Var(X)$ ou σ^2 por,*

$$Var(X) = E(X^2) - \left[E(X)\right]^2 = \int_{-\infty}^{\infty} x^2 f_X\left(x\right) dx - \left[E(X)\right]^2. \qquad (3.4)$$

A função de distribuição acumulada fornece uma maneira de descrever como as probabilidades são associadas aos valores ou aos intervalos de valores de uma variável aleatória.

Definição 3.18. *A função de distribuição acumulada de uma variável aleatória X é uma função que a cada número real x associa o valor,*

$$F(x) = P(X \leq x), x \in \mathbb{R}. \tag{3.5}$$

A notação $(X \leq x)$ é usada para designar o conjunto $\{\omega \in \Omega \mid X(\omega) \leq x\}$, isto é, denota a imagem inversa do intervalo pela variável aleatória X. Com isso, pode-se observar que a função de distribuição acumulada F tem como domínio os números reais $\mathbb{R}$ e imagem o intervalo $(-\infty, x]$. O conhecimento da função de distribuição acumulada é suficiente para entender o comportamento de uma variável aleatória. Mesmo que a variável assuma valores apenas num subconjunto dos reais, a função de distribuição é definida em toda a reta. Ela é chamada de função de distribuição acumulada, pois acumula as probabilidades dos valores inferiores ou iguais a x.

Definição 3.19. *As variáveis X e Y são independentes se para quaisquer dois conjuntos de números reais A e B, $P(X \in A\ ,\ Y \in B) = P(X \in A)P(Y \in B)$.*

Quando X e Y são v.a. discretas, a condição de independência é equivalente a $p(x, y) = p_X(x)\, p_Y(y)$ para todo x, y.

Quando X e Y são v.a. contínuas, a condição de independência é equivalente a $\int\int f(x, y)\, dxdy = \int f(x)\, dx \int f(y)\, dy$ para todo x, y.

Se X_i e X_j são independentes e possuem mesma função de probabilidade (ou função densidade de probabilidade), para todo $i \neq j$, dizemos que as v.a. são independentes e identicamente distribuídas (*i.i.d.*).

A característica da população na qual estamos interessados, em geral pode ser representada por uma variável aletória X. Se $F(x)$ é conhecida completamente não há necessidade de colher uma amostra pois toda informação desejada é obtida através da distribuição da variável. Mas isso raramente acontece, ou seja a informação a respeito da variável, é apenas parcial. Por exemplo, podemos conhecer a forma da distribuição mas desconhecer os parâmetros que a caracterizam (por exemplo: média e variância) ou de forma contrária, podemos ter ideia da média e da variância, mas desconhecer a distribuição da variável, ou ainda o que é mais comum, não possuírmos informações nem sobre os parâmetros, nem sobre a forma da distribuição da variável. Um dos objetivos da inferência estatística é estimar um ou vários parâmetros desconhecidos da população. De uma forma geral denotaremos o parâmetro a estimar por θ.

Definição 3.20. *Um estimador $\widehat{\theta}$ é não viciado ou não viesado para o parâmetro θ se $E\left(\widehat{\theta}\right) = \theta$. Ou seja, um estimador é não viciado se o seu valor esperado coincide com o valor do parâmetro de interesse.*

Definição 3.21. *Um estimador $\widehat{\theta}$ é consistente, se a medida que o tamanho da amostra aumenta, seu valor esperado converge para o valor do parâmetro de interesse e sua variância converge para zero. Ou seja $\widehat{\theta}$ é consistente se as duas propriedades seguintes são satisfeitas:*

$$a) \lim_{n\to\infty} E\left(\widehat{\theta}\right) = \theta, \qquad b) \lim_{n\to\infty} Var\left(\widehat{\theta}\right) = 0. \tag{3.6}$$

Definição 3.22. *O erro padrão ou de desvio padrão, é definido como sendo a raiz quadrada de sua variância.*

Definição 3.23. *O erro padrão estimado também chamado de desvio padrão estimado para a média $\overline{x}$ com base em n dados independentes, $x_1, x_2, ..., x_n$, é dado por:*

$$dp\left(\overline{x}\right) = \sqrt{\frac{S^2}{n}}, \tag{3.7}$$

onde

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2.$$

O erro padrão dado por (3.7) é a mais simples medida de precisão estatística para a média. O S^2 é um estimador não viesado e consistente para a variância σ^2.

O erro padrão não é uma estimativa de uma quantidade pertinente a uma população, mas sim uma medida da incerteza da média amostral vista como uma estimativa da média populacional. A expressão (3.7) indica que a magnitude desta incerteza diminui conforme o tamanho da amostra n aumenta.

No caso de estimadores para outos parâmetros pode não existir uma fórmula como enunciada em (3.7) para obter o erro padrão estimado. Em outras palavras pode ser inviável avaliar a precisão de alguns estimadores.

A função de distribuição empírica (definição 3.24), que denotaremos por $\widehat{F}$, é uma estimativa de toda a distribuição F. Uma maneira de estimar algum aspecto interessante de F, além da média, variância e mediana é usar o aspecto correspondente de $\widehat{F}$.

Definição 3.24. *Considere uma amostra aleatória de tamanho n e uma distribuição de probabilidade F. A função de distribuição empírica $\widehat{F}$ é definida como sendo a distribuição discreta que atribui probabilidade $\dfrac{1}{n}$ em cada x_i, $i = 1, 2, ..., n$.*

A função de distribuição empírica fornece a porcentagem ou proporção de valores da amostra que são menores ou iguais a um determinado valor x. Assim dado um valor p qualquer, entre 0 e 1, podemos obter o valor V_p que divide a amostra em duas partes tais que $100\%p$ dos elementos são menores ou iguais a V_p e os restantes são maiores ou iguais a V_p . Este valor V_p é denominado de percentil de ordem p ou porcentagem $100\%p$.

O princípio do "plug-in", definido em 3.25, é um método para estimar parâmetros a partir de amostras. Este princípio é adequado quando a única informação disponível de F vem da amostra $x = (x_1, x_2, ..., x_n)$, através de $\widehat{F}$. Esse princípio consiste em estimar um parâmetro, uma quantidade que descreve a população, utilizando a estatística que é a quantidade correspondente para a amostra.

Definição 3.25. *A estimativa "plug-in" para o parâmetro $\theta = t(F)$ é definida como:*

$$\widehat{\theta} = t\left(\widehat{F}\right). \tag{3.8}$$

Em outras palavras, estima-se a função $\theta = t(F)$ utilizando a distribuição empírica $\widehat{F}$.

Em geral, a estimativa 'plug-in' para a esperança $\theta = E_{\widehat{F}}(X)$ é:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

Segundo BUSSAB e MORETTIN [3], assumir que os dados tenham uma distribuição normal é conveniente tanto do ponto de vista teórico como do ponto de vista computacional. Mas isto fornece um problema de fundamental importância: sob quais circunstâncias é razoável assumir que a distribuição normal pode ser usada? Gauss trabalhou neste problema por muito tempo, mas é um resultado de Laplace que é utilizado hoje. Enunciado em 1810, Laplace chamou este resultado de Teorema Central do Limite, que diz que sob a hipótese de amostragem aleatória, quando o tamanho da amostra aumenta, a distribuição da média amostral se aproxima de uma distribuição normal com média μ e variância $\frac{\sigma^2}{n}$, ou seja, se o tamanho da amostra é suficientemente grande, pode-se assumir que a média amostral segue uma distribuição normal.

Teorema 3.1. *(Teorema Central do Limite - TCL) Para uma amostra casual simples $(x_1, x_2, ..., x_n)$, retirada de uma população com média μ e variância σ^2, a distribuição amostral da média $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ aproxima-se de uma distribuição normal com média μ e variância $\frac{\sigma^2}{n}$ quando $n \to \infty$.*

$$\overline{x} \simeq N\left(\mu, \frac{\sigma^2}{n}\right). \tag{3.9}$$

Corolário 3.1. *Se* $(x_1, x_2, ..., x_n)$ *é uma amostra casual simples de* X *com média* μ *e variância* σ^2 *e* $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *, então* $Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \simeq N(0\ ,\ 1)$.

Ou seja, a distribuição de Z se aproxima da distribuição normal padrão quando n tende ao infinito. O mais impressionante do teorema central do limite é o fato de que nada é dito a respeito da função distribuição. Ou seja, qualquer que seja a função de distribuição, desde que ela tenha variância finita, a média amostral terá uma distribuição aproximadamente normal para amostras suficientemente grandes.

3.2 *Jackknife*: o método pioneiro para estimar o erro padrão

Jackknife é um método de reamostragem também denominado computacional intensivo, não paramétrico que pode ser utilizado para avaliar alguma medida de variabilidade, por exemplo o erro padrão. Introduzido por volta de 1950, por QUENOUILLE [20] e aprimorado anos depois por TUKEY [27]. O metódo *Jackknife* é também chamado de "leave-one-out", utilizado para estimar a variância de estimadores em condições teoricamente complexas ou em que não se tem confiança no modelo especificado. Nesse método recalcula-se o valor da estatística de interesse em cada uma das n amostras denominadas amostras *jackknife* de tamanho $(n-1)$ que são formadas a partir da amostra geral $x = (x_1, x_2, ..., x_n)$ da seguinte forma:

A primeira amostra *jackknife* é formada por todas as observações da amostra original, exceto a primeira.

A segunda amostra *jackknife* é formada por todas as observações da amostra original, exceto a segunda. E assim sucessivamente.

A n-ésima amostra *jackknife* é formada por todas as observações da amostra original, exceto a n-ésima.

Nota-se que o *jackknife* é um método que incorpora todas as n amostras extraídas sem reposição de tamanho $(n-1)$ da amostra original (EFRON), [7]. Ou seja, sendo uma amostra original de tamanho n, $x = (x_1, x_2, ..., x_{n-1}, x_n)$ são geradas as amostras *jackknife*:

$$x^1 = (x_2, x_3, ..., x_{n-1}, x_n),$$

$$x^2 = (x_1, x_3, ..., x_{n-1}, x_n),$$
$$\vdots$$
$$x^n = (x_1, x_2, ..., x_{n-1}).$$

A i-ésima amostra *jackknife* consiste no conjunto de dados observados com o i-ésimo termo removido.

Da amostra original $x = (x_1, x_2, ..., x_n)$, estima-se $\widehat{\theta} = s(x)$, em que $s(x)$ é uma estatística de interesse.

Em cada uma destas amostras *jackknife* geradas, estima-se $\widehat{\theta}^J_{(i)} = s\left(x^i\right)$, para $i = 1, 2, 3, ..., n$ denominada réplica *jackknife* da estatística de interesse.

Na Figura 3.1 foram geradas n amostras *jackknife* da amostra original. Cada amostra *jackknife* tem tamanho *(n-1)* e é obtida suprindo o i-ésimo valor das observações. Em posse das amostras *jackknife* são calculadas as réplicas *jackknife* $s\left(x^i\right)$ $i = 1, 2, ..., n$.

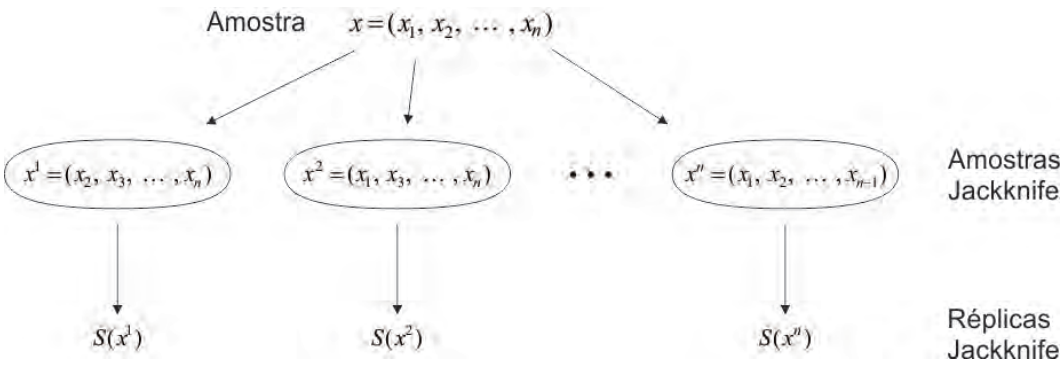


Figura 3.1: Réplicas *Jackknife*

Definição 3.26. *Considere* $\widehat{\theta}^J_{(\cdot)} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\theta}^J_{(i)}$. *O estimador jackknife do viés é definido por,*

$$\widehat{vi\acute{e}s}_{jackk} = (n-1)\left(\widehat{\theta}^J_{(\cdot)} - \widehat{\theta}\right). \tag{3.10}$$

Definição 3.27. *O estimador jackknife do erro padrão é dado por,*

$$\widehat{dp}_{jackk} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}^J_{(i)} - \widehat{\theta}^J_{(\cdot)}\right)^2}, \tag{3.11}$$

onde $\widehat{\theta}^J_{(\cdot)} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\theta}^J_{(i)}$.

Segundo MARTINEZ e LOUZADA [14] é possível obter uma expressão mais simples para (3.11) quando a estatística for a média amostral. Considere então $\widehat{\theta} = s(x) = \overline{x}$. Obtemos a média amostral sem considerar a i-ésima observação, dada pela expressão:

$$\overline{x}^{(i)} = \frac{n\overline{x} - x_i}{n-1}, \qquad i = 1, 2, ..., n. \tag{3.12}$$

A média das n médias, $\overline{x}^{(i)}$ de (3.12) é igual a média amostral.

De fato,

$$\widehat{\theta}(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\overline{x}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{n\overline{x} - x_i}{n-1} =$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}(n\overline{x} - x_i) = \frac{1}{n(n-1)}\left[n^2\overline{x} - \sum_{i=1}^{n}x_i\right] =$$

$$= \frac{1}{n(n-1)}\left[n^2\overline{x} - n\overline{x}\right] = \frac{1}{n(n-1)}n(n-1)\overline{x} = \overline{x}$$

Portanto,

$$\widehat{\theta}(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\overline{x}^{(i)} = \overline{x}. \tag{3.13}$$

Usando a definição do estimador *jackknife* do erro padrão (3.11) para a média $\widehat{\theta} = \overline{x}$ temos,

$$\widehat{dp}_{jack}\left(\widehat{\theta}\right) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\overline{x}^{(i)} - \overline{x}\right)^2}. \tag{3.14}$$

Substituindo (3.12)em (3.14) segue que,

$$\widehat{dp}_{jack}(\overline{x}) = \sqrt{\left(\frac{n-1}{n}\right)\sum_{i=1}^{n}\left[\frac{n\overline{x} - x_i - (n-1)\overline{x}}{n-1}\right]^2} =$$

$$= \sqrt{\left(\frac{n-1}{n}\right)\left(\frac{1}{n-1}\right)^2\sum_{i=1}^{n}(n\overline{x} - x_i - n\overline{x} + \overline{x})^2} =$$

$$= \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\overline{x} - x_i)^2} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Portanto,

$$\widehat{dp}_{jack}(\overline{x}) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.15}$$

A expressão (3.15) é idêntica à expressão (3.7), o que motiva o termo $\frac{(n-1)}{n}$ em (3.11). Este termo pode ser interpretado como uma correção, para que a expressão (3.14) seja um estimador não viesado de θ.

Exemplo 3.2. Considere os dados a seguir:

Tabela 3.1: Dados para encontrar o estimador *jackknife* do erro padrão da média

52	104	146	10	51	30	40	27	46

Vamos determinar o estimador *jackknife* do erro padrão de $\widehat{\theta} = \overline{x}$.

Observe a Tabela 3.2, que ilustra os passos recorrentes para encontrarmos o estimador *jackknife* do erro padrão para a média.

Tabela 3.2: Construção do erro padrão *jackknife* da média, exemplo 3.6

i	x_i	$\overline{x}^{(i)}$	$\left(\overline{x}^{(i)} - \overline{x}_{(\cdot)}\right)^2$
1	52	56,74	0,27
2	104	50,24	35,66
3	146	44,99	125,93
4	10	61,99	33,38
5	51	56,87	0,42
6	30	59,49	10,71
7	40	58,24	4,11
8	27	59,87	13,34
9	46	57,49	1,63
	$\overline{x} = 56,22$	$\widehat{\theta}_{(\cdot)} = 56,21$	soma=225,49

Tem-se,

$$\overline{x}^{(i)} = \frac{n\overline{x} - x_i}{n-1} = \frac{9\overline{x} - x_i}{8} = \frac{9 \cdot 56,22 - x_i}{8} = \frac{505,98 - x_i}{8},$$

$$\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\overline{x}_i = 56,21.$$

Logo,

$$\widehat{dp}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\overline{x}^i - \overline{x}\right)^2} = \sqrt{\frac{8}{9}\,(225,49)} = 14,15.$$

Observa-se que o estimador *jackknife* do erro padrão de $\widehat{\theta} = s(x) = \overline{x}$ é igual ao erro padrão da média estimado por $\widehat{dp}(\overline{x}) = \sqrt{\dfrac{s^2}{n}}$ o que era de se esperar.

Descreve-se a seguir o algoritmo *jackknife* para o cálculo do erro padrão *jackknife* de um estimador geral.
Considere uma amostra original $x = (x_1, x_2, .., x_n)$ de tamanho n.

O algoritmo para estimativa do erro padrão *jackknife*,

1) defina a estatística de interesse;
$\widehat{\theta} = s(x)$;

2) gere as amostras *jackknife*;
gere a amostra *jackknife* 1;
$x^1 = (x_2, x_3, ..., x_n)$
gera a amostra *jackknife* 2;
$x^2 = (x_1, x_3, ..., x_n)$
$\vdots$
gere a amostra *jackknife* n;
$x^n = (x_1, x_2, ..., x_{n-1})$

3) calcule a estatística de interesse da cada amostra *jackknife* obtido em 2;
$\widehat{\theta}_{(1)} = s\left(x^1\right)$
$\widehat{\theta}_{(2)} = s\left(x^2\right)$
.
.
.
$\widehat{\theta}_{(n)} = s\left(x^n\right)$

4) estime o erro padrão *jackknife* da estatística utilizando a seguinte expressão,

$$\widehat{dp}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(i)} - \widehat{\theta}(\cdot)\right)^2},$$

onde $\widehat{\theta}(\cdot) = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}$, sendo θ o valor que a estatística assume na amostra original.

A Figura 3.2 esquematiza o diagrama do algoritmo *jackknife* do erro padrão.

