O MÉTODO BOOTSTRAP E SUA APLICAÇÃO EM ANÁLISE DE DADOS AGROFLORESTAIS COM VARIÁVEIS ALEATÓRIAS DO TIPO RAZÃO

ANTONIO CLAUDIO ALMEIDA DE CARVALHO
Engº. Agrônomo

Orientador: Prof. Dr. HILTON THADEU ZARATE DO COUTO

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", da Universidade de São Paulo, para obtenção do título de Mestre em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
Estado de São Paulo - Brasil
Novembro - 1996

O MÉTODO BOOTSTRAP E SUA APLICAÇÃO EM ANÁLISE DE DADOS AGROFLORESTAIS COM VARIÁVEIS ALEATÓRIAS DO TIPO RAZÃO

ANTONIO CLAUDIO ALMEIDA DE CARVALHO

Aprovado em: 24/02/1997

Comissão julgadora:

Prof. Dr. Hilton Thadeu Zarate do Couto	ESALQ/USP
Prof^{a}. Dr^{a}. Clarice Garcia Borges Demétrio	ESALQ/USP
Pesq. Dr. Antonio Carlos de Oliveira	CNPMS/EMBRAPA

Prof. Dr. Hilton Thadeu Zarate do Couto
Orientador

A meus pais:

Neuza Maria Almeida de Carvalho (in memorian)

José Monteiro de Carvalho,

dedico com amor e carinho.

A meus irmãos:

Lindomar, Luzimar, Londimar, João Bosco,

Lamartine, Artur, Maria da Graças e Andréa

dedico com gratidão.

À Silvia e Laiane,

dedico com amor.

AGRADECIMENTOS

Expresso aqui meus agradecimentos a todos aqueles que de alguma forma colaboraram para a realização deste trabalho, em especial:

Ao Professor Hiton Thadeu Zarate do Couto, a orientação e o estímulo.

Ao Professor Décio Barbin, a consideração e o apoio.

Aos Professores Doutores, Décio Barbin, Clarice Garcia Borges Demétrio, Maria Cristina S. Nogueira, José Eduardo Corrente, Humberto de Campos e Antonio Francisco Iemma, os ensinamentos e o convívio.

Aos Colegas de curso, Silvano, Beth, Renata, Rui, Maria Cristina, Gino, Dalberto e José Fernando, a amizade.

À Luciane, Solange e Rosa, a presteza e simpatia.

À EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA em especial ao CENTRO DE PESQUISA AGROFLORESTAL DO AMAPÁ - CPAF/AMAPÁ, as condições oferecidas para realização deste mestrado.

Ao CENTRO DE PESQUISA AGROFLORESTAL DA AMAZÔNIA ORIENTAL - EMBRAPA/CPATU, a cessão dos dados experimentais utilizados neste trabalho.

Aos Colegas pesquisadores do CPAF/AMAPÁ, o companheirimo, incentivo e o apoio.

À COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR - CAPES, o apoio financeiro que me foi conferido à realização deste curso.

SUMÁRIO

O MÉTODO BOOTSTRAP E SUA APLICAÇÃO EM ANÁLISE DE DADOS AGROFLORESTAIS COM VARIÁVEIS ALEATÓRIAS DO TIPO RAZÃO

Autor: Antonio Claudio Almeida de Carvalho

Orientador: Prof. Dr. Hilton Thadeu Zarate do Couto

RESUMO

O conceito de produção sustentada, baseada no princípio de diversificação de culturas é consolidado através dos sistemas agroflorestais, que é a denominação recente para os cultivos consorciados que envolvem um componente arbóreo, culturas agrícolas e/ou animais.

Devido à possibilidade de múltiplas interações entre os componentes, a análise e a interpretação dos dados experimentais de um sistema agroflorestal pode tornar-se complexa. Uma abordagem muito encontrada na literatura, para análise de cultivos consorciados é feita através do LER (*Land Equivalent Ratio*), que representa uma medida de equivalência do uso da terra do consórcio em relação ao monocultivo. Do ponto de vista estatístico, o LER representa uma variável aleatória formada pela razão de duas variáveis aleatórias e, conseqüentemente, sua distribuição de probabilidades nem sempre segue a distribuição normal. Esse fato, impossibilita a aplicação dos métodos paramétricos, comumente empregados na experimentação agronômica.

Os métodos computacionalmente intensivos como "Jackknife" e "Bootstrap" possibilitam análise estatísticas livres de suposições de modelos teóri-

cos, tornando possível a exploração das propriedades amostrais, independentemente de suas formas analíticas.