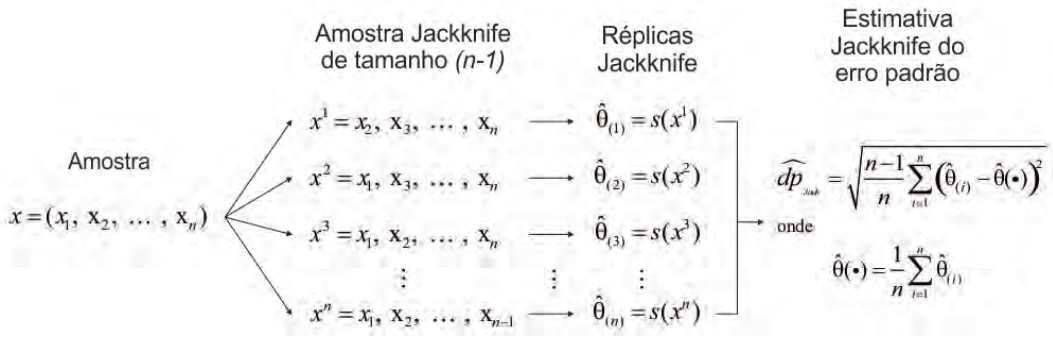


Figura 3.2: Algoritmo para estimativa do erro padrão *jackknife*

Nota-se que o algoritmo *jackknife* multiplica a informação inicial, pois além das estimativas de θ, obtém-se n valores para a mesma estatística.

3.3 Método *bootstrap*

Para calcular os intervalos de confiança *bootstrap*, descritos no capítulo 4, é necessário calcular primeiramente o erro padrão *bootstrap*, já que é utilizado no lugar do erro padrão da amostra original. Antes de descrever como encontrar o valor do erro padrão *bootstrap*, (seção 3.5), nesta seção, descreve-se como encontrar as réplicas *bootstrap*.

A ideia básica do método *bootstrap* é reamostrar de um conjunto de dados diretamente ou via um modelo ajustado, a fim de criar réplicas dos dados, a partir das quais pode-se avaliar a variabilidade de quantidade de interesse, sem usar cálculos analíticos. Assim pode-se classificar o método *bootstrap* em paramétrico e não paramétrico, como veremos a seguir.

3.3.1 Método *bootstrap* não paramétrico

O termo *bootstrap* não paramétrico é utilizado quando não se tem nenhuma suposição sobre a distribuição do conjunto de dados, ou seja, a distribuição da qual os dados vieram é desconhecida. Neste caso utiliza-se o conjunto de dados diretamente para gerar as réplicas.

3.3.2 Método *bootstrap* paramétrico

O termo *bootstrap* paramétrico é utilizado quando se tem alguma suposição sobre a distribuição de probabilidade dos dados. Ou seja, no caso do método *bootstrap*

paramétrico, existe uma suposição sobre a distribuição que originou os dados e as B amostras *boostrap* são geradas utilizando esse modelo, a partir dos parâmetros estimados com os dados da amostra original.

Tanto o *bootstrap* não paramétrico como o *bootstrap* paramétrico geram uma grande quantidade de amostras. A diferença chave entre eles é que o método *bootstrap* não paramétrico utiliza-se a amostra original para gerar as amostras *bootstrap* já o método *bootstrap* paramétrico utiliza-se informações que tem-se em mãos, por exemplo: depois de ajustado o modelo conseguimos estimar os parâmetros dessa distribuição de interesse e utiliza-se esses parâmetros para gerar as amostras. Ou seja, no caso paramétrico cada amostra *bootstrap* é obtida da distribuição paramétrica que originou os dados, ao invés de reamostrar-se as observações disponíveis. A forma de obtenção das réplicas *bootstrap* é a mesma para o caso não paramétrico e paramétrico.

3.4 Obtenção das réplicas *bootstrap*

Considere uma amostra de tamanho n, $x = (x_1, x_2, ..., x_n)$, oriunda de uma distribuição F que chamaremos de amostra original. O objetivo é estimar um parâmetro de interesse $\theta = t(F)$ com base em x. Para isso, pode-se calcular uma estimativa $\widehat{\theta} = s(x)$. Tal estatística de interesse, pode ser o erro padrão. Segundo HESTERBERG [12], a amostra original representa a população (objeto de estudo).

Definição 3.28. *Uma amostra bootstrap* $x^* = (x_1^*, x_2^*, ..., x_n^*)$ *é obtida por amostragem aleatória n vezes, com reposição a partir dos dados originais* $x = (x_1, x_2, ..., x_n)$.

A notação * indica que x^* não é um conjunto de dados reais e sim uma reorganização de $x = (x_1, x_2, ..., x_n)$. Por exemplo, pode-se obter, $x_1^* = x_7$, $x_2^* = x_3$, $x_3^* = x_3$, $x_4^* = x_2$, $\cdots$, $x_n^* = x_9$.

Alguns elementos podem não aparecer e outros aparecer com mais frequência.

A distribuição *bootstrap* se baseia em muitas amostras *bootstrap* e representa uma distribuição amostral desta estatística. Para obter resultados confiáveis, é preciso realizar várias amostras *bootstrap* do mesmo tamanho n. Essas amostras *bootstrap* devem ser feitas com reposição e de forma aleatória. Para MONTGOMERY e RUNGER [18], o número de amostras *bootstrap* pode ser estipulado verificando a variação do desvio padrão para a estimativa do parâmetro em questão calculado para as amostras *bootstrap* à medida que estas são realizadas. Quando esse valor ficar estável teremos a quantidade de amostras *bootstrap* adequada. A variabilidade do *bootstrap* é dada pela escolha da amostra original e pelas amostras *bootstrap*. Deve-se gerar um número finito B de amostras *bootstrap*. Em relação ao número de réplicas *bootstrap* B EFRON

e TIBSHIRANI [7], KENDALL e STUARD [13] e HALL [11] discutem a quantidade de replicações *bootstrap* necessárias para uma boa estimativa do erro padrão e do intervalo de confiança. EFRON e TIBSHIRANI [7] afirmam que para obter-se uma boa estimativa do erro padrão através do *bootstrap* são necessárias entre 25 e 200 replicações e que, para uma boa estimativa dos limites de confiança seriam necessárias mais de 500 replicações. Utiliza-se neste trabalho 1000 replicações para a construção de intervalos de confiança *bootstrap* e 40 replicações para estimar o erro padrão *bootstrap*.

Definição 3.29. *Para cada amostra bootstrap estima-se θ através das réplicas bootstrap $\widehat{\theta}_b^* = s\left(x_b^*\right)$, $b = 1, 2, ..., B$, onde B é o número de amostras bootstrap geradas.*

Resumidamente, o *bootstrap* gera um grande número de amostras *bootstrap* B independentes $x_1^*, x_2^*, ..., x_B^*$, por amostragens de tamanho n igual ao da amostra original com reposição da mesma. Correspondente a cada amostra *bootstrap* tem-se uma réplica *bootstrap*, que é o valor da estatística avaliada, denotada por $\widehat{\theta}_b^* = s\left(x_b^*\right)$, $b = 1, 2, ..., B$. A Figura 3.3 ilustra esse procedimento.

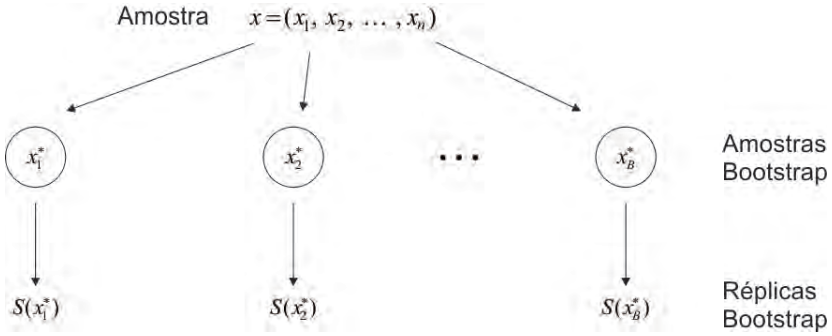


Figura 3.3: Réplicas *Bootstrap*

3.5 Erro padrão *bootstrap*

O objetivo desta seção é descrever o erro padrão *bootstrap*, bem como a estimativa *bootstrap* do viés, valores úteis para os cálculos dos intervalos de confiança *bootstrap*, descrito no capítulo 4.

Definição 3.30. *A estimativa bootstrap do erro padrão é o desvio padrão das réplicas bootstrap, ou seja, define-se a estimativa bootstrap do erro padrão $\widehat{dp}_{boot}$ por,*

$$\widehat{dp}_{boot}\left(\widehat{\theta}^*\right) = \sqrt{\left\{\frac{1}{B-1}\sum_{b=1}^{B}[s\left(x_b^*\right) - s\left(\cdot\right)]^2\right\}}, \tag{3.16}$$

com $s\left(x_b^*\right)$ igual ao valor da estatística para cada amostra bootstrap e B é o número de amostras bootstrap geradas, onde

$$s(\cdot) = \frac{1}{B}\sum_{b=1}^{B} s(x_b^*).$$

A estimativa do erro padrão *bootstrap* mostra o quanto de variação ou dispersão existe em relação a média (ou valor esperado).

Definição 3.31. *Define-se o estimador bootstrap ideal de* $dp_F\left(\widehat{\theta}^*\right)$, *como sendo o limite de* $\widehat{dp}_{boot}$ *quando* B *tende ao infinito, ou seja,*

$$\lim_{B\to\infty} \widehat{dp}_{boot} = dp_{\widehat{F}}\left(\widehat{\theta}^*\right). \tag{3.17}$$

Quando $B \to \infty$ tem-se a oportunidade de verificar melhor a variabilidade real. O quanto se deve aumentar B vai depender do método. O estimador *bootstrap* ideal e sua aproximação $\widehat{dp}_{boot}$ são chamados estimadores *bootstrap* não paramétricos, já que se baseiam em $\widehat{F}$, um estimador não paramétrico de F.

Apresenta-se em seguida o algoritmo *bootstrap* para o cálculo do erro padrão *bootstrap* de um estimador geral. Considere uma amostra original $x_n = (x_1, x_2, .., x_n)$ de tamanho n.

1) selecione B amostras *bootstrap* independentes $x_1^*, x_2^*, ..., x_B^*$, cada amostra constituindo de n valores retiradas com reposição de $x = (x_1, x_2, .., x_n)$;

2) calcule a réplica *bootstrap* $\widehat{\theta}_b^*$ para cada amostra *bootstrap*;

$$\widehat{\theta}_b^* = s(x_b^*), \qquad b = 1, 2, 3..., B;$$

3) utilize as B estimações $s(x_b^*)$ para calcular a estimação do erro padrão *bootstrap* da seguinte forma,

$$\widehat{dp}_{boot} = \sqrt{\left\{\frac{1}{B-1}\sum_{b=1}^{B}[\widehat{\theta}_b^* - \widehat{\theta}^*(\cdot)]^2\right\}},$$

onde $\widehat{\theta}^*(\cdot) = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\widehat{\theta}_b^*$.

A Figura 3.4 esquematiza o diagrama do algoritmo *bootstrap* do erro padrão.

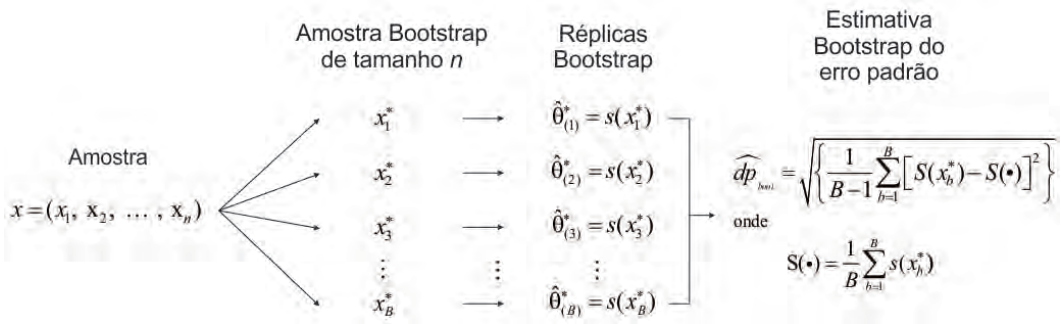


Figura 3.4: Algoritmo para estimativa do erro padrão *bootstrap*

O algoritmo descrito na Figura 3.4 aplica-se ao caso do *bootstrap* não paramétrico. No caso do *bootstrap* paramétrico, procede-se de maneira semelhante com uma única diferença: cada amostra *bootstrap* é obtida da distribuição paramétrica que originou os dados, ao invés de reamostrar-se as observações disponíveis. O livro de SHAO e WU [26], apresenta um apanhado geral dos teoremas sobre a consistência e precisão dos estimadores. Os autores mostram que a aproximação *bootstrap* é válida para a maioria das estatísticas de interesse e que seus estimadores são consistentes.

A técnica *bootstrap* nos permite verificar o viés olhando se a distribuição *bootstrap* da estatística está centrada na estatística da amostra original. Se o valor do viés é pequeno, há uma indicação de que os valores estimados devem se encontrar próximos dos valores verdadeiros.

Definição 3.32. *Um estimador utilizado para estimar um parâmetro apresenta viés, ou seja é viesado, quando a distribuição amostral não estiver centrada no verdadeiro valor do parâmetro.*

$$E\left(\widehat{\theta}^*(\cdot)\right) \neq \widehat{\theta}. \tag{3.18}$$

Definição 3.33. *A estimativa bootstrap do viés é definida dado por,*

$$\widehat{viés}_{boot}\left(\widehat{\theta}\right) = \widehat{\theta}^*(\cdot) - \widehat{\theta}, \tag{3.19}$$

onde $\widehat{\theta}^*(\cdot) = \dfrac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^*$.

Segundo EFRON [7] podemos considerar o viés pequeno se é menor que 25% de seu desvio padrão, isto é se $\left|\widehat{viés}_{boot}\left(\widehat{\theta}\right)\right| < 0,25\widehat{dp}[\left(\widehat{\theta}\right)]$.

A Tabela 3.3 mostra a classificação do viés.

Tabela 3.3: Classificação do viés

$\widehat{viés}_{boot}\left(\widehat{\theta}\right) < 0$	$\widehat{\theta}^* < \widehat{\theta}$	Subestima a estatística
$\widehat{viés}_{boot}\left(\widehat{\theta}\right) = 0$	$\widehat{\theta}^* = \widehat{\theta}$	Não há viés
$\widehat{viés}_{boot}\left(\widehat{\theta}\right) > 0$	$\widehat{\theta}^* > \widehat{\theta}$	Superestima a estatística

Definição 3.34. *O estimador de θ corrigido é definido por,*

$$\widetilde{\theta}_{boot} = \widehat{\theta} - \widehat{viés}_{boot}\left(\widehat{\theta}\right). \tag{3.20}$$

Os cálculos de (3.16), (3.19) e (3.20) podem ser feitos utilizando o código 1 do anexo A.1 (BORKOWSKI), [2].

4 Intervalos de confiança baseados no método *bootstrap*

Neste capítulo apresenta-se os intervalos de confiança baseados no método *bootstrap*. Antes descreve-se brevemente os intervalos de confiança assintóticos.

No processo de inferência estatística, geralmente não é suficiente somente avaliar determinadas características com o procedimento pontual. É necessário recorrer a um mecanismo de estimação por intervalo, que possibilita avaliar o erro que se comete na estimação pontual. A avaliação deste erro pode ser feita com a construção de um intervalo de confiança para o parâmetro de interesse.

4.1 Intervalos de confiança assintóticos

Uma estimativa pontual de um parâmetro de interesse θ, pode não ser suficiente no auxílio de deduções. Daí, as medidas de precisão desta estimativa possibilitam que o pesquisador conjecture e tire conclusões baseadas em suas observações. O intervalo de confiança produz uma avaliação do erro que se comete na estimação, sob ponto de vista diferente do erro padrão.

Quando utilizamos um estimador pontual como por exemplo a média amostral $\overline{X}$, para estimar um parâmetro como a média populacional μ, não temos meios de julgar qual a possível magnitude do erro que estamos cometendo. Daí surge a ideia de construirmos intervalos de confiança para os parâmetros populacionais os quais são baseados nas distribuições amostrais dos estimadores pontuais (MAGALHÃES), [15].

A estimação intervalar, como o próprio nome diz, consiste na determinação de um intervalo onde, com uma certa confiança, esteja o parâmetro θ desconhecido, tendo-se em conta seu estimador. A vantagem da estimação intervalar é a possibilidade de determinar o erro máximo cometido na estimação, com uma certa confiança.

Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. O quanto estas estimativas são prováveis será determinado pelo coeficiente de confiança $(1-\alpha)$, para $\alpha \in (0,1)$.

Definição 4.1. *Considere uma amostra da variável X retirada de uma população com*

distribuição que depende de θ. Denota-se o intervalo de confiança para θ da seguinte forma,

$$[L_I(x) \ , \ L_S(x)] \tag{4.1}$$

onde, $L_I(x)$ é o limite inferior do intervalo e $L_S(x)$ é o limite superior do intervalo.

Definição 4.2. *O intervalo de confiança é simétrico quando a quantidade*

$$q(x) = \left| \frac{L_I(x) - \widehat{\theta}}{L_S(x) - \widehat{\theta}} \right| = 1,$$

e dizemos que $q(x)$ é o índice de simetria de um intervalo de confiança.

Dentre um grande número de possíveis realizações de um experimento que objetiva gerar uma estimativa $\widehat{\theta}$ de θ, todas considerando amostras de tamanho n, de cada 100, $100(1-\alpha)$ produzirão intervalos de confiança que vão conter θ, o verdadeiro valor do parâmetro. A amostra obtida no experimento é apenas uma dentre várias ou infinitas possibilidades e a estimativa $\widehat{\theta}$ obtida na amostra poderia ser diferente em alguma outra amostra obtida da mesma população, através do mesmo processo de amostragem. Realizado o experimento uma única vez, o pesquisador tem em mãos apenas uma destas possíveis estimativas e daí surge a seguinte questão: qual a confiança que pode-se depositar nesta estimativa? O coeficiente de confiança ajuda a responder essa questão.

A interpretação de (4.1) pode ser dada, por

$$P_F[L_I(x) \leq \theta \leq L_S(x)] = (1-\alpha), \qquad 0 \leq \alpha \leq 1. \tag{4.2}$$

O intervalo de confiança descrito em (4.2) não é uma aproximação, ele é dito exato, ou seja a probabilidade em questão é igual a $(1-\alpha)$.

Em geral os intervalos de confiança são construídos com áreas iguais sob a curva normal,

$$P_F[\theta \leq L_I(x)] = \frac{\alpha}{2}, \qquad P_F[\theta \geq L_S(x)] = \frac{\alpha}{2}. \tag{4.3}$$

No entanto as vezes não é possível construir intervalos de confiança como em (4.2). Os intervalos de confiança mais utilizados são os aproximados, ou seja,

$$P_F[L_I(x) \leq \theta \leq L_S(x)] \approx (1-\alpha). \tag{4.4}$$

A principal limitação para a construção de intervalos conforme enunciado em (4.4) é a necessidade de uma amostra suficientemente grande para garantir a validade das

aproximações em questão, pois a teoria utilizada para a construção de tais intervalos é a teoria assintótica. Os intervalos de confiança exatos são construídos através de soluções analíticas complexas, já os intervalos de confiança aproximados dependem de aproximações assintóticas como a normalidade assintótica dos estimadores de máxima verossimilhança, (EMV).

4.1.1 Intervalos de confiança para a média

Quando queremos estimar a média de uma população através de uma amostra temos dois casos distintos a considerar: quando a variância da população é conhecida e quando ela é desconhecida.

Para interpretar o intervalo de confiança da média, assume-se que os valores são amostrados de forma independente e aleatória de uma população normal com média μ e variância σ^2. Dado que estas suposições são válidas, temos $100(1-\alpha)$ de "chance" do intervalo conter o verdadeiro valor da média populacional. Em outras palavras, se produzirmos diversos intervalos de confiança provenientes de diferentes amostras independentes de mesmo tamanho, podemos esperar que aproximadamente $100(1-\alpha)\%$ destes intervalos devem conter o verdadeiro valor da média populacional.

Considere uma amostra aleatória simples $x_1, x_2, ..., x_n$ obtida de uma população com distribuição Normal, com média μ e variância σ^2 conhecida. Desta forma, a distribuição amostral da média também é Normal com média μ e variância $\dfrac{\sigma^2}{n}$. Logo, temos

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

A variável Z segue distribuição Normal padronizada.

Considere que a probabilidade da variável Z com valores entre $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$ é $(1-\alpha)$. Os valores $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$ são obtidos na tabela da distribuição Normal. Assim o intervalo de confiança da média com variância conhecida é dado por,

$$IC(\mu\ ;\ 1-\alpha) = \left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right). \tag{4.5}$$

Por exemplo, se considerarmos $100(1-\alpha) = 0,95$, retirada uma amostra e encontrada sua média e, admitindo conhecer σ^2 podemos construir o intervalo,

$$\left(\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right). \tag{4.6}$$

Esse intervalo de confiança pode ou não conter o parâmetro μ mas pelo exposto acima temos 95% de confiança de que contenha, Figura 4.1.

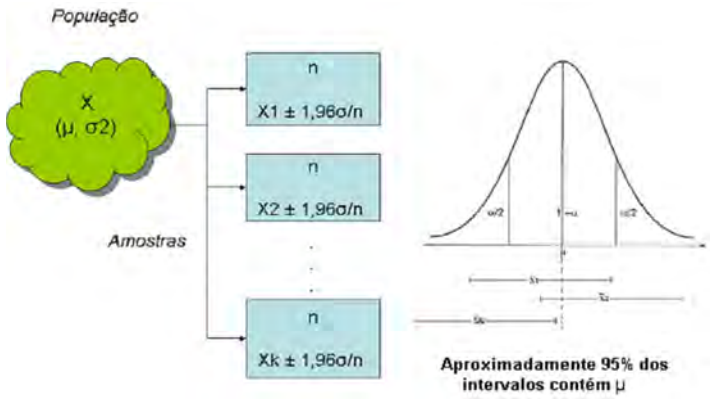


Figura 4.1: Esquema-Intervalo de confiança

Na Figura 4.1 $X1$ é a média da amostra $X1$, $X2$ é a média da amostra $X2, ..., Xk$ é a média da amostra Xk.

Para ilustrar o que foi dito acima, considere-se o seguinte experimento de simulação BUSSAB e MORETTIN [3]. Foram geradas 20 amostras de tamanho $n = 25$ de uma distribuição normal com média $\mu = 5$ e desvio padrão $\sigma = 3$. Para cada amostra foram construidos os intervalos de confiança para μ com coeficiente de confiança de 95%, utilizando a equação . Na Figura 4.2 tem-se esses intervalos representados e notamos que três deles não contêm a média $\mu = 5$.

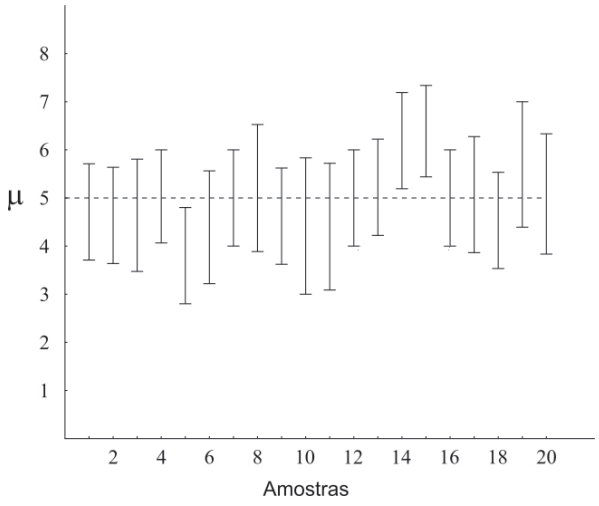


Figura 4.2: Intervalos de confiança para a média de uma N(5 , 9), para 20 amostras de tamanho n=25

No caso em que a variância populacional é desconhecida, utiliza-se a variância amostral S^2 como estimativa de σ^2. Assim,

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t - Student_{(n-1)},$$

A variável T segue distribuição t de Student com $n-1$ graus de liberdade.

Então, ao fixar-se o coeficiente de confiança em $(1-\alpha)$, obtém-se da Tabela da distribuição $t-Student$, o valor $t_{\left((n-1)\ ,\ \frac{\alpha}{2}\right)}$. Logo, o intervalo com $100(1-\alpha)\%$ de confiança para μ, com variância desconhecida, é dado por:

$$IC\left(\mu\ ;\ 1-\alpha\right)=\left(\overline{X}-t_{(n-1),\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\ ,\ \overline{X}+t_{(n-1),\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right). \tag{4.7}$$

4.1.2 Intervalo de confiança para a variância

Considere-se uma amostra aleatória x=$(x_1, x_2, ..., x_n)$, de tamanho n de uma população com distribuição normal com média μ e variância σ^2. Um estimador para σ^2 é a variância amostral S^2.

Seja $x=(x_1, x_2, ..., x_n)$ uma amostra aleatória de uma população normal $X\sim N\left(\mu, \sigma^2\right)$ com média populacional μ e variância populacional σ^2 desconhecidas. Pode-se mostrar que $\dfrac{(n-1)\,S^2}{\sigma^2}\sim\chi^2_{n-1}$, onde χ^2_{n-1} é a distribuição qui-quadrado com $(n-1)$ graus de liberdade.

A partir deste resultado, obtém-se o intervalo com $100(1-\alpha)\%$ de confiança para a variância, dado por:

$$IC\left(\sigma^2\ ;\ 1-\alpha\right)=\left(\frac{(n-1)\,S^2}{Q_{1-\frac{\alpha}{2}}}\ ,\ \frac{(n-1)\,S^2}{Q_{\frac{\alpha}{2}}}\right). \tag{4.8}$$

sendo $Q_{1-\frac{\alpha}{2}}$ e $Q_{\frac{\alpha}{2}}$ obtidos na tabela da distribuição Qui-quadrado.

4.2 Intervalos de confiança *bootstrap*

O método *bootstrap* é uma técnica robusta na construção de um intervalo de confiança. A seguir apresenta-se os diferentes métodos de obtenção dos intervalos de confiança *bootstrap*, sendo eles o intervalo *bootstrap* padrão, o intervalo *bootstrap-t*, o intervalo *bootstrap* percentil, BCBP e o BC_a sendo os dois últimos uma adaptação do intervalo percentil para casos especiais. O cálculo do erro padrão *bootstrap* é utilizado nos intervalos de confiança *bootstrap* aproximados para um parâmetro de interesse θ. Utiliza-se em todos os intervalos de confiança *bootstrap* mencionados a notação $[L_I(x)\ ,\ L_S(x)]$ em que $L_I(x)$ é o limite inferior do intervalo e $L_S(x)$ é o limite superior do intervalo, conforme (4.1).

4.2.1 Intervalo *bootstrap* padrão

Como na maioria dos intervalos de confiança *bootstrap* usa-se o erro padrão *bootstrap* $\widehat{dp}_{boot}$, deve-se primeiramente encontrá-lo utilizando o código 1 que se encontra

no anexo A.1, (BORKOWSKI), [2]. Neste código, também encontra-se o valor do viés que é utilizado nos intervalos de confiança *bootstrap* padrão e *bootstrap-t* com viés ajustado. Considere X uma variável aleatória com distribuição F desconhecida, $x = (x_1, x_2, ..., x_n)$ uma amostra de X. A partir das B amostras *bootstrap* $x^* = (x_1^*, x_2^*, ..., x_B^*)$ calculamos as réplicas *bootstrap* $\widehat{\theta}_b^* = s(x^*)\ b = 1, 2, ..., B$ e depois encontramos $\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)$. Segundo EFRON e TIBSHIRANI [7] podemos definir a variável,

$$Z_b^* = \frac{\widehat{\theta} - \theta}{\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)} \sim N(0\ ,\ 1). \tag{4.9}$$

A variável Z_b^* segue a distribuição Normal padrão.

Adotando-se o coeficiente de confiança $100(1-\alpha)\,\%$ com $0 < \alpha < 1$ é possível obter o valor z_c na tabela da distribuição normal padrão que satisfaz,

$$P\left[-z_c \leq Z_b^* \leq +z_c\right] = (1-\alpha).$$

Na tabela z_c é o valor crítico tal que,

$$P(0 < Z_b^* < z_c) = (1-\alpha)/2.$$

Então pode-se escrever

$$P\left[-z_c \leq \frac{\widehat{\theta} - \theta}{\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)} \leq +z_c\right] = (1-\alpha),$$

o que implica,

$$P\left[\widehat{\theta} - z_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right) \leq \theta \leq \widehat{\theta} + z_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right] = (1-\alpha).$$

Logo, pode-se definir o intervalo de confiança *bootstrap* padrão conforme a definição 4.3.

Definição 4.3. *Considere* $\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)$*, a estimativa bootstrap do erro padrão de* $\widehat{\theta}^*$*. O intervalo de confiança bootstrap padrão para* θ *com coeficiente de confiança* $100(1-\alpha)\,\%$ *é dado por,*

$$\left[\widehat{\theta} - z_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\ ,\ \widehat{\theta} + z_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right]. \tag{4.10}$$

Uma vantagem desse método é a facilidade algébrica para obter o intervalo de confiança para θ.

Usando o fato dessa aproximação ser assintótica, o intervalo de confiança *bootstrap* padrão pode ser impreciso. Às vezes, a assimetria e o viés podem estar presentes na distribuição de Z.

Definição 4.4. *O intervalo de confiança bootstrap padrão com viés ajustado, com coeficiente de confiança* $100\,(1-\alpha)\,\%$ *é dado por,*

$$\left[\left(\widehat{\theta} - z_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right) - \widehat{vi\acute{e}s}_{boot}\left(\widehat{\theta}\right)\ ,\left(\widehat{\theta} + z_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right) - \widehat{vi\acute{e}s}_{boot}\left(\widehat{\theta}\right)\right]. \qquad (4.11)$$

4.2.2 Intervalo *bootstrap-t*

Segundo EFRON e TIBSHIRANI [6] e HALL [10] o procedimento *bootstrap-t* também é uma generalização do método t-Student, sendo em particular, aplicado às estatísticas de locação tais como a média amostral, a mediana ou ainda os percentis amostrais. Estes autores citam que pelo menos em sua forma tradicional, o método *bootstrap-t* não é uma boa construção de intervalos para outras estatísticas, como por exemplo, o coeficiente de correlação.

Da mesma forma que o intervalo *bootstrap* padrão pode-se encontrar o intervalo de confiança *bootstrap-t* e o intervalo *bootstrap-t* com viés ajustado, ambos utiliza-se o erro padrão *bootstrap* $\widehat{dp}_{boot}\left(\widehat{\theta}\right)$.

Quando $\widehat{\theta}$ é a média amostral, Gosset (1876-1937) mostrou que a expressão (4.9) tem uma aproximação $t-Student$

$$T_b^* = \frac{\widehat{\theta} - \theta}{\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)} \sim t_{n-1}. \qquad (4.12)$$

em que t_{n-1} representa a distribuição $t-Student$ com $(n-1)$ graus de liberdade.

O lado direito da expressão (4.12) é interpretado como uma quantidade pivotal aproximada. Note-se que em seu denominador tem-se um estimador do desvio padrão de $\widehat{\theta}_b^*$ para a amostra bootstrap x^*. Caso não tivéssemos em mãos esse estimador, um novo esquema de reamostragem seria necessário para estimar o erro padrão, para cada uma das B amostras *bootstrap*. Neste caso, o número de amostras *bootstrap* necessárias para a construção de intervalos de confiança *bootstrap-t* seria dado pelo produto de B pelo número de novas amostras *bootstrap* requeridas para cada estimação de $\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)$. Em alguns casos esse número de amostras requer um poder computacional muito alto. Para Karlsson & Löthgren (2000) (apud MARTINEZ e LOUZADA-NETO), [14] este processo recebe o nome de "douple *bootstrap*".

Ao fixar-se o coeficiente de confiança em $100\,(1-\alpha)\,\%$ com $0 < \alpha < 1$, obtém-se na tabela da distribuição $t - Student$ o valor t_c que satisfaz,

$$P\left[-t_c \leq T_b^* \leq +t_c\right] = (1-\alpha)\,.$$

Este valor é encontrado na tabela considerando $(n-1)$ graus de liberdade e consultando a coluna $100\alpha\%$.

Então pode-se escrever,

$$P\left[-t_c \leq \frac{\widehat{\theta} - \theta}{\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)} \leq +t_c\right] = (1-\alpha)\,,$$

o que implica

$$P\left[\widehat{\theta} - t_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right) \leq \theta \leq \widehat{\theta} + t_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right] = (1-\alpha)\,.$$

Assim, pode-se definir o intervalo de confiança *bootstrap*-t conforme a definição 4.5.

Definição 4.5. *Considere* $\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)$*, a estimativa bootstrap do erro padrão de* $\widehat{\theta}$ *. O intervalo de confiança bootstrap-t para* θ *com coeficiente de confiança* $100\,(1-\alpha)\,\%$ *é dado por,*

$$\left[\widehat{\theta} - t_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\ ,\ \widehat{\theta} + t_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right]. \tag{4.13}$$

Definição 4.6. *O intervalo de confiança bootstrap-t com viés ajustado com coeficiente de confiança* $100\,(1-\alpha)\,\%$ *é dado por,*

$$\left[\left(\widehat{\theta} - t_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right) - \widehat{vi\acute{e}s}_{boot}\left(\widehat{\theta}\right)\ ,\ \left(\widehat{\theta} + t_c\widehat{dp}_{boot}\left(\widehat{\theta}^*\right)\right) - \widehat{vi\acute{e}s}_{boot}\left(\widehat{\theta}\right)\right]. \tag{4.14}$$

O intervalo *bootstrap-t* funciona bem quando a distribuição da estatística é aproximadamente normal e a estatística apresenta viés pequeno. Respeitando essas restrições o intervalo de confiança *bootstrap-t* pode ser calculado na estimação de diversos parâmetros, tais como a média populacional, a proporção e a variância.

Há uma tendência geral para que os intervalos *bootstrap-t* tenham amplitudes menores do que os intervalos baseados na tabela normal. No entanto, este ganho de precisão faz-se à custo de uma perda de generalidade, como já acontece em certa medida, quando se aplicam as tabelas da distribuição t de Student relativamente a normal.

4.2.3 Intervalo *bootstrap* percentil

Será descrita uma outra abordagem para o intervalo de confiança *bootstrap*, baseada em percentis de uma estatística *bootstrap*. A ideia dos intervalos de confiança percentil é muito simples e se baseia na utilização dos percentis do histograma *bootstrap* para definir os limites de confiança. No intervalo de confiança *bootstrap* percentil do tipo 1 utiliza-se o percentil $\left(\frac{\alpha}{2}\right)100\%$ e o percentil $\left(1-\frac{\alpha}{2}\right)100\%$ das estatísticas amostrais no *bootstrap*.

Considere X uma v.a. com distribuição F desconhecida e $x=(x_1,x_2,...,x_n)$ uma amostra de X. Encontradas as B amostras *bootstrap* $x^*=(x_1^*,x_2^*,...,x_n^*)$ e consequentemente as réplicas *bootstrap* $\widehat{\theta}_b^*=s\left(x^*\right)$ $b=1,2,...,B$, pode-se definir o intervalo *bootstrap* percentil do tipo 1 da seguinte forma:

Definição 4.7. *O intervalo de confiança bootstrap percentil do tipo 1 com coeficiente de confiança* $100\left(1-\alpha\right)\%$ *é obtido pelos* $\left(\frac{\alpha}{2}\right)-$*ésimo e* $\left(1-\frac{\alpha}{2}\right)-$*ésimo percentis de* $\widehat{F}$*, em que* $\widehat{F}$ *é a função de distribuição empírica de* $\widehat{\theta}^*$*.*

Uma expressão para esse intervalo de confiança é:

$$\left[L_I\left(x\right)\ ,\ L_S\left(x\right)\right]=\left[\widehat{F}^{-1}_{\left(\frac{\alpha}{2}\right)}\ ,\ \widehat{F}^{-1}_{\left(1-\frac{\alpha}{2}\right)}\right]. \tag{4.15}$$

Como $\widehat{F}^{-1}_{\left(\frac{\alpha}{2}\right)}=\widehat{\theta}^*_{\left(\frac{\alpha}{2}\right)}$, representa o percentil de ordem $\left(\frac{\alpha}{2}\right)$ da distribuição *bootstrap*, pode-se reescrever a expressão (4.15) da seguinte forma,

$$\left[L_I\left(x\right)\ ,\ L_S\left(x\right)\right]=\left[\widehat{\theta}^*_{\left(\frac{\alpha}{2}\right)}\ ,\ \widehat{\theta}^*_{\left(1-\frac{\alpha}{2}\right)}\right]. \tag{4.16}$$

Na prática utiliza-se um número finito B de replicações, gerando as amostras $x_1^*,x_2^*,...,x_n^*$ e a partir destas calcula-se as replicações *bootstrap* $\widehat{\theta}_b^*=s\left(x_b^*\right)$, $b=1,2,...,B$.

Um exemplo para entendimento dos detalhes deste método *bootstrap* percentil, encontra-se no Anexo B.

A distribuição *bootstrap* da média $\widehat{\theta}^*$ aproxima-se da distribuição normal quando $n\rightarrow\infty$. Nesse caso, os intervalos de confiança *bootstrap* percentil e o intervalo de confiança *bootstrap* padrão serão semelhantes. Quando n é reduzido, o histograma *bootstrap* pode afastar-se da normalidade e os dois tipos de intervalos serão diferentes. O intervalo *bootstrap* percentil estima automaticamente as transformações e faz a sua incorporação direta no cálculo dos intervalos. EFRON e TIBSHIRANI [7] apresentam o exemplo em que se pretende estimar o parâmetro $\theta=e^{\mu}$ a partir de $\widehat{\theta}=e^{\overline{x}}$. O

intervalo de confiança padrão para θ terá problemas de simetria e cobertura. Neste caso pode ser feita uma transformação $\widehat{\phi} = \log \widehat{\theta}$ e em seguida, o calculo do intervalo de confiança padrão. Depois, remeter os valores para a escala original dos valores de θ. Assim os pontos críticos do intervalo de confiança de θ serão deslocados para a direita. O método percentil fornece diretamente um intervalo de confiança preciso para θ e o histograma obtido é próximo da distribuição exponencial. Aplica-se o mesmo intervalo percentil para ϕ e obtém-se um histograma próximo da distribuição normal. Neste sentido, o intervalo percentil pode ser interpretado como um algoritmo que introduz automaticamente as transformações necessárias no cálculo dos intervalos de confiança, o que permite uma utilização mais geral e eficaz, dado que não necessitamos conhecer a transformação apropriada a utilizar. Isto nos leva ao seguinte lema, EFRON e TIBSHIRANI [7].