O método "Bootstrap" é mais versátil que o método "Jackknife" e pode ser implementado facilmente, tanto na forma não-paramétrica quanto paramétrica, para uma grande variedades de situações. A idéia básica dos procedimentos "Bootstrap" baseia-se no fato de se obter uma distribuição empírica, que reproduza o mecanismo probabilístico gerador dos dados amostrais e assim, a partir de grande quantidades de reamostras, obtêm-se as estimativas das estatísticas de interesse.

Encontra-se neste trabalho uma sucinta descrição do método "Jackknife". Os conceitos e algoritmos que envolvem os procedimentos "Bootstrap" não-paramétricos, são descritos e executados através de dados simulados. A análise de um sistema agroflorestal, com a variável aleatória LER, foi realizada com o uso dos procedimentos "Bootstrap" e através dos *software* SAS e S-PLUS obtiveram-se limites de confiança e teste de hipótese para os parâmetros populacionais.

BOOTSTRAP APPLICATION TO DATA ANALYSIS OF AGROFORESTRY EXPERIMENTS FOR RATIO TYPE RANDOM VARIABLES

Author: Antonio Claudio Almeida de Carvalho

Adviser: Prof. Dr. Hilton Thadeu Zarate do Couto

SUMMARY

The concept of sustainable yield, based on the principle of crop diversification, is consolidated through agroforestry systems, that is the up-to-date denomination for mixture of crops that include a tree component, crops and/or livestock.

Due to the possibility of multiple interactions among the systems components, the analysis and interpretation of the experimental data of an agroforestry systems many became complex. An approach found in the literature for analysis of intercropping experiments is to use LER (Land Equivalent Ratio), that represent a measurement of equivalence of land use by the intercropping and cropping. Under statistical point of view, LER represents a random variable representing the ratio of two random variables, and its distribution not always follow the normal distribution. This fact do not allow the application of parametric methods, frequently used in agricultural experiments.

Methods which use intensively computer, as Jackknife and Bootstrap, allow for analysis of data without following the assumptions of theoretical models. This, it is possible to explore sample properties, independent of analitical

forms.

Bootstrap is more freely used as the Jackknife method and can be implemented as parametric and non parametric forms, for wide situations. The basic idiea of Bootstrap procedures is that one can get empirical distribution, that mimic the mechanism of generating sample data, and then, with the large amount of resampling data, it is, possible to get statistics of interest.

In this paper it was presented a brief description of the Jackknife method and the concepts underlying Bootstrap procedures under non parametic form are also described and executed with simulated data. The analysis of a Agro-forestry system, using LER as random variable was analised through the use of Bootstrap procedure and SAS and S-PLUS softwares. It is presented the confidence interval for the parameters and the respective hypothesis test.

1 Introdução

Sistema Agroflorestal é uma denominação recente para as práticas agrícolas, algumas bastante antigas, em que estão associadas na mesma área, espécies florestais com culturas agrícolas, espécies florestais com animais ou os três componentes juntos, numa associação simultânea ou seqüencial. Nesse sistema é claro a interação entre os componentes envolvidos e, devido às ações complementares dos mesmos, na maioria das vezes obtêm-se efeitos positivos com relação aos cultivos monocultivos.

A complexidade das interrelações que ocorrem nos sistemas consorciados, mais especificamente os sistemas agroflorestais, têm dificultado a análise e interpretação dos dados experimentais dos mesmos. Pois, devido ao caráter multivariado dos dados, do ponto de vista experimental, admitem-se inúmeras alternativas de modelagem e conseqüentemente, várias técnicas de análises quantitativas.

Para os ensaios envolvendo consórcios com duas culturas anuais, encontram-se na literaturas vários trabalhos enfocando o assunto, inclusive um livro publicado recentemente, de Walter T. Federer, dedicado especificamente ao tema (*Statistical Design and Analysis for Intercropping Experiments - Two Crops*, 1993).

No entanto, para os consóricios onde estão associadas mais de duas culturas ou cultivos de ciclo longo, há escassez de trabalhos referindo-se à análise estatística dos mesmos. Segundo BARNETT & RILEY (1995), é necessário testar a aplicabilidade das metodologias desenvolvidas para consórcios com culturas anuais na análise de sistemas agroflorestais pois, esse sistema envolve maior complexidade, dada às diferentes mudanças nos modelos de desenvolvimento e ao fato de que as relações entre espécies podem sofrer alterações no decorrer do tempo.