Lema 4.1. *Suponha a transformação* $\widehat{\phi} = m\left(\widehat{\theta}\right)$ *que normaliza a distribuição de* $\widehat{\theta}$*:*

$$\widehat{\phi} \sim N\left(\phi\ ,\ c^2\right),$$

para toda a escolha de $\widehat{F}$*, o mecanismo de probalidade. Então o intervalo percentil baseado em* $\widehat{\phi}$ *é,*

$$\left[m^{-1}(\widehat{\phi} - z_{\left(\frac{\alpha}{2}\right)}c)\ ,\ m^{-1}\left(\widehat{\phi} - z_{\left(1-\frac{\alpha}{2}\right)}c\right)\right]. \tag{4.17}$$

Uma ótima propriedade do intervalo construído pelo método percentil é a invariância de transformações monótonas descrita na seção 4.3.

Para o Intervalo *bootstrap* percentil do tipo 2, utiliza-se os percentis das diferenças dos valores da estatísticas das reamostragens em relação ao valor médio desta mesma estatística. Considere X uma v.a. com distribuição F desconhecida e $x = (x_1, x_2, ..., x_n)$ uma amostra de X. Encontra-se as B amostras *bootstrap* $x^* = (x_1^*, x_2^*, ..., x_n^*)$ e consequentemente as réplicas *bootstrap* $\widehat{\theta}_b^* = s\left(x^*\right)$, $b = 1, 2, ..., B$. Em seguida calcula-se a média das réplicas *bootstrap* que denotaremos por $\overline{\widehat{\theta}_b^*}$. Após, encontra-se, para cada réplica *bootstrap*, $\widehat{\theta}_b^* = s\left(x^*\right)$, $b = 1, 2, ..., B$ a diferença $\triangle_b^*$ entre o valor de cada estimativa da réplica *bootstrap* e a média de todas as estimativas *bootstrap* $\overline{\widehat{\theta}_b^*}$, ou seja,

$$\triangle_b^* = \widehat{\theta}_b^* - \overline{\widehat{\theta}_b^*}. \tag{4.18}$$

Após este cálculo, essas diferenças são ordenadas de forma crescente, obtendo-se B diferenças. Assim pode-se encontrar o percentil desejado para as diferenças de maneira análoga ao tipo 1.

Definição 4.8. *O intervalo de confiança bootstrap percentil do tipo 2 com coeficiente de confiança* $100\,(1-\alpha)\,\%$ *é obtido pelos percentis das diferenças dos valores das estatísticas nas réplicas bootstrap, com sua média.*

Uma expressão para esse intervalo de confiança é:

$$[L_I\,(x)\ ,\ L_S\,(x)] = \left[\triangle^*_{\left(\frac{\alpha}{2}\right)\%}\ ,\ \triangle^*_{\left(1-\frac{\alpha}{2}\right)\%}\right]. \tag{4.19}$$

Por exemplo: Para um intervalo de confiança de 95%, encontram-se os percentis de 2,5% e 97.5% das diferenças e calcula-se o intervalo de confiança *bootstrap* percentil da seguinte forma:

$$[L_I\,(x)\ ,\ L_S\,(x)] = \left[\triangle^*_{2,5\%}\ ,\ \triangle^*_{97,5\%}\right]. \tag{4.20}$$

Para verificar se o intervalo de confiança t calculado é confiável, podemos compará-lo com o intervalo de confiança percentil. Se o viés for pequeno e a distribuição *bootstrap* for aproximadamente normal, os dois intervalos irão apresentar valores muito próximos. Segundo HESTERBERG [12], caso os intervalos de confiança *bootstrap* calculados pela t e pelo percentil não tiverem valores próximos, nenhum destes métodos deve ser utilizado. Entretanto EFRON e TIBSHIRANI [7], afirma que se a distribuição *bootstrap* não for aproximadamente normal, mas se existir uma transformação monótona possível que a torne normal, pode-se calcular o intervalo *bootstrap* percentil para os dados transformados e posteriormente desfazer a transformação para os limites do intervalo encontrado. Isto é possível uma vez que a transformação utilizada é uma transformação monótona, portanto o intervalo de confiança *bootstrap* pelo método percentil assim calculado coincide com o intervalo de confiança *bootstrap* pelo método percentil para os dados transformados.

Segundo EFRON e TIBSHIRANI [7], se o viés e a assimetria estão presentes de forma muito forte é mais recomendável que se utilize os métodos de *bootstrap* de correção como o método BCPB e o método BC_a .

4.2.4 Intervalo *bootstrap* BCPB

O intervalo de Confiança Percentil Corrigido em Relação ao Viés (BCPB) "Biased-Corrected Percentilt Bootstrap" segundo RIZZO e CYMROT [23], é um método *bootstrap* que faz correções substanciais. Neste método os extremos do intervalo serão os percentis da distribuição *bootstrap* ajustados, para corrigir o viés e a assimetria da distribuição. Por exemplo, para encontrar um intervalo de confiança BCPB com 95% de confiança, é preciso ajustar os percentis , que para um cálculo de intervalo de confiança Percentil seriam 2,5% e 97,5%, como ilustrado no anexo B. No intervalo de confiança

bootstrap BCPB serão outros valores de percentis. O intervalo de confiança BCPB possui as seguintes propriedades:

Propriedade 4.1. Se a estatística for viesada "para cima" o BCPB move os extremos para a esquerda.

Propriedade 4.2. Se a estatística for viesada "para baixo" o BCPB move os extremos para a direita.

Considere X uma v.a. com distribuição F desconhecida e $x = (x_1, x_2, ..., x_n)$ uma amostra de X. Encontradas as B amostras *bootstrap* $x^* = (x_1^*, x_2^*, ..., x_n^*)$ e as réplicas *bootstrap* $\widehat{\theta}_b^* = s(x^*)$, $b = 1, 2, ..., B$, deve-se ordenar as B estimativas $\widehat{\theta}_b^* = s(x^*)$, $b = 1, 2, ..., B$, de forma crescente. Então encontra-se a proporção das réplicas *bootstrap*, menores que $\widehat{\theta}$, denotada por p_0.

Definição 4.9. *A proporção p_0 é definida como a probabilidade de uma estimativa ser inferior à estimativa da amostra original $\widehat{\theta}$.*

Uma expressão para a proporção p_0 das réplicas *bootstrap* é dada por,

$$p_0 = P\left[\widehat{\theta}_b^* \leq \widehat{\theta}\right], \qquad b = 1, 2, ..., B. \tag{4.21}$$

Tendo o valor de p_0 em mãos, encontra-se o parâmetro de correção do viés que é denotado por z_0.

Definição 4.10. *O parâmetro de correção do viés z_0 é a inversa da normal padrão aplicada no ponto p_0 de modo que*

$$z_0 = \Phi^{-1}(p_0). \tag{4.22}$$

Assim tem-se B valores de z_0 e em seguida calcula-se a média para esses valores obtendo um único valor, denotado por $\overline{z_0}$.

Selecionado um coeficiente de confiança $(1 - \alpha)\,100\%$ para a estimativa do parâmetro e encontrado $z_{\left(\frac{\alpha}{2}\right)}$ obtém-se as correções para os percentis.

Definição 4.11. *As correções P_I e P_S são definidas por,*

$$P_I = \Phi\left(2\overline{z_0} - z_{\left(\frac{\alpha}{2}\right)}\right), \qquad P_S = \Phi\left(2\overline{z_0} + z_{\left(\frac{\alpha}{2}\right)}\right). \tag{4.23}$$

Definição 4.12. *O intervalo de confiança bootstrap BCPB é definido por:*

$$[L_I(x)\ ,\ L_S(x)] = \left(\widehat{\theta}^*{}_{(PI)}\ ,\ \widehat{\theta}^*{}_{(PS)}\right). \tag{4.24}$$

4.2.5 Intervalo *bootstrap* BCa

O método de Correção do Viés Acelerado BC_a ("Bias-Corrected and Acceleration") permite encontrar o intervalo de confiança quando a assimetria estiver presente de maneira muito forte. Esta metodologia de construção de intervalos de confiança constitui um aperfeiçoamento na concepção dos intervalos de confiança *bootstrap*, EFRON e TIBSHIRANI [7].

Tal método permite ultrapassar e sanar alguns problemas colocados pelos intervalos de confiança *bootstrap-t* e percentil quando a distribuição é bem assimétrica. No entanto, os intervalos de confiança BC_a exigem um volume de cálculo elevado, fator que constitui um grande inconveniente para este método.

Assim como nos casos dos intervalos de confiança percentil e BCPB, o BC_a utiliza os percentis da distribuição *bootstrap* para a construção dos intervalos de confiança para parâmetros de interesse. Neste caso utiliza-se percentis que dependem de duas constantes: $\widehat{z}_0$ denominada correção para a tendência e $\widehat{a}$ denominada constante de aceleração, que ajusta o intervalo de confiança em relação a assimetria. A constante $\widehat{z}_0$ representa a magnitude da tendenciosidade mediana de $\widehat{\theta}^*$ ou seja a discrepância entre a mediana de $\widehat{\theta}^*$e $\widehat{\theta}$. Já a constante de aceleração $\widehat{a}$ representa a taxa de variação do desvio padrão de $\widehat{\theta}$ relativamente ao verdadeiro valor do parâmetro θ. Segundo EFRON e TIBSHIRANI [7] o método BC_a é mais indicado que é o método BCPB. Antes de definir o intervalo *bootstrap* BC_a, vamos definir as constantes $\widehat{z}_0$ e $\widehat{a}$.

Definição 4.13. *O valor de $\widehat{z}_0$ que representa o enviesamento mediano de $\widehat{\theta}^*$, isto é, o desvio entre a mediana dos valores de $\widehat{\theta}^*$ e o valor obtido para $\widehat{\theta}$ em unidades normais, é calculado diretamente a partir da proporção das réplicas bootstrap inferiores à estimativa original $\widehat{\theta}$, ou seja,*

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\widehat{\theta}_b^* < \widehat{\theta}\right\}}{B}\right), \tag{4.25}$$

onde Φ^{-1} é a inversa da função de distribuição acumulada normal padrão.

Se a metade das medianas de $\widehat{\theta}_b^* = s\left(x^*\right)$, $b = 1, 2, ..., B$ forem menores ou iguais a $\widehat{\theta}$ obtemos $\widehat{z}_0 = 0$.

A constante de aceleração $\widehat{a}$ diz respeito à taxa de variação do desvio padrão de $\widehat{\theta}$ relativamente ao verdadeiro valor do parâmetro θ. Essa constante corrige a hipótese não realista exigida pela aproximação normal $\widehat{\theta} \sim N\left(\theta\ ,\ dp^2\left(\widehat{\theta}\right)\right)$, segundo a qual o desvio padrão de $\widehat{\theta}$ é igual para qualquer θ. Existem várias maneiras de se obter a constante de aceleração. EFRON e TIBSHIRANI [7] propõe o cálculo de $\widehat{a}$ em termos de valores *jackknife* de uma estatística $\widehat{\theta} = s\left(x\right)$.

Definição 4.14. *Seja $x^{(i)}$ uma amostra jackknife (após eliminar a observação i) e $\widehat{\theta}_{(i)} = s\left(x^{i}\right)$. O fator de aceleração $\widehat{a}$ é obtido por*

$$\widehat{a} = \frac{\sum_{i=1}^{n}\left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\right)^{3}}{6\left[\sum_{i=1}^{n}\left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\right)^{2}\right]^{\frac{3}{2}}}, \tag{4.26}$$

onde $\widehat{\theta}_{(\cdot)} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)}$.

Considere X uma v.a. com distribuição F desconhecida e $x = (x_1, x_2, ..., x_n)$ uma amostra de X. Encontradas as B amostras *bootstrap*, $x^* = (x_1^*, x_2^*, ..., x_n^*)$ e as réplicas *bootstrap* $\widehat{\theta}_b^* = s\left(x^*\right)$, $b = 1, 2, ..., B$, pode-se encontrar o intervalo de confiança BC_a utilizando a definição 4.15.

Definição 4.15. *O intervalo de confiança bootstrap BC_a com coeficiente de confiança de $(1-\alpha)\,100\%$ é dado por,*

$$\left[L_I\left(x\right)\ ,\ L_S\left(x\right)\right] = \left(\widehat{\theta^*}_{(PI)}\ ,\ \widehat{\theta^*}_{(PS)}\right), \tag{4.27}$$

onde

$$P_I = \phi\left(\widehat{z}_0 - \frac{\widehat{z}_0 + z_{\left(\frac{\alpha}{2}\right)}}{1 - \widehat{a}\left(\widehat{z}_0 + z_{\left(\frac{\alpha}{2}\right)}\right)}\right),$$

$$P_S = \phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{\left(1-\frac{\alpha}{2}\right)}}{1 - \widehat{a}\left(\widehat{z}_0 + z_{\left(1-\frac{\alpha}{2}\right)}\right)}\right),$$

sendo $\phi(\cdot)$ função de distribuição padrão normal e $z_{\left(\frac{\alpha}{2}\right)}$ é o percentil de ordem $100z_{\left(\frac{\alpha}{2}\right)}$ da distribuição normal.

4.3 Propriedades dos intervalos de confiança

Nesta seção, são descritas as propriedades dos intervalos de confiança tais como: a preservação da amplitude, a acurácia de ordem 1 e 2, a corretividade de ordem 1 e 2 e a propriedade de invariância a transformações monótonas, descrevendo seu comportamento diante dos intervalos de confiança *bootstrap*.

A propriedade de preservação da amplitude respeita as restrições aos valores possíveis para um parâmetro. Os métodos que satisfazem essa propriedade são geralmente mais precisos e confiáveis.

Propriedade 4.3. Preservação da Amplitude (PA): Considere o parâmetro θ, que assume valores em um conjunto A, de forma que $A = [a\ ,\ b]$, tal que $A = [a\ ,\ b] \subset \mathbb{R}$, com $a < b$. Diz-se que o intervalo de confiança para θ possui a propriedade de preservação da amplitude se seus limites inferior e superior estão contidos em A. Ou seja,

$$[L_I(x), L_S(x)] \subset A. \tag{4.28}$$

Exemplo 4.1. O coeficiente de correlação ρ assume valores no intervalo $[-1, 1]$. Espera-se que um intervalo de confiança para ρ, construído através de uma amostra aleatória (a.a), tenha limite inferior não menor que -1 e limite superior não maior que 1. Ou seja, deseja-se obter,

$$[L_I(x), L_S(x)] \subset [-1, 1].$$

Os conceitos de acurácia e corretividade mencionados a seguir, são descritos por EFRON e TIBSHIRANI [7]. Por simplicidade, segundo MARTINEZ e LOUZADA-NETO [14], para definir tais propriedades, utiliza-se um intervalo de confiança unilateral para um parâmetro de interesse θ, em que o único limite de confiança é dado por $\widehat{\theta}_{(\gamma)}$, onde γ é a probabilidade de cobertura deste intervalo. Ou seja,

$$P\left[\theta \leq \widehat{\theta}_{(\gamma)}\right] \approx \gamma, \forall\gamma.$$

A noção de acurácia está relacionada com a proximidade entre as probabilidades de cobertura dos intervalos $(-\infty\ ,\ \widehat{\theta}_{(\gamma)}]$ e $[\widehat{\theta}_{(\gamma)}\ ,\ +\infty)$.

Propriedade 4.4. Acurácia de Ordem 1: Um limite de confiança é acurado de primeira ordem se,

$$P\left[\theta \leq \widehat{\theta}_{(\gamma)}\right] = \gamma + O\left(\frac{1}{\sqrt{n}}\right). \tag{4.29}$$

Propriedade 4.5. Acurácia de Ordem 2: Um limite de confiança é acurado de segunda ordem se,

$$P\left[\theta \leq \widehat{\theta}_{(\gamma)}\right] = \gamma + O\left(\frac{1}{n}\right). \tag{4.30}$$

O termo $O(\cdot)$ denota a ordem da magnitude de uma sequência $\{a_n\}_{n\geq 1}$ de números reais, sendo que $a_n = O\{b_n\}$ indica existir um número real $k > 0$ e um número positivo inteiro $n_0 = n_0(K)$ tal que $|a_n \setminus b_n| \leq K$, para todo $n \geq n_0$ e $\{b_n\}_{n\geq 1}$ uma sequência de números reais.

A propriedade (4.4) estabelece que os erros dos intervalos de ordem 1 tendem a zero, a medida que n tende para o infinito, a uma taxa de $\frac{1}{\sqrt{n}}$. A propriedade (4.5) estabelece que os erros dos intervalos de ordem 2 tendem a zero, a medida que n tende para o infinito, a uma taxa de $\frac{1}{n}$. Uma comparação destas duas taxas fornece não somente a magnitude do erro que se comete nos intervalos, mas também, um critério de escolha entre eles, visto que a segunda taxa converge para zero bem mais rapidamente que a primeira.

A noção de corretividade se refere à proximidade de um limite de confiança aproximado para um limite de confiança exato.

Propriedade 4.6. Corretividade de ordem 1: Um limite de confiança é correto de primeira ordem se,

$$\widehat{\theta}_{(\gamma)} = \widehat{\theta}_{(\gamma)}^{EXATO} + O_p\left(\frac{1}{n}\right). \tag{4.31}$$

onde $\widehat{\theta}_{(\gamma)}^{EXATO}$ é o ponto de confiança exato que satisfaz, $P\left[\theta \leq \widehat{\theta}_{(\gamma)}\right] = \gamma$.

Propriedade 4.7. Corretividade de ordem 2: Um limite de confiança é correto de segunda ordem se,

$$\widehat{\theta}_{(\gamma)} = \widehat{\theta}_{(\gamma)}^{EXATO} + O_p\left(\frac{1}{\sqrt{n^3}}\right). \tag{4.32}$$

onde $\widehat{\theta}_{(\gamma)}^{EXATO}$ é o ponto de confiança exato que satisfaz, $P\left[\theta \leq \widehat{\theta}_{(\gamma)}\right] = \gamma$.

Se $X_n = O_p(a_n)$ e $Y_n = O_p(b_n)$, onde $\{a_n\}_{n\geq 1}$ e $\{b_n\}_{n\geq 1}$ são duas sequências de números reais (ou variáveis aleatórias), tem-se que $X_n\, Y_n = O_p(a_n b_n)$. Logo as expressões (4.31) e (4.32) são equivalentes a dizer que, o limite de confiança $\widehat{\theta}_{(\gamma)}$ é correto de primeira ordem se,

$$\widehat{\theta}_{(\gamma)} = \widehat{\theta}_{(\gamma)}^{EXATO} + O_p\left(\frac{1}{\sqrt{n}}\right)\widehat{\sigma}, \tag{4.33}$$

e correto de segunda ordem se,

$$widehat\theta_{(\gamma)} = \widehat{\theta}_{(\gamma)}^{EXATO} + O_p\left(\frac{1}{n}\right)\widehat{\sigma}, \tag{4.34}$$

onde $\widehat{\sigma}$ é um plausível estimador do erro padrão de $\widehat{\theta}$ que usualmente é de ordem $\frac{1}{\sqrt{n}}$.

A corretividade de uma determinada ordem implica acurácia da mesma ordem.

Propriedade 4.8. Invariância a Transformações Monótonas (ITM): Um intervalo de confiança para um parâmetro θ possui a propriedade de invariância a transformações monótonas se o intervalo obtido para um novo parâmetro $\phi = g(\theta)$, onde $g(\cdot)$ é uma transformação monótona, corresponde ao intervalo para θ transformado por $g(\theta)$.

Se $g(\cdot)$ é monótona crescente então o intervalo para o novo parâmetro é dado por,

$$[L_I(x), L_S(x)] = \left[g\left(\widehat{\theta}^*_{\left(\frac{\alpha}{2}\right)}\right), g\left(\widehat{\theta}^*_{\left(1-\frac{\alpha}{2}\right)}\right)\right]. \tag{4.35}$$

Se $g(\cdot)$ é monótona decrescente então o intervalo para o novo parâmetro é dado por,

$$[L_I(x), L_S(x)] = \left[g\left(\widehat{\theta}^*_{\left(1-\frac{\alpha}{2}\right)}\right), g\left(\widehat{\theta}^*_{\left(\frac{\alpha}{2}\right)}\right)\right]. \tag{4.36}$$

Apresenta-se na Tabela 4.1 a informação que resume o comportamento de cada método *bootstrap* no que diz respeito a cada uma das propriedades estudadas acima. Nesta tabela abreviamos "não necessariamente"por NC.

Tabela 4.1: Propriedades dos intervalos de confiança *bootstrap*

I.C	P.A	Acur.	Corr.	I.T.M
boot. padrão	NC	$1^o ordem$	$1^o ordem$	NC
boot.-t	NC	$2^o ordem$	$2^o ordem$	NC
percentil	SIM	$1^o ordem$	$1^o ordem$	SIM
boot. BCPB	SIM	$2^o ordem$	$2^o ordem$	SIM
boot. BC_a	SIM	$2^o ordem$	$2^o ordem$	SIM

5 Aplicações dos métodos *bootstrap*

Neste capítulo, são descritos exemplos dos métodos *bootstrap* aplicados em 4 dados simulados e um exemplo a dados reais. É realizado uma comparação dos intervalos de confiança *bootstrap* com os intervalos de confiança usuais e também é verificado, dentre os intervalos de confiança *bootstrap*, qual apresenta o melhor resultado.

5.1 Exemplos - Comparação dos intervalos de confiança

Na maioria dos casos, como visto no capítulo 4, há necessidade de utilizar resultados assintóticos para o cálculo dos intervalos de confiança (por exemplo, a normalidade assintótica dos estimadores de máxima verossimilhança) e muitas vezes as amostras utilizadas não são de tamanho suficiente para o uso destes resultados. Como consequência disto, pode-se obter intervalos de confiança muito amplos e, em alguns casos em que há maior complexidade do modelo, o intervalo de confiança fora do domínio do parâmetro. Veja o exemplo a seguir.

Exemplo 5.1. Considere 10 conjuntos de dados gerados com $X \sim bino\,(n\ ,\ p)$, usando 5 tamanhos de amostras: $n = 10$, $n = 12$, $n = 15$, $n = 25$ e $n = 50$ e dois valores para a probabilidade de sucesso $p = 0,2$ ou $p = 0,8$. A Tabela 5.1 ilustra os dez conjuntos de dados gerados, o estimador de máxima verossimilhança (EMV) para o parâmetro p e o respectivo intervalo de confiança usual para p com 95% de confiança.

Analisando a Tabela 5.1, percebe-se que quanto maior é o valor de n, o EMV para p aproxima-se mais do verdadeiro valor de p. Neste exemplo, nem sempre os intervalos de confiança estarão fora do domínio do parâmetro, mas com n pequeno isso acontece muitas vezes. Com n relativamente pequeno, o resultado assintótico (usual) leva a intervalos de confiança fora do domínio do parâmetro, uma vez que p é probabilidade, $p \in [0, 1]$. E ainda tem-se que os intervalos de confiança estão muito amplos. Uma alternativa para esse caso é a utilização das técnicas *bootstrap*.

Tabela 5.1: Parâmetros dos dados gerados, EMV para o parâmetro p e respectivo IC assintótico

Amostras	Parâmetros	EMV(p)	IC(p , 95%)
Dados 1:	p=0,2 e n=10	0,1200	$[-0,0814\ ,\ 0,3214]$
Dados 2:	p=0,8 e n=10	0,8100	$[\ 0,5668\ ,\ 1,0532]$
Dados 3:	p=0,2 e n=12	0,1528	$[-0,1301\ ,\ 04357]$
Dados 4:	p=0,8 e n=12	0,8542	$[\ 0,5713\ ,\ 1,1371]$
Dados 5:	p=0,2 e n=15	0,1511	$[-0,1019\ ,\ 0,4041]$
Dados 6	p=0,8 e n=15	0,8578	$[\ 0.6047\ ,\ 1,1108]$
Dados 7:	p=0,2 e n=25	0,1872	$[-0,0088\ ,\ 0,3832]$
Dados 8:	p=0,8 e n=25	0,8240	$[\ 0,6280\ ,\ 1,0200]$
Dados 9:	p=0,2 e n=50	0,1992	$[\ 0,0885\ ,\ 0,3099]$
Dados10:	p=0,8 e n=50	0,8060	$[\ 0,6964\ ,\ 0,9156]$

Para mostrar que os intervalos de confiança *bootstrap* preservam a amplitude dos intervalos, geramos novamente uma amostra binomial $X \sim bin(n,p)$, com paramêtros $n = 10$ e $p = 0,2$ e obtivemos os resultados demonstrados na Tabela 5.2.

Tabela 5.2: Intervalos de confiança assintótico e *bootstrap*

Método	$IC\,(p\ ,\ 95\%)$
Assintótico	$(-0,0628\ ,\ 0,4028)$
boot padrão	$(0,0840\ ,\ 0,2560)$
boot-t	$(0,0708\ ,\ 0,2692)$
boot percentil	$(0,1000\ ,\ 0,2400)$
BCBP	$(0,1000\ ,\ 0,2800)$

Pensando em uma forma fácil e rápida de implementação do método *bootstrap*, ou seja, obter as amostras *bootstrap*, as réplicas *bootstrap* e também encontrar o erro padrão *bootstrap* e a estimativa *bootstrap* do viés, pode-se utilizar o código 1 que se encontra no anexo A.1. O código 1 foi elaborado por (BORKOWSKI) [2], para a implementação no software Matlab. Dado uma amostra original de tamanho n, o código 1 calcula a média, o desvio padrão, a variância e a mediana para essa amostra denotadas por $\widehat{\theta}\,(\cdot)$, em que $(\cdot)$ representa a estatística de interesse. O código dispoe as B amostras *bootstrap* e calcula as réplicas *bootstrap* $\widehat{\theta}^{*}\,(\cdot)$ tais como a média, o desvio padrão, a variância e a mediana de cada amostra *bootstrap* e, em seguida, calcula a média de cada estatística de interesse, o erro padrão *bootstrap*, $\widehat{dp}_{boot}$ e a estimativa *bootstrap*

do viés, $\widehat{vi\acute{e}s}_{boot}\left(\widehat{\theta}\right)$. Além disso, o código 1 reorganiza as réplicas *bootstrap* de forma crescente, juntamente com o percentil, ferramenta útil para os cálculos dos intervalos de confiança *bootstrap*. Os exemplos a seguir ilustram na prática a técnica *bootstrap*.

Exemplo 5.2. Foi gerado no software Matlab uma amostra de $X \sim bin(n\ ,\ p)$, binomial com parâme-tros $n = 10$ e $p = 0,2$, obtendo a amostra original $x = [4, 2, 0, 2, 4, 0, 1, 1, 1, 2]$. Aplicando o código 1, foram obtidas as informações listadas na Tabela 5.3.

Tabela 5.3: Aplicação do código 1, exemplo 5.2

O valor de $\widehat{\theta}(\cdot)$ para a amostra original.

$\widehat{\theta}(\cdot)$	Média	Desvio Padrão	Variância	Mediana
	1,7000	1,4181	2,0111	1,5000

As 40 amostras *bootstrap*.

Observação	1	2	3	4	5	6	7	8	9	10
Amostra Original	4	2	0	2	4	0	1	1	1	2
Amostra *bootstrap* 1	1	2	0	4	0	2	2	1	1	0
Amostra *bootstrap* 2	0	1	1	2	1	2	1	4	2	4
Amostra *bootstrap* 3	1	1	2	1	0	0	2	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Amostra *bootstrap* 40	2	2	2	2	2	4	4	2	4	1

As 40 réplicas *bootstrap*.

Amostra boot	Média	Desvio Padrão	Variância	Mediana
1	1,3000	1,2517	1,5667	1,0000
2	1,8000	1,3166	1,7333	1,5000
3	0,8000	0,7888	0,6222	1,0000
⋮	⋮	⋮	⋮	⋮
40	2,5000	1,0801	1,1667	2,0000

A média das B=40 réplicas bootstrap.

$\widehat{\theta}^*(\cdot)$	Média	Desvio Padrão	Variância	Mediana
	1,6225	1,2492	1,6447	1,4750

O erro padrão bootstrap.

$\widehat{dp}_{boot}\left(\widehat{\theta}\right)$	Média	Desvio Padrão	Variância	Mediana
	0,4233	0,2937	0,6953	0,4929

A estimativa bootstrap do viés.

$\widehat{vi\acute{e}s}_{boot}\left(\widehat{\theta}\right)$	Média	Desvio Padrão	Variância	Mediana
	-0,0775	-0,1689	-0,3664	-0,0250

Aplicando o código 2, que se encontra no anexo A.2, para os dados da amostra original do Exemplo 5.2, obtém-se os intervalos de confiança: *bootstrap* padrão, *bootstrap-t*, *bootstrap* Percentil, o intervalo de confiança Percentil Corrigido em relação ao Viés (BCPB) e o intervalo de confiança de Correção do Viés Acelerado (Biased Corrected Accelerated) BC_a, conforme as informações listadas na Tabela 5.4.

Tabela 5.4: Intervalos de confiança *bootstrap*, exemplo 5.2

IC boot		Média	Var	DP	Mediana
padrão		$(0,84\ ,\ 2,56)$	$(0,66\ ,\ 3,35)$	$(0,87\ ,\ 1,95)$	$(0,37\ ,\ 2,63)$
boot-t		$(0,78\ ,2,61)$	$(0,57\ ,\ 3,44)$	$(0,84\ ,\ 1,99)$	$(0,29\ ,\ 2,71)$
perc I		$(1,00\ ,\ 2,40)$	$(0,55\ ,\ 2,90)$	$(0,73\ ,\ 1,70)$	$(1,00\ ,\ 2,00)$
perc II		$(1,00\ ,\ 2,40)$	$(1,12\ ,\ 3,47)$	$(1,33\ ,\ 2,10)$	$(1,00\ ,\ 2,00)$
BCPB	p_0	0,4350	0,3650	0,3650	0,3650
	z_0	0,1637	0,3451	0,3451	0,3505
	P	$(0,0513\ ,\ 0,9889)$	$(0,1021\ ,\ 0,9960)$	$(0,1021\ ,\ 0,9960)$	$(0,1040\ ,\ 0,9961)$
	IC	$(1,00\ ,\ 2,80)$	$(0,84\ ,\ 3,43)$	$(0,91\ ,\ 1,85)$	$(1,00\ ,\ 4,00)$
BC_a	$\widehat{a}$	0,0293	0,0439	0,0439	0,0000
	P	$(0,0802\ ,\ 0,9954)$	$(0,1109\ ,\ 0,9973)$	$(0,1109\ ,\ 0,9973)$	$(0,1119\ ,\ 0,9974)$
	IC	$(1,00\ ,3,00)$	$(0,84\ ,\ 3,56)$	$(0,91\ ,\ 1,88)$	$(1,00\ ,\ 4,00)$

A seguir, fazemos uma comparação entre os intervalos de confiança *bootstrap*, e os intervalos de confiança usuais, destacando, se o tamanho da amostra n é pequeno (problema comum encontrado nos problemas biológicos) o objetivo é calcular a estimativa da variabilidade de forma mais precisa do que usando resultados assintóticos. Uma forma de verificar a qualidade dos intervalos de confiança é comparar as amplitudes destes, sendo que todos são calculados com a mesma probabilidade de cobertura $100(1-\alpha)\% = 95\%$. Quanto menor essa amplitude, melhor é o intervalo do ponto de vista prático.

Exemplo 5.3. Foi gerada no software Matlab uma amostra de $X \sim bin(n\ ,\ p)$, binomial com parâmetros $n = 10$ e $p = 0,4$. Com esses parâmetros sabemos que $E(X) = 4$, $V(X) = 2,4$, $dp(X) = 1,5492$ e $Md(X) = 3,5$. A amostra original é $x = [4,5,1,3,6,1,5,3,5,2]$. Aplicando o código 2 que se encontra no anexo A.2, obtém-se as informações listadas na Tabela 5.5.

Tabela 5.5: Intervalos de confiança *bootstrap*, exemplo 5.3

IC boot		Média	Var	DP	Mediana
padrão		$(2,44\ ;\ 4,56)$	$(1,52\ ;\ 4,81)$	$(1,27\ ;\ 2,29)$	$(1,68\ ;\ 5,32)$
boot-t		$(2,36\ ;4,63)$	$(1,41\ ;\ 4,92)$	$(1,23\ ;\ 2,32)$	$(1,56\ ;\ 5,43)$
perc I		$(2,60\ ;\ 4,30)$	$(1,37\ ;\ 4,10)$	$(1,17\ ;\ 2,02)$	$(2,00\ ;\ 5,00)$
perc II		$(2,70\ ;\ 4,40)$	$(2,23\ ;\ 4,95)$	$(1,53\ ;\ 2,38)$	$(2,00\ ;\ 5,00)$
BCPB	p_0	0,4600	0,3160	0,3160	0,4460
	z_0	0,0853	0,4789	0,4789	0,1358
	P	$(0,0368\ ;\ 0,9834)$	$(0,1581\ ;\ 0,9982)$	$(0,1581\ ;\ 0,9982)$	$(0,0457\ ;\ 0,9872)$
	IC	$(2,50\ ;\ 4,60)$	$(2,00\ ;\ 5,21)$	$(1,41\ ;\ 2,28)$	$(2,00\ ;\ 5,00)$
BC_a	$\widehat{a}$	-0,0099	0,0253	0,0272	0,0000
	P	$(0,0660\ ;\ 0,9928)$	$(0,1329\ ;\ 0,9978)$	$(0,1329\ ;\ 0,9978)$	$(0,0727\ ;\ 0,9938)$
	IC	$(2,70\ ;\ 4,70)$	$(1,87\ ;\ 5,11)$	$(1,37\ ;\ 2,26)$	$(2,50\ ;\ 5,00)$

O intervalo de confiança assintótico para a média do Exemplo 5.3 com 95% de confiança é $[2,3577\ ;\ 4,6423]$, com amplitude de 2,2846. Comparamos com os intervalos de confiança *bootstrap*, para a média, e constatamos que a amplitude de todos os intervalos de confiança *bootstrap* para a média são menores que o intervalo de confiança assintótico. O intervalo de confiança para a variância com 95% de confiança é $[1,3916\ ;\ 8,7690]$ com amplitude de 7,3774. Também comparando com os intervalos de confiança *bootstrap* para a variância vemos que todos os intervalos *bootstrap* para a variância são menores que a amplitude do intervalo de confiança assintótico da variância. A amplitude dos intervalos *bootstrap* em geral são menores que dos intervalos assintóticos, conforme a Tabela 5.6.

Tabela 5.6: Comparação das amplitudes dos IC assintótico e *bootstrap*, exemplo 5.3

Método	*IC*. Média	*A*	*IC*. Var	*A*
Assintótico	(2,3577 ; 4,6423)	**2,29**	(1,3916 ; 8,7690)	**7,37**
Boot-z	(2,4375 ; 4,5625)	**2,12**	(1,5242 ; 4,8091)	**3,28**
Boot-t	(2,3654 ; 4,6346)	**2,26**	(1,4128 ; 4,9206)	**3,50**
Percentil	(3,2326 ; 7,7366)	**1,70**	(1,3778 ; 4,1000)	**2,72**
BCPB	(2,5000 ; 4,6000)	**2,10**	(2,0000 ; 5,2111)	**3,21**
BC_a	(2,7000 ; 4,7000)	**2,00**	(2.5000 ; 5,0000)	**2,50**

5.2 Análise dos métodos *bootstrap*

Dentre os intervalos de confiança *bootstrap*, existem intervalos de confiança que apresentam melhores resultados, ou seja, é possível estabelecer o intervalo de confiança

bootstrap adequado para cada tipo de problema.

Exemplo 5.4. Foi realizada, num determinado período do ano, uma pesquisa para se estimar a média da taxa de hemoglobina no sangue (em gramas cm^3) dos operários de uma companhia de construção civil, MAGALHÃES [16]. Foram feitas medidas em 30 operários dessa companhia. Os 30 dados coletados apresentados na Tabela 5.7 formaram a amostra original. Com base nesta amostra original, foram realizadas 1000 amostras *bootstrap* do mesmo tamanho e aplicada a técnica *bootstrap*, a fim de calcular os intervalos de confiança *bootstrap* para a média da taxa de hemoglobina no sangue. Estes resultados foram comparados com o intervalo de confiança usual. Foi também calculado o intervalo de confiança *bootstrap* para a variância. Os dados foram analisados utilizando o software Matlab e os códigos 1 e 2, que se encontram respectivamente nos anexos A.1 e A.2.

Tabela 5.7: Dados da taxa de hemoglobina de 30 operários

11,1	12,2	11,7	12,5	13,9	12,3	14,4	13,6	12,7	12,7
12,6	11,3	11,7	13,4	15,2	13,2	13,0	16,9	15,8	14,7
13,5	12,7	12,3	13,5	15,4	16,3	15,2	12,3	13,7	14,1

Na Figura 5.1 é apresentado o histograma das 1000 amostras *bootstrap* das médias no qual foi verificado que a forma da distribuição é próxima da normal.