Uma abordagem largamente encontrada na literatura, que tem grande apelo intuitivo, para análise estatística de cultivos consorciados é feita através da utilização da variável LER (*Land Equivalent Ratio*), que representa uma medida de eficiência do uso da terra do sistema consorciado em relação ao monocultivos. No entanto, do ponto de vista estatístico, o LER é uma variável aleatória, formada pela razão de duas variáveis aleatórias; tanto o numerador quanto o denominador estão sujeitos à variação experimental e, portanto, a distribuição de probabilidades do LER não segue necessariamente a distribuição normal. Além disso, segundo GONÇALVES (1982), obtém-se uma estrutura correlacionada para as observações das parcelas e isso torna duvidoso que a suposição de linearidade se verifique. Sendo assim, os testes estatísticos como a análise de variância, não podem ser aplicados.

MEAD & RILEY, (1981) propõem algumas alternativas para minimizar os problemas que o LER apresenta sobre as correlações e não normalidade. Essas alternativas consistem em alterar a forma de obtenção do LER, de maneira que, o denominador não seja mais uma variável aleatória e assim, o LER passa a ter distribuição aproximadamente normal. O problema é que os melhores resultados, segundo esses autores, ocorrem quando são usados nos denominadores valores externos ao experimento, como a produção média da região ou valores ótimos pré-fixados. Todavia, os ensaios agroflorestais envolvem culturas perenes e suas avaliações são feitas ao longo do tempo e a obtenção do LER, conforme sugerido acima, pode promover o aparecimento de outras correlações.

Uma solução para os problemas apresentados pelo a variável aleatória LER, pode ser encontrada quando se faz uso dos métodos computacionalmente intensivos como o método "Jackknife" desenvolvido por QUENOUILLE[1], citado por GREGOIRE (1984) e o método "Bootstrap", proposto por EFRON (1979).

---

[1]QUENOUILLE, M. H. 1956. Notes on bias in estimation. **Biometrika. 43**: 353-360.

Graças aos avanços na capacidade de processamento dos computadores, as análise estatísticas através dos referidos métodos tornaram-se uma ferramenta atraente e uma alternativa para os modelos estatísticos tradicionais, quando estes não são adequadamente ajustados. A principal vantagem dos métodos computacionalmente intensivos está no fato de possibilitarem análises estatísticas livres de suposições de modelos teóricos para a distribuição dos dados. Isso torna possível a obtenção de estatísticas de interesse, como estimativas de variâncias, tendenciosidades, intervalos de confiança e testes de hipóteses, sem o uso das expressões analíticas dessas estatísticas. Em outras palavras, os métodos estatísticos computacionalmente intensivos substituem o poder analítico das expressões teóricas pelo poder de processamento dos computadores.

Nesse trabalho, o objetivo concentra-se no método "Bootstrap", por sua versatilidade e facilidade de implementação através dos *softwares* disponíveis comercialmente. Todavia, será abordado de forma suscinta o procedimento "Jackknife".

Através de um exemplo simulado serão abordadas as idéias básicas da inferência estatística tradicional, conjuntamente com as alternativas apresentadas através da aplicação do método "Bootstrap". Dados de um experimento agroflorestal, envolvendo a variável aleatória LER serão analisados com o uso do procedimento "Bootstrap" e, no Apêndice encontram-se os programas computacionais desenvolvidos para implementação da referida análise.

2 Revisão da Literatura

2.1 Definição de Um Sistema Agroflorestal

Sistema Agroflorestal é uma denominação recente para práticas culturais, algumas bastante antigas, em que estão associadas no mesmo espaço, espécies florestais com culturas, espécies florestais com animais, ou os três componetes juntos, em uma associação simultânea ou sequencial, em faixas ou em misturas. Em todas essas associações existe uma evidente interação dos componentes com os fatores ambientais, principalmente no que tange às condições edafoclimáticas. Esses sistemas têm como objetivo a diversificação da produção, aumento dos níveis de materia orgânica no solo, fixação do nitrogênio atmosférico, ciclagem de nutrientes, etc.. Em síntese, busca-se a otimização da produtividade através do conceito de produção sustentada. Segundo NAIR (1985 E 1990) os sistemas agroflorestais podem ser classificados conforme a Figura 1.

Os Sistemas Agroflorestais não são uma panacéia; no entanto, têm um grande potencial de expansão como sistema de uso da terra na região amazônica, principalmente entre os pequenos e médios