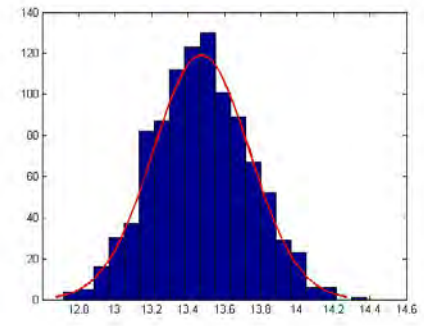


Figura 5.1: Histograma das 1000 médias *bootstrap*, exemplo 5.4

O gráfico apresentado na Figura 5.1 sugere a normalidade da distribuição do estimador da média nas amostras *bootstrap*. Isto pode ser visto através da forma muito próxima de uma distribuição normal no histograma. Neste caso, o intervalo de confiança *bootstrap*-t pode ser utilizado e deve apresentar limites próximos aos intervalos de confiança *bootstrap* percentil, ou seja os intervalos de confiança *bootstrap*-t e *bootstrap* percentil são adequados. A amostra original apresentou uma média estimada

para a taxa de hemoglobina no sangue igual a 13,4600 e variância igual a 2,2135. As amostras *bootstrap* apresentaram uma média igual a 13,4501 e variância igual a 2,1300. Os intervalos de confiança *bootstrap* para a média da taxa de hemoglobina no sangue foram calculadas através de 3 formas distintas do método *bootstrap* descritas anteriormente, a saber: intervalo de confiança *bootstrap* padrão [12,9424 , 13,9776], intervalo de confiança *bootstrap*-t [12,9073 , 14,01273] e o intervalo de confiança *bootstrap* percentil [13,0167 , 13,8767]. Os intervalos de confiança *bootstrap* para a média calculados através dos três métodos citados acima, revelaram-se muito próximos. Foi também calculado o intervalo de confiança assintótico para a média. Para este cálculo foram utilizados os dados da amostra original, com 95% de confiança, tendo sido obtido o intervalo [12,6363 , 14,2864]. A amplitude dos intervalos de confiança *bootstrap* calculados para a média foram menores que o intervalo de confiança assintótico para a média. A estimativa *bootstrap* do viés para a média é igual a $13,4501 - 13,4600 = -0,0099 < 0,25\widehat{dp}\,(média) = 0,0184$ considerado pequeno. Sendo a estimativa *bootstrap* do viés pequena para a média, reforça ainda mais que os intervalos de confiança *bootstrap*-t e o intervalo de confiança *bootstrap* percentil são adequados.

De modo análogo, foram obtidos os intervalos de confiança para a variância da taxa de hemoglobina no sangue dos operários. Na Figura 5.2 é apresentado o histograma das variâncias obtidas nas 1000 amostras *bootstrap* no qual não foi verificado a forma aproximadamente normal da distribuição.

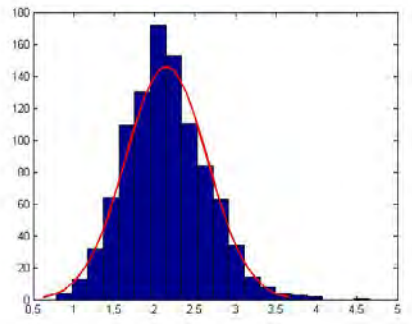


Figura 5.2: Histograma das 1000 variâncias *bootstrap*, exemplo 5.4

Desta forma os intervalos de confiança *bootstrap-t* e *bootstrap* percentil não são muito confiáveis uma vez que a suposição de normalidade não se verificou. A estimativa *bootstrap* do viés para a variância é igual a $2,1300 - 2,2135 = -0,0853 < 0,25\widehat{dp}\,(variância) = 0,1240$, considerado pequeno. Como a distribuição do estimador da variância não é aproximadamente normal os métodos BCPB e BC_a são adequados. Os intervalos de confiança *bootstrap* para a variância, calculados através desses dois métodos revelaram-se muito próximos a saber: intervalo de confiança *bootstrap* BCPB [1,4027 , 3,4070] e o intervalo de confiança *bootstrap* BC_a [1,4350 , 3,4656]. O valor da

constante de aceleração para a variância é 0,0590. Como a estimativa *bootstrap* do viés é negativa, a estimativa *bootstrap* está subestimando o valor da estatística. Pode-se observar que o intervalo de confiança *bootstrap* BCPB corrige o intervalo de confiança para a direita (Propriedade 4.2). O mesmo acontece com o intervalo de confiança *bootstrap* BC_a só que além de corrigir o intervalo para a direita este intervalo de confiança amplia seu tamanho por causa da constante de aceleração a. O intervalo assintótico para a variância com 95% de confiança é [1,3664 , 3,8229]. A amplitude desse intervalo é 2,5959. A amplitude de todos os intervalos de confiança *bootstrap* para a variância são menores que a amplitude do intervalo de confiança assintótico da variância. O intervalo de confiança *bootstrap* para a variância reduziu bem a amplitude em relação ao intervalo assintótico.

Assim, se o estimador do parâmetro de interesse tiver distribuição aproximadamente normal e apresentar viés pequeno, os intervalos de confiança *bootstrap*-t e o *bootstrap* percentil são adequados e os valores dos limites de confiança são próximos. Caso os intervalos de confiança *bootstrap*-t e o *bootstrap* percentil não apresentarem valores próximos os métodos não são adequados, ou seja, existem métodos melhores. Se a distribuição da estatística de interesse não for normal e/ou apresentar viés muito grande os métodos BCPB e BC_a são adequados. Esses métodos também são adequados quando a distribuição apresentar assimetria de maneira muito forte. Quando houver mais de um tipo de intervalo de confiança *bootstrap* adequado, esses intervalos de confiança são bem próximos.

Exemplo 5.5. Foi gerada uma amostra de variável aleatória $X \sim Exp(1/5)$, portanto sabemos que $E(X) = 5$ e $V(X) = 25$, com o objetivo de testar os métodos *bootstrap* de estimação por intervalos. Os dados estão apresentados na Tabela 5.8 e no histograma da Figura 5.3.

Tabela 5.8: Dados gerados com modelo Exponencial(1/5)

1.5675	20.8504	1.9071	7.1948	8.8270	0.0224	0.0766	4.8808
10.4564	10.1107	1.1764	5.9008	2.5569	3.1698	0.9831	

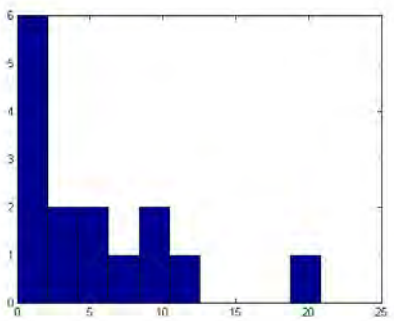
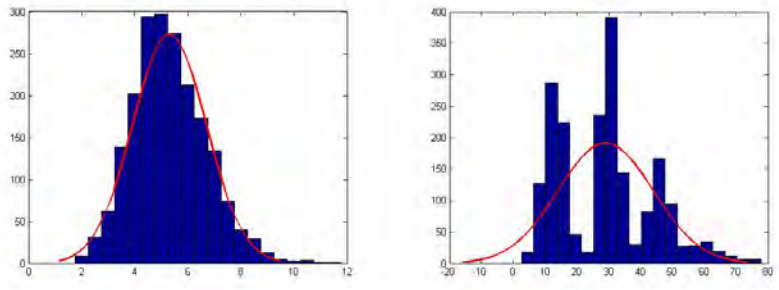


Figura 5.3: Histograma da amostra gerada, Exponencial(1/5)

Com base nesta amostra original, foram geradas 2000 amostras *bootstrap* do mesmo tamanho e aplicadas as técnicas de *bootstrap* afim de calcular os intervalos de confiança para a média desta variável. Todos os cálculos dos intervalos foram obtidos utilizando o software Matlab e os códigos obtidos em [2].
Os resultados foram comparados com o intervalo de confiança tradicional paramétrico, considerando a normalidade assintótica do estimador de máxima verossimilhança da média e da variância. Neste caso, se $X \sim Exp(\beta)$ então o estimador de máxima verossimilhança $EMV(\beta) = \widehat{\beta}$ segue distribuição assintótica Normal com média $E(\widehat{\beta}) = \beta$ e variânica $V(\widehat{\beta}) = n/\beta^2$.



(a) Histograma da média (b) Histograma da variância

Figura 5.4: Histogramas distribuições *bootstrap* da média e da variância, exemplo 5.5

Se estamos interessados nos intervalos de confiança para os estimadores da média e da variância de X, temos que $E(X) = g(\beta) = 1/\beta$ e $V(X) = h(\beta) = 1/\beta^2$ e pelo princípio da invariância, $EMV[g(\beta)] = g(\widehat{\beta}) = 1/\widehat{\beta}$ e $EMV[h(\beta)] = h(\widehat{\beta}) = 1/\widehat{\beta}^2$.

Para estimar a variância do estimador da média $g(\widehat{\beta})$ temos que $\widehat{V}[g(\widehat{\beta})] = [g'(\beta)]^2/V(\widehat{\beta})$ e para estimar a variância do estimador da variância temos $\widehat{V}[h(\widehat{\beta})] = [h'(\beta)]^2/V(\widehat{\beta})$, CASELLA e BERGER, [4]. Com isso as distribuições assintóticas para os estimadores

da média e da variância de X são, respectivamente:

$$1/\widehat{\beta} \sim N\left(\frac{1}{\beta}, \frac{1}{n\beta^2}\right) \qquad 1/\widehat{\beta}^2 \sim N\left(\frac{1}{\beta^2}, \frac{4}{n\beta^4}\right). \tag{5.1}$$

Na Tabela 5.9 apresentamos os intervalos de confiança para a média e para variância calculados utilizando normalidade assintótica, *bootstrap* padrão-z, *bootstrap-t* , *bootstrap* percentil , BCPB e BC_a. Uma forma de se comparar a qualidade dos intervalos é verificando-se a amplitude (A) destes, sendo a probabilidade de cobertura igual para todos. Neste trabalho, consideramos a probabilidade de cobertura igual a 95%.

Tabela 5.9: Comparação dos IC assintótico e *bootstrap*, exemplo 5.5

Método	$IC(g(\widehat{\beta}), 95\%)$	A	$IC(h(\widehat{\beta}), 95\%)$	A
Assintótico	(3,3922 ; 9,4910)	**6,10**	(0,0000 ; 56,7782)	**56,78**
Boot-z	(2,5875 ; 8,0366)	5,45	(2,0020 ; 60,3398)	58,64
Boot-t	(2,4026 ; 8,2215)	5,82	(0,0227 ; 62,3191)	62,30
Percentil	$(3{,}2326 \ ; \ 7{,}7366)^{I}$	**4,50**	$(5{,}9404 \ ; \ 53{,}2078)^{II}$	**47,27**
BCPB	(3,0304 ; 8,6513)	5,62	(10,4234 ; 69,6711)	59,55
BC_a	(3,5195 ; 10,2543)	6,73	(11,4405 ; 76,2715)	64,83

Sabemos que para $X \sim Exp(1/5)$ o verdadeiro valor da média é 5 e da variância é 25. As estimativas pontuais para média e variância, calculadas a partir da amostra original são $\overline{X} = 5,312$ e $S^2 = 31,17$ com $V(\overline{X}) = 2,08$ e $dp(\overline{X}) = 1,44$. A estimativa pontual para a média calculada por *bootstrap,* como esperado, não é muito diferente $\widehat{\theta}^*(\cdot) = 5,308$. O interessante é notar que a estimativa da variabilidade deste estimador é menor, isto é, o desvio padrão bootstrap é $\widehat{dp}_{boot}\left(\widehat{\theta}^*\right) = 1,39$. No caso da estimativa pontual para a variância calculada por *bootstrap,* temos o valor V_{boot} =28,98 com $\widehat{dp}_{boot}(V_{boot}) = 14,88$.

A estimativa *bootstrap* do viés para a média é $\widehat{viés}_{boot}\left(\widehat{\theta}\right) = -0,004$ considerado pequeno ($< 0,25\widehat{dp}[\left(\widehat{\theta}\right)] = 0,3475$). Além disso temos a simetria na distribuição *bootstrap* da média , como vemos no histograma da Figura 5.4. Esses resultados indicam que o método *bootstrap* percentil Ié adequado para essa aplicação e apresenta intervalo com a menor amplitude. No caso da variância, temos $\widehat{viés}_{boot} = -2,19$ que não é tão pequeno e não há simetria na distribuição do estimador, ver Figura 5.5. Isso indica que o método percentil do tipo II é mais adequado para o cálculo de intervalos de confiança para a variância. Vemos na Tabela 5.8 que ao comparar a amplitude do intervalo assintótico para a média com as amplitudes dos intervalos calculados por *bootstrap,* observamos que apenas o intervalo calculado com o método BC_a apresenta amplitude

maior que o assintótico, todos os outros apresentam amplitudes menores e essa é uma propriedade interessante do ponto de vista prático. A amplitude dos intervalos para variância não são menores, a não ser o intervalo percentil tipo II.

6 Aplicação a dados reais de natureza biológica

Neste capítulo apresentamos a aplicação dos métodos de cálculo de intervalos de confiança utilizando *bootstrap* paramétrico e não paramétrico e a comparação dos resultados destes métodos aplicados ao problema de contagem e estimação da média de indivíduos da espécie *Brachycephalus pitanga,* em atividade de vocalização, no ano de 2011, em área da Mata Atlântica (São Paulo, Brasil).

Trabalhos sobre ecologia das espécies (problemas biológicos) são de extrema importância para o entendimento dos fatores que afetam sua distribuição, atividade e relação com fatores bióticos e abióticos, e servem de base para outros estudos e delineamento de estratégias de conservação e preservação.

O delineamento de estratégias de conservação e preservação dos animais tem sido cada vez mais discutido na atualidade. Com o aumento da preocupação com o meio ambiente e com os impactos da ação do homem na natureza, os estudos elaborados têm apontado que as consequências das extinções prematuras de espécies, causadas pelo homem, incidem diretamente sobre seus habitats e também sobre a qualidade de vida das populações.

O bioma é constituído de fauna e flora, formando um sistema que precisou de milhões de anos para se desenvolver, e que, portanto, é interdependente entre suas espécies. A ação do homem, por meio da poluição, do desmatamento de habitats nativos, tem causado desequilíbrio nos biomas. A interferência da destruição e extinção prematura de espécies incide diretamente na vegetação e, por consequência, em todo o bioma, afetando também os rios e cursos d'água e a qualidade do ar, além das populações em gerais. Oficialmente o bioma mais sofrido do país é a mata atlântica. Neste contexto e com a preocupação social e cultural, cientistas encontraram nos últimos anos algumas espécies raras em áreas degradadas da mata atlântica, das quais se destacam os anfíbios anuros e se enquadram na questão do delineamento de estratégias de conservação e preservação. Um dos levantamentos mais recentes, divulgado em março de 2012 pela Sociedade Brasileira de Herpetologia (que reúne especialistas em anfíbios e répteis), mostra um grande aumento na quantidade de espécies de anfíbios, sendo que o Brasil é o campeão mundial de diversidade do grupo. A espécie *Brachycephalus pi-*

tanga é endêmica da mata atlântica. Isso significa que eles só existem ali, e em nenhum outro lugar do mundo. Nesse cenário encontramos o gênero de anuros miniaturizados *Brachycephalus*, que são a espécie alvo deste trabalho. Seu nome científico é *Brachycephalus pitanga*, SBH, [24]. Assim os problemas biológicos enfrentados são, perda de habitat por desmatamento e extinção das espécies, informações precisas para traçar o delineamento de estratégias de conservação e preservação.

6.1 Aplicação com dados reais

Uma equipe integrada pelo biólogo Célio Haddad, do Departamento de Zoologia da Unesp (Universidade Estadual Paulista) de Rio Claro, um dos maiores especialistas em anfíbios do país, descreveu recentemente o *Brachycephalus* pitanga. É um Anfíbio do grupo dos Anuros, minúsculo, mede cerca de 1 cm, bem pequeno conforme a Figura 6.1 em relação a mão do pesquisador, venenoso e é uma das novas espécies de sapos da Mata Atlântica. Tem cores como amarelo, marrom, laranja e vermelho. Suas cores lembram uma pitanga, daí a origem de seu nome, ESCOBAR [9]. Alimenta-se de pequenos animais, como insetos. Usa a fuga, toxinas venenosas ou se esconde entre plantas para se defender de predadores.

Figura 6.1: *Brachycephalus* pitanga. Fotografia: Célio FB Haddad

Segundo Haddad, "Várias espécies de *Brachycephalus* vêm sofrendo perdas de habitat por desmatamento e deverão enfrentar sérios problemas com as mudanças climáticas".

A espécie *Brachycephalus pitanga* desempenha um papel ecológico importante, controlando a população de insetos e representa também uma promessa biomédica, pois possue princípios bioativos na pele, que não foram bem estudados e que poderão ter aplicação no desenvolvimento de fármacos. Assim, na perda dessa biodiversidade, poderemos estar perdendo também uma série de medicamentos para doenças hoje incuráveis.

Na região no Núcleo Santa Virgína do parque da Mata Atlântica os pesquisadores Célio Haddad e Eliziane Garcia de Oliveira confirmaram que a espécie *Brachycephalus pitanga* é abundante. Os dados para este estudo foram coletados contemplando todas

as estações do ano, compreendidos entre os meses de fevereiro de 2011 até janeiro de 2012, sendo que em alguns meses não houve coletas de dados. Ao todo foram 9 meses de coletas. Essas coletas foram realizadas em um único dia de cada mês. Em cada dia de coleta foram considerados os fatores: período, que se subdivide em manhã (M), meio do dia (MD) e tarde (T), temperatura e umidade. A Figura 6.2 mostra um exemplar da espécie.

Figura 6.2: *Brachycephalus* pitanga. Fotografia: Carlos Gussoni

6.2 Intervalos de confiança *bootstrap* não paramétrico

A amostra original é dada pelas observações apresentadas no gráfico da Figura 6.3, que consistem na variável aletaória X: número de indivíduos em atividade de vocalização, em 27 ocasiões diferentes. Denotaremos

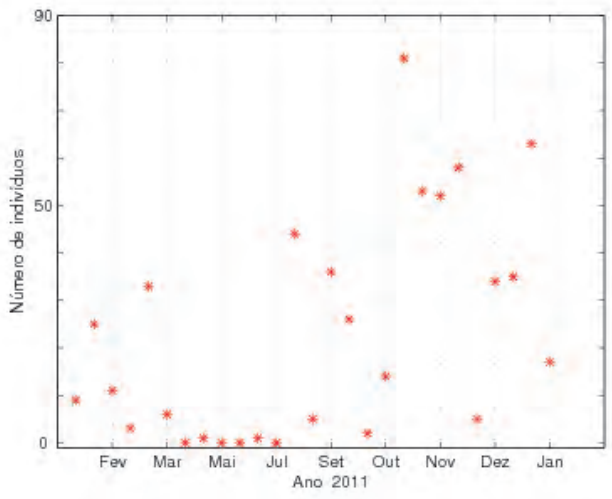


Figura 6.3: Número de indivíduos em atividade de vocalização

Aplicando o código 1 para os dados do Gráfico da Figura 6.3 obtemos os resultados descritos na Tabela 6.1.

Tabela 6.1: Aplicação do código 1, dados das espécies *Brachycephalus* pitanga do número de indivíduos em atividade de vocalização

O valor de $\widehat{\theta}(\cdot)$ para a amostra original.				
$\widehat{\theta}(\cdot)$	Média	Desvio Padrão	Variância	Mediana
	22,7407	23,3989	547,5071	14,0000
A média das B=40 réplicas bootstrap.				
$\widehat{\theta}^*(\cdot)$	Média	Desvio Padrão	Variância	Mediana
	22,7994	22,9406	534,5910	15,7680
O erro padrão bootstrap.				
$\widehat{dp}_{boot}\left(\widehat{\theta}\right)$	Média	Desvio Padrão	Variância	Mediana
	4,2752	2,8862	132,4821	8,4041
A estimativa bootstrap do viés.				
$\widehat{viés}_{boot}\left(\widehat{\theta}\right)$	Média	Desvio Padrão	Variância	Mediana
	0,0586	-0,4583	-12,9161	1,7680

Para implementar o *bootstrap* não paramétrico a amostra com 27 observações foi reamostrada e obtivemos B=1000 amostras *bootstrap* com as quais calculamos os intervalos de confiança utilizando o código 2, apresentados na Tabela 6.2.

Tabela 6.2: IC *bootstrap* não paramétrico, dados do número de indivíduos em atividade de vocalização

IC boot		Média	Var	DP	Mediana
padrão		$(14,36\ ;\ 31,12)$	$(287,84\ ;\ 807,17)$	$(17,74\ ;\ 29,06)$	$(-2,47\ ;\ 30,47)$
boot-t		$(13,95\ ;31,53)$	$(275,12\ ;\ 819,84)$	$(17,63\ ;\ 29,17)$	$(-2,80\ ;\ 30,80)$
perc I		$(15,81\ ;\ 29,85)$	$(327,76\ ;\ 767,06)$	$(18,10\ ;\ 27,69)$	$(5,00\ ;\ 33,00)$
perc II		$(15,63\ ;\ 29,67)$	$(327,76\ ;\ 767,05)$	$(19,09\ ;\ 28,69)$	$(-5,00\ ;\ 23,00)$
BCPB	p_0	0,4910	0,4450	0,4450	0,4130
	z_0	0,0226	0,1383	0,1383	0,2198
	P	$(0,0278\ ;\ 0,9775)$	$(0,0461\ ;\ 0,9873)$	$(0,0461\ ;\ 0,9873)$	$(0,0642\ ;\ 0,9918)$
	IC	$(14,77\ ;\ 31,63)$	$(323,53\ ;\ 861,95)$	$(17,98\ ;\ 29,35)$	$(5,00\ ;\ 35,00)$
BC_a	$\widehat{a}$	0,0266	0,0953	0,0989	0,0000
	P	$(0,0540\ ;\ 0,9943)$	$(0,0679\ ;\ 0,9943)$	$(0,0679\ ;\ 0,9943)$	$(0,0793\ ;\ 0,9955)$
	IC	$(15,93\ ;\ 33,97)$	$(342,31\ ;\ 901,84)$	$(18,50\ ;\ 30,03)$	$(6.00\ ;\ 35,00)$

Na Figura 6.4 é apresentado o histograma das 1000 amostras *bootstrap* das médias, na qual foi verificado que a forma da distribuição é próxima da normal. O viés *bootstrap* para a média é igual a $22,7994 - 22,7407 = 0,0587$ considerado pequeno ($<0,25\widehat{dp}\,(média) = 0,2166$). Desta forma os intervalos de confiança *bootstrap* pelo método *bootstrap-t* e pelo método *bootstrap* percentil são confiáveis uma vez que a

suposição de normalidade se verificou e o viés é considerado pequeno. Neste caso os métodos*bootstrap-t* e *bootstrap* percentil são adequados. Os intervalos de confiança *bootstrap* para a média, calculados através desses dois métodos revelaram-se muito próximos a saber: intervalo de confiança *bootstrap-t* $[13,9510\ ,\ 31,5305]$ e o intervalo de confiança *bootstrap* percentil $[15,8148\ ,\ 29,8619]$. O intervalo assintótico para a média com 95% de confiança é $[13,4820\ ,\ 31,9988]$. A amplitude desse intervalo é $18,5168$. A amplitude de todos os intervalos de confiança *bootstrap* para a média são menores que a amplitude do intervalo de confiança assintótico da média, como consta a Tabela 6.3.

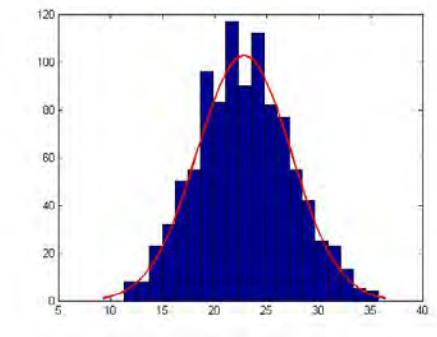


Figura 6.4: Histograma das 1000 médias *bootstrap*, dados de vocalização

De modo análogo foram obtidos os intervalos de confiança para a variância. Na Figura 6.5 é apresentado o histograma das variâncias obtidas nas 1000 amostras *bootstrap* no qual não foi verificado a forma aproximadamente normal da distribuição.

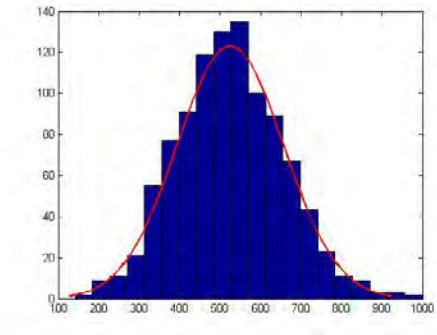


Figura 6.5: Histograma das 1000 variâncias *bootstrap*, dados de vocalização

Desta forma os intervalos de confiança *bootstrap* pelo método t e percentil não são muito confiáveis uma vez que a suposição de normalidade não se verificou. O viés *bootstrap* para a variância foi igual a $534,5910 - 547.5071 = -12,9161$ considerado grande $> 0,25\widehat{dp}\,(variância) = 1,2266$. Como a distribuição do estimador da variância não é normal e o viés é grande os métodos BCPB e BC_a são adequados. Os intervalos de confiança *bootstrap* para a variância , calculados através desses dois métodos revelaram-se

muito próximos a saber: intervalo de confiança *bootstrap* BCPB $[323,5328\ ,\ 861,9516]$ e o intervalo de confiança *bootstrap* BC_a $[342,3077\ ,\ 901,8490]$. O valor da constante de aceleração para a variância é $0,0953$. Como o viés foi negativo, a estimativa *bootstrap* está subestimando o valor da estatística. Pode-se observar que o intervalo de confiança bootstrap BCPB corrige o intervalo de confiança para a direita. O mesmo acontece com o intervalo de confiança bootstrap BC_a só que além de corrigir o intervalo para a direita este intervalo de confiança amplia seu tamanho por causa da constante de aceleração a. O intervalo assintótico para a variância com 95% de confiança é $[329,5945\ ,\ 998,9603]$. A amplitude desse intervalo é $669,3658$. A amplitude de todos os intervalos de confiança *bootstrap* para a variância são menores que a amplitude do intervalo de confiança assintótico da variância, como vemos na Tabela 6.3. O intervalo de confiança *bootstrap* para a variância reduziu bem a amplitude em relação ao intervalo assintótico.

Tabela 6.3: Comparação dos IC assintótico e os IC *bootstrap* para a média e variância, dados do número de indivíduos em atividade de vocalização

Método	*IC.* Média	*A*	*IC.* Var	*A*
Assintótico	(13,4820 ; 31,9988)	**18,52**	(329,5945 ; 998,9603)	**669,37**
Boot-z	(14,3614 ; 31,1200)	**16,75**	(287,8422 ; 807,1721)	**519,32**
Boot-t	(13,9510 ; 31,5305)	**17,57**	(275,1239 ; 819,8904)	**544,76**
Percentil	(15,8148 ; 29,8619)	**14,03**	(327,7635 ; 767,0541)	**439,29**
BCPB	(14,7778 ; 31,6296)	**16,85**	(323,5328 ; 861,9516)	**538,42**
BC_a	(15,9259 ; 33,9630)	**18,03**	(342,3077 ; 901,8490)	**559,54**

6.3 Intervalos de confiança *bootstrap* paramétrico

No caso do método *bootstrap* paramétrico, existe uma suposição sobre a distribuição que originou os dados e as B amostras *bootstrap* são geradas utilizando esse modelo, a partir dos parâmetros estimados com os dados da amostra original. Consideramos que, em geral, quando trabalha-se com dados de contagem, o primeiro modelo a ser testado é o Poisson. Mas para os dados desse trabalho esse modelo não é adequado devido à sobredispersão (variância maior que a média) apresentada: Média amostral $\overline{X} = 22,7407$ e variância amostral $S^2 = 547,5071$. Esse fenômeno ocorre provavelmente por causa da heterogeneidade das unidades amostrais. O modelo indicado para esse caso é o binomial negativo. Foi feito um teste de aderência e confirmou-se o modelo com $p - value \simeq 0.2275$.

Para verificar a afirmação de que os dados do número de sapinhos vocalizando, segue a distribuição binomial negativa realizamos o teste de aderência. Considere a

variável aleatória X: número de indivíduos em atividade de vocalização. As hipóteses a serem testadas são:

$H_0 : X$ segue distribuição binomial negativa;
$H_1 : X$ não segue distribuição binomial negativa.

Efetuando a estatística do teste usando a expressão,

$$Q^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

onde K representa o número de categorias, O_i a frequência observada, Tabela 6.4 e E_i a frequência esperada, Tabela 6.5 para a categoria i, tem-se $Q^2 = 8,1468$. Logo, utilizando $\alpha = 0,05$ e com o auxílio da tabela da Qui-Quadrado com 6 graus de liberdade, obtém-se a região crítica,

$$RC = \left\{\chi^2 \geq 12,592\right\}.$$

Tabela 6.4: Frequências Observadas

Ind. vocalizando	freq.obs.
$0 \vdash 10$	12
$10 \vdash 20$	3
$20 \vdash 30$	2
$30 \vdash 40$	4
$40 \vdash 50$	1
$50 \vdash 60$	3
$60 \vdash 70$	1
$70 \vdash 80$	0
$80 \vdash 90$	1

Tabela 6.5: Frequências Esperadas

Categoria	freq.esp.
1	10,2505
2	5,8399
3	3,7941
4	2,4712
5	1,6115
6	1,0517
7	0,6867
8	0,4837
9	0,8107

Concluimos pela aceitação do modelo binomial negativo para a variável aleatória X.

Assumindo que o número de indivíduos vocalizando segue o modelo binomial negativo, CASELLA e BERGER, [4] com parâmetros $w > 0$ e $0 < p < 1$, foi utilizado o método dos momentos para estimar os parâmetros com os dados reais. Temos que a função de probabilidade de X é dada por:

$$P(X = x) = \frac{\Gamma(x + w)}{\Gamma(w + 1)\Gamma(x)} p^w (1 - p)^x, x = 0, 1, ...$$

O método dos momentos consiste na solução de um sistema de equações que é construído igualando os momentos populacionais aos momentos amostrais, ou seja:

$$\begin{cases} E(X) = & \frac{w(1-p)}{p} = & \overline{X} = & 22,7407 \\ V(X) = & \frac{w(1-p)}{p^2} = & S^2 = & 547,5071 \end{cases} \Rightarrow \begin{cases} \widehat{p} = & 0,0415 \\ \widehat{w} = & 0,9853 \end{cases}$$

Com o modelo binomial negativo e esses valores estimados, geramos então $B = 1000$ amostras de tamanho 27 com as quais calculamos novamente os intervalos de confiança apresentados na Tabela 6.6.

Tabela 6.6: IC *bootstrap* paramétrico, dados do número de indivíduos em atividade de vocalização

IC boot		Média	Var	DP	Mediana
padrão		(13,86 , 27,76)	(63,00 , 661,46)	(11,01 , 27,04)	(9,80 , 32,20)
boot-t		(17,16 , 29,57)	(42,69 , 681,76)	(10,47 , 27,60)	(09,05 , 32,95)
perc I		(14,88 , 26,48)	(139,77 , 626,03)	(11,82 , 25,02)	(10,00 , 24,00)
perc II		(15,14 , 26,74)	(98,44 , 584,69)	(13,14 , 26,24)	(18,00 , 32,00)
BCPB	p_0	0,4860	0,4320	0,4320	0,4120
	z_0	0,0351	0,1662	0,1662	0,2224
	P	(0,0294 , 0,9788)	(0,0518 , 0,9891)	(0,0518 , 0,9891)	(0,0649 , 0,9919)
	IC	(14,44 , 28,11)	(141,23 , 759,56)	(11,88 , 27,04)	(10,00 , 27,00)
BC_a	$\widehat{a}$	0,0503	0,1429	0,1460	-0,0022
	P	(0,0679 , 0,9956)	(0,0869 , 0,9977)	(0,0869 , 0,9977)	(0,0961 , 0,9975)
	IC	(15,56 , 30,70)	(160,56 , 822,42)	(12,67 , 28,67)	(10,00 , 28,00)

Foi realizada a comparação dos resultados destes métodos tanto na sua forma paramétrica quanto na sua forma não paramétrica, aplicados ao problema de contagem e estimação da média de indivíduos da espécie *Brachycephalus pitanga,* em atividade de vocalização, no ano de 2011, em área da Mata Atlântica (São Paulo, Brasil), conforme os dados da Tabela 6.7.

Tabela 6.7: Comparação entre o IC assintótico e os IC *bootstrap* para a média do número de indivíduos em atividade de vocalização

Método Assintótico	(13,4820 ; 31,9988)		$A = 18,5168$	
	bootstrap não paramétrico	A	*bootstrap* paramétrico	A
Boot-z	(14,36 ; 31,12)	16,76	(13,86 ; 27,74)	13,90
Boot-t	(13,95 ; 31,53)	17,58	(17,16 ; 29,57)	12,41
Percentil	(15,81 ; 29,85)	14,04	(14,88 ; 26,48)	11,60
BCPB	(14,77 ; 31,63)	16,86	(14,44 ; 28,11)	13,67
BC_a	(15,93 ; 33,97)	18,04	(15,56 ; 30,70)	15,14

Uma forma de verificar a qualidade dos intervalos de confiança é comparar as amplitudes (A) destes, sendo que todos foram calculados com a mesma probabilidade de cobertura $100(1-\alpha)\% = 95\%$. Quanto menor essa amplitude, melhor é o intervalo do ponto de vista prático. Vemos que, em ambos os métodos (não paramétrico ou paramétrico) as amplitudes dos intervalos são todas menores que a amplitude do intervalo assintótico. Comparando entre o método não paramétrico e paramétrico observamos que o método paramétrico produz intervalos com amplitudes menores do que

o não paramétrico. Isso indica que quando é possível assumir um modelo para os dados é preferível do que fazer apenas a reamostragem da amostra original. Do ponto de vista da ecologia vemos que o resultado pode ser muito útil para avaliação da média do número de indivíduos em atividade de vocalização, fenômeno importante no estudo do acasalamento da espécie *Brachycephalus pitanga*.

7 Considerações finais

Métodos de *bootstrap* são métodos computacionais de análise estatística que usam simulação para calcular erros-padrão, viés e intervalos de confiança. São muito úteis por não necessitarem de muitas suposições para estimar parâmetros das distribuições de interesse , fornecem respostas mais precisas e de fácil entendimento e implementação. Uma grande vantagem do *bootstrap* é que essa técnica não depende do teorema central do limite, já em que em suas aplicações, medidas de precisão são obtidas diretamente dos dados. Os métodos são aplicados tanto na ánalise paramétrica quanto na não paramétrica.

Em geral as distribuições *bootstrap* tem aproximadamente a mesma forma e amplitude que as distribuições amostrais dos estimadores, no entanto elas estão centradas nas estimativas calculadas com os dados originais (amostras originais). Já as distribuições amostrais está centradas nos parâmetros da população.

O método de simulação *bootstrap* mostrou uma grande eficiência ao estimar os intervalos de confiança. A amplitude dos intervalos *bootstrap* foram menores em relação aos intervalos assintóticos, mesmo quando o tamanho da amostra era razoavelmente grande. Na aplicação do método *bootstrap* foram utilizadas 1000 replicações, pode-se utilizar um número maior de reamostragens, mas durante os testes, para valores de B perto e acima das 1000 replicações, os valores dos intervalos começaram a ficar iguais, apenas aumentando o tempo computacional demandado. Dentre os intervalos de confiança *bootstrap* existem intervalos de confiança que apresentam melhores resultados, ou seja é possível estabelecer o intervalo de confiança *bootstrap* adequado para cada tipo de situação dependendo do tipo de distribuição, viés e assimetria da estatística do parâmetro estudado.

Se o estimador do parâmetro de interesse tiver distribuição aproximadamente normal e tiver viés pequeno, os intervalos de confiança *bootstrap*-t e o *bootstrap* percentil são adequados e os valores dos limites de confiança são próximos. Caso os intervalos de confiança de *bootstrap*-t e o *bootstrap* percentil não apresentem valores próximos, os métodos não são adequados ou seja existem métodos melhores. Se a distribuição do estimador de interesse não for aproximadamente normal e/ou apresentar viés muito grande, os métodos BCPB e BC_a são adequados. Esses métodos também são adequados quando a distribuição apresentar assimetria de maneira muito forte. Quando

houver mais de um tipo de intervalo de confiança *bootstrap* adequado, esses intervalos de confiança são bem próximos. Enfim o método *bootstrap* permite que o cálculo do intervalo de confiança seja realizado de modo mais simples e abrangente para diversos estimadores, mesmo quando as distribuições dos mesmos não são conhecidas. Foi possível observar a generalidade de aplicação desta técnica visto que a mesma se adequa a qualquer situação, sendo seus cálculos rápidos e seus resultados muito eficientes como demonstrado na aplicação do método *bootstrap* ao problema de contagem e estimação da média de indivíduos da espécie *Brachycephalus pitanga,* em atividade de vocalização.

7.1 Trabalhos Futuros

Efetuar análises e aplicações de um outro método de intervalo de confiança *bootstrap* chamado ABC. Aprofundar os estudos sobre as propriedades dos intervalos de confiança e aplicar o métodos *bootstrap* para outras áreas de estudo, tais como medicina e economia.

8 Referências bibliográficas

[1] BABU, G. J. and SINGH, K. Inference on means using the *bootstrap*. Ann. Statist. 11 999-1003, 1984.

[2] BORKOWSKI, J. Notas de curso, disponível em www.math.montana.edu/vjobo/st431/index.html

[3] BUSSAB, W. O. e MORETTIN, P.A. Estatística básica - 5. Ed. - : Saraiva, São Paulo, 2002.

[4] CASELLA, G. e BERGER, R. L. Inferência Estatística. Tradução da segunda edição norte - americana, Cengage Learning - 2010.

[5] DAVISON, A.C. and HINKLEY, D.V. *Bootstrap* methods and their application, Cambridge University Press, 1997.

[6] DIACONIS, P. and EFRON, B. Computer-intensive methods in statistics. Sci. Amer. 113-130, 1983.

[7] EFRON, B. and TIBSHIRANI, R. An Introduction to the *bootstrap*. Chapman and Hall, New York,1983.

[8] EFRON, B. *Bootstrap* methods: Another look at the *jackknife*, Ann. Statist 7, 1-26, 1979.

[9] ESCOBAR, H. Em busca dos sapos-miniaturas no topo da mata atlântica. Matéria retirada do Jornal Estado de SP, 25 de março de 2012.

[10] HALL, L. Gray e W. R. Schucany, The Generalized *jackknife* Statistic, Marcel Dekker Inc., New York, 1972.

[11] HALL, P. On the number of *bootstrap* simulations required to construct a confidence interval. AS, 14, 1453-62, 1986.

[12] HESTERBERG, T. *Bootstrap* methods an permutation tests. In: The practice of business statistics: using data for decisions. New York, cap 18, 2003.

[13] KENDALL, M.G. and STUARD, A. The Advanced Theory of Statistics, 4th Edition. Griffin, London, 1977.

[14] MARTINEZ, E. Z. e LOUZADA, N. F. Estimação intervalar via *bootstrap*, Revista Mat-Estat.,São Paulo 19: 217-251, 2001.

[15] MANTEIGA,W. G., SÁNCHES, J. M. P, ROMO, J . "The *bootstrap* - A Review". In: Computational Statistics. Vol. 9. 165-205, 1994.

[16] MAGALHÃES, M. N. LIMA, A. C. P. Noções de Probabilidade e estatística - 7 ed.m 1.reimpr. - São Paulo: Editora da Universidade de São Paulo, 2011.

[17] MEMÓRIA, J. M. P. Breve história da estatística - Brasília, DF : Embrapa Informação Tecnológica, 111p, 2004.

[18] MONTGOMERY, D. C. ; RUNGER, G. C, Estatística aplicada e probabilidade para engenheiros, 2.ed . Rio de Janeiro: LTC, 2003.

[19] NAVIDI, W. C. Statistics for engineer and scientists. Boston: McGraw-Hill,c. 2006.

[20] QUENOUILLE, M. Approximate tests of correlation in time series. J. Royal. Statist. Soc. B 11, 18-44, 1949.

[21] RAMOS, E.M.L. Estatística poderosa ciência ao alcance de todos. Disponível em <http://www.ufpa.br/beiradorio/arquivo/Beira21/opiniao.html>, 2007.

[22] RAO, J.N.K. and WU, C.F.J. Resampling inference with complex survey data. J. Amer. Statist. Assoc. 231-241, 1988.

[23] RIZZO, A. L. T.; CYMROT, R. Estudo e aplicações da técnica *bootstrap*. II Jornada de Iniciação Científica, Universidade Presbiteriana Mackenzie.

[24] SBH. 2010. Brazilian amphibians – List of species. Sociedade Brasileira de Herpetologia (SBH). Disponível em: http://www.sbherpetologia.org.br/checklist/anfibios.htm, acessado em 25 de abril de 2013.

[25] SCHUCANY, W.R., GRAY, H.L. and OWEN, A. B. The Generalized *Jackknife* Statiscs, Marcel Dekker, New York, 1972.

[26] SHAO, J. and WU, C.F.J. A general theory for jackknife variance estimation. Ann. Statistic. c17, 1176-1197, 1989.

[27] TUKEY, J Bias and confidence in not quite large samples, Ann. Math. Statist. 29 - 614, 1958.

A Códigos *bootstrap* para a implementação no software Matlab

A.1 Código 1

disp(‘Código 1’); disp(‘ ’);

n = ; % n = tamanho da amostra;
B = 40; % B = número de amostras *bootstrap*;

rand(‘seed’,8005241); % semente

$x = [x_1, x_2, ..., x_n]$;
disp(‘Os dados’);
disp(x); disp(‘ ’);

% Histograma da amostra original
histfit(x)

xtmp = zeros(n,1); % vetor inicial;
bmn = zeros(B,1);
bstd=bmn; bmed = bmn; bvar= bmn;

xbar = mean(x); % cálculo das estimativas ;
xstddev = std(x);
xvr = xstddev^2;
xmedian = median(x);

disp(‘theta chapéu valor para a média, desvio padrão, variância e mediana’);
disp(‘ média , desvio padrão , variância e mediana’);
xsum = [xbar xstddev xvr xmedian];
disp(xsum); disp(‘ ’);

disp([‘O número de amostras *bootstrap* B= ’,int2str(B)]);

```
disp(‘ ’);

disp(‘As amostras bootstrap’);
for b = 1:B;
xvec = ceil(n*rand(n,1));
for i=1:n; % Gerar amostras bootstrap;
xtmp(i) = x(xvec(i));
end;
disp(xtmp’);
bmn(b) = mean(xtmp); % b-ésima média bootstrap
bstd(b) = std(xtmp); % desvio padrão,;
bvar(b) = std(xtmp)^2; % variância ;
bmed(b) = median(xtmp); % mediana ;
end;
disp(‘ ’);

disp(‘Replicações bootstrap: theta chapéu estrela’);
disp(‘ média, desvio padrão ,variância e mediana’)
bout = [bmn bstd bvar bmed];
disp(bout); disp(‘ ’);

disp(‘Média das B replicações bootstrap : theta chapéu estrela ’)
disp(‘ média , desvio padrão , variância e mediana’)
bmean = mean(bout);
disp(bmean); disp(‘ ’);

disp(‘Erro Padrão bootstrap: dp chapéu ’)
disp(‘ média , desvio padrão , variância e mediana’)
bstderr = std(bout);
disp(bstderr); disp(‘ ’);

disp(‘Estimativa bootstrap do viés: viés boot ’)
disp(‘ média , desvio padrão , variância e mediana’)
bbias = bmean - xsum ;
disp(bbias); disp(‘ ’);

bmn = sort(bmn);
bstd = sort(bstd);
bvar = sort(bvar);
bmed = sort(bmed);

ECDF = [];
for b = 1:B; ECDF(b,1) = b/B; end;
```

```
disp('réplicas bootstrap ordenadas')
disp(' Percentil, média desvio padrão variância mediana')
bout = [ECDF bmn bstd bvar bmed];
disp(bout)
```

A.2 Código 2

Código 2

```
disp('Código 2'); disp(' ');

n = ; % n = tamanho da amostra
B = 1000; % B = número de réplicas bootstrap B

cl = 95; % Nível de confiança em porcentagem;
t = 2.0930; % valor utilizado no bootstrap-t;
z = 1.960; % valor utilizado no bootstrap padrão ;

rand('seed',52789033); % semente

x = [x_1, x_2, ..., x_n];
disp('Os dados');
disp(x'); disp(' ');

histfit(x);

xtmp = zeros(n,1); %vetor inicial;
bmn = zeros(B,1);

bstd = bmn; bmed = bmn; bvar = bmn; bstdn= bmn;

xbar = mean(x); % cálculo das estimativas ;
xstddev = std(x);
xstdn = xstddev*sqrt((n-1)/n);
xvr = xstddev^2;
xmedian = median(x);

disp('theta(chapéu) valor para média, desvio padrão, variância, s(n), mediana');
disp(' média desvio padrão variância s(n) mediana');
xsum = [xbar xstddev xvr xstdn xmedian];
disp(xsum); disp(' ');

disp(['O número de amostras bootstrap B = ',int2str(B)]); disp(' ');
```

```
for b = 1:B;
xvec = ceil(n*rand(n,1));
for i=1:n; % Gerar amostras bootstrap;
xtmp(i) = x(xvec(i));
end;
bmn(b) = mean(xtmp); % b-ésima média bootstrap ;
bstd(b) = std(xtmp); %desvio padrão;
bstdn(b)= std(xtmp)*sqrt((n-1)/n); % desvio padrão denominador
n;
bvar(b) = std(xtmp)^2; % variância ;
bmed(b) = median(xtmp); % mediana ;
end;

disp('Replicações bootstrap: theta chapéu estrela');
disp(' média desvio padrão variância s(n) mediana');
bout = [bmn bstd bvar bstdn bmed];
disp(bout);
disp(' ');

% Exibir histogramas das amostras bootstrap

bmn = sort(bmn); histfit(bmn,20);
% print -dwin
pause;

bstd = sort(bstd); histfit(bstd,20);
% print -dwin
pause;

bvar = sort(bvar); histfit(bvar,20);
% print -dwin
pause;

bstdn= sort(bstdn); histfit(bstdn,20);
% print -dwin
pause;

bmed = sort(bmed); histfit(bmed,20);
% print -dwin
pause;

disp('Média das B replicações bootstrap: theta chapéu estrela)')
disp(' média desvio padrão variância s(n) mediana');
```

```
bmean = mean(bout);
disp(bmean); disp(‘ ’);

disp(‘Erro padrão bootstrap: dp chapéu ’)
disp(‘ média desvio padrão variância s(n) mediana’);
bstderr = std(bout);
disp(bstderr); disp(‘ ’);

disp(‘Esimativa bootstrap do viés: viés boot ’)
disp(‘ média desvio padrão variância s(n) mediana’);
bbias = bmean - xsum ;
disp(bbias); disp(‘ ’);

bmn = sort(bmn);
bstd = sort(bstd);
bvar = sort(bvar);
bstdn=sort(bstdn);
bmed = sort(bmed);

ECDF = [];
for b = 1:B; ECDF(b,1) = b/B; end;

disp(‘replicações bootstrap ordenadas’)
disp(‘ ECDF média desvio padrão variância s(n) mediana’)
bout = [ECDF bmn bstd bvar bmed];
disp(bout)

disp(‘————————————————————————————————’);
disp(‘INTERVALO DE CONFIANÇA bootstrap PADRÃO ’);
disp(‘————————————————————————————————’);

zl = xsum-z*bstderr; zu = xsum+z*bstderr;

disp([int2str(cl),’% intervalo de confiança bootstrap padrão z’]);
disp(‘ média desvio padrão variância s(n) mediana’);
disp([zl ; zu]); disp(‘ ’);

disp(‘Intervalo de confiança padrão z com viés ajustado’);
disp(‘ média desvio padrão variância s(n) mediana’);
disp([zl-bbias ; zu-bbias]); disp(‘ ’);

disp(‘————————————————————————————————’);
disp(‘ INTERVALO DE CONFIANÇA bootstrap-t ’);
disp(‘————————————————————————————————’);
```

```
tl = xsum-t*bstderr; tu = xsum+t*bstderr;

disp([int2str(cl),'% intervalo de confiaça bootstrap-t']);
disp(' média desvio padrão variância s(n) mediana');
disp([tl ; tu]); disp(' ');

disp('Intervalo de confiança bootstrap-t com viés ajustado');
disp(' média desvio padrão variância s(n) mediana');
disp([tl-bbias ; tu-bbias]); disp(' ');

disp('————————————————————————————————————————');
disp('INTERVALO DE CONFIANÇA bootstrap PERCENTIL');
disp('————————————————————————————————————————');

lpct = B*(100-cl)/100; upct = B*(cl/100)+1;

bsrt = [bmn bstd bvar bstdn bmed ];
pl = bsrt(lpct,:); pu = bsrt(upct,:);

disp([int2str(cl),'% intervalo de confiança bootstrap percentil– Tipo I']);
disp(' mean s variance s(n) median')
disp([pl ; pu]); disp(' ');

disp([int2str(cl),'% intervalo de confiança bootstrap percentil Tipo II']);
disp(' mean s variance s(n) median')
pl2 = 2*xsum - pu; pu2 = 2*xsum - pl;
disp([pl2 ; pu2]); disp(' ');

disp'————————————————————————————————————————');
disp('INTERVALO DE CONFIANÇA bootstrap CORRIGIDO EM RELAÇÃO AO
VIÉS (BCPB)');
disp('————————————————————————————————————————');

pvec = zeros(1,5);
for b = 1:B;
for est = 1:5;
if bsrt(b,est) xsum(1,est); pvec(1,est) = pvec(1,est) + 1; end;
end; end;

disp( 1: A proporção p das réplicas bootstrap theta(chapeu));
disp(' média desvio padrão variância s(n) mediana');
pvec = pvec/B;
disp(pvec); disp(' ');
```

```
disp(‘ z0 valores associados a proporção p’);
disp(‘ média desvio padrão variância s(n) mediana’);
z0 = -norminv(pvec,0,1);
disp(z0); disp(‘ ’);

disp(‘Passo 2: obter os percentis’);
disp(’ média desvio padrão variância s(n) mediana’);
phiL = normcdf(2*z0-z*ones(1,5));
phiU = normcdf(2*z0+z*ones(1,5));
disp( [ phiL; phiU ]); disp(’ ’);
BCL = ceil(B*phiL); BCU = floor(B*phiU);
BCCIL = zeros(1,5); BCCIU = zeros(1,5);
for est = 1:5;
BCCIL(1,est) = bsrt(BCL(1,est),est);
BCCIU(1,est) = bsrt(BCU(1,est),est);
end;

disp([‘Passo 3: ’, int2str(cl),’% intervalo de confiança bootstrap corrigido em relação
ao viés’]);
disp(‘ média desvio padrão variância s(n) mediana’);
disp([BCCIL ; BCCIU]); disp(‘ ’);

disp(‘————————————————————————————’);
disp(‘BCa’);
disp(‘————————————————————————————’);

for i=1:n; % remove row i from x and store;
     jtmp = x; % remaining n-1 values in xtmp ;
     jtmp([i],:) = [];
     jrmn(i,1) = mean(jtmp); % i-ésima réplica jackknife ;
     jrstd(i,1) = std(jtmp);
     jrvar(i,1) = var(jtmp);
     jrstdn(i,1) = sqrt(n-2)*std(jtmp)/sqrt(n-1);
     jrmd(i,1) = median(jtmp);
end;
jout = [jrmn jrstd jrvar jrstdn jrmd];
jmean = mean(jout);
for i=1:n; % Estimativa da constante de aceleração ;
     aout(i,:) = jmean - jout(i,:);
     theta2(i,:) = aout(i,:).*aout(i,:);
     theta3(i,:) = theta2(i,:).*aout(i,:);
end;
```

```
disp('Estimativa da constante de aceleração a');

disp(' média desvio padrão variância s(n) mediana');
ac = sum(theta3)./(6* ((sum(theta2)).^1.5 ));
disp(ac); disp(' ');

% ('Cálculos dos percentis');
disp(' média desvio padrão variância s(n) mediana')
phiL2 = normcdf(z0 + (z0-z*ones(1,5))/(1-ac.*(z0-z*ones(1,5)))) ;
phiU2 = normcdf(z0 + (z0+z*ones(1,5))/(1-ac.*(z0+z*ones(1,5)))) ;
disp( [ phiL2; phiU2 ]); disp(' ');
BCL2 = ceil(B*phiL2); BCU2 = floor(B*phiU2);
BCCIL2 = zeros(1,5); BCCIU2 = zeros(1,5);
for est = 1:5;
BCCIL2(1,est) = bsrt(BCL2(1,est),est);
BCCIU2(1,est) = bsrt(BCU2(1,est),est);
end;

disp('O Intervalo BCa');
disp(' média desvio padrão variância s(n) mediana')
disp([BCCIL2 ; BCCIU2]); disp(' ');
```

B Exemplo do intervalo de confiança *bootstrap* Percentil

Exemplo B.1. Considere a amostra original x=$(8, 3, 7, 4, 6, 2, 7, 3, 9, 5)$. Utilizando O código 1, A.1 (BORKOWSKI), [2], com $B = 40$ réplicas *bootstrap*, obtém-se ao final deste código as réplicas *bootstrap* ordenadas para cada estatística de interesse e também os percentis em ordem crescente, conforme listado a seguir. Adotando o coeficiente de confiança de 95%, ou seja $\alpha = 0,05$, tem-se que o intervalo de confiança *bootstrap* percentil do tipo 1 é dado pelos valores dos percentis $2,5\%$ e $97,5\%$, sendo o valor do percentil $2,5\%$ como sendo o limite inferior do intervalo de confiança e o valor do percentil $97,5\%$ como sendo o limite superior desse intervalo. Por exemplo, o intervalo de confiança bootstrap percentil para a média é dado por:

$$\left[\widehat{\theta}^*_{\left(\frac{0,05}{2}\right)} \ , \ \widehat{\theta}^*_{\left(1-\frac{0,05}{2}\right)}\right] = \left[\widehat{\theta}^*_{(0,025)} \ , \ \widehat{\theta}^*_{(0,975)}\right] = [4,1000 \ , \ 6,6000] \, .$$

PERCENTIS	MÉDIA	DP	VAR	MEDIANA
0.0250	4.1000	1.4491	2.1000	3.0000
0.0500	4.2000	1.5055	2.2667	3.5000
0.0750	4.2000	1.6193	2.6222	3.5000
0.1000	4.7000	1.6865	2.8444	3.5000
0.1250	4.8000	1.7512	3.0667	4.0000
0.1500	4.8000	1.8135	3.2889	4.0000
0.1750	4.8000	1.8738	3.5111	4.0000
0.2000	4.9000	1.9322	3.7333	4.0000
0.2250	4.9000	1.9465	3.7889	4.5000
0.2500	5.0000	2.0028	4.0111	4.5000
0.2750	5.1000	2.0111	4.0444	4.5000
0.3000	5.1000	2.0790	4.3222	4.5000
0.3250	5.2000	2.0790	4.3222	4.5000
0.3500	5.2000	2.0790	4.3222	5.0000
0.3750	5.2000	2.1082	4.4444	5.0000
0.4000	5.2000	2.1108	4.4556	5.0000

0.4250	5.2000	2.1187	4.4889	5.0000
0.4500	5.2000	2.1318	4.5444	5.0000
0.4750	5.3000	2.1499	4.6222	5.0000
0.5000	5.3000	2.1833	4.7667	5.0000
0.5250	5.3000	2.2211	4.9333	5.5000
0.5500	5.4000	2.2608	5.1111	5.5000
0.5750	5.5000	2.2706	5.1556	5.5000
0.6000	5.6000	2.2706	5.1556	5.5000
0.6250	5.6000	2.3214	5.3889	6.0000
0.6500	5.6000	2.3476	5.5111	6.0000
0.6750	5.6000	2.3944	5.7333	6.0000
0.7000	5.6000	2.4404	5.9556	6.0000
0.7250	5.6000	2.4495	6.0000	6.0000
0.7500	5.9000	2.4967	6.2333	6.5000
0.7750	5.9000	2.5033	6.2667	6.5000
0.8000	5.9000	2.5033	6.2667	6.5000
0.8250	6.0000	2.5144	6.3222	6.5000
0.8500	6.0000	2.5582	6.5444	6.5000
0.8750	6.2000	2.5734	6.6222	6.5000
0.9000	6.3000	2.5841	6.6778	7.0000
0.9250	6.4000	2.7406	7.5111	7.0000
0.9500	6.5000	2.7988	7.8333	7.0000
0.9750	6.6000	3.0258	9.1556	7.0000
1.0000	7.2000	3.1198	9.7333	7.5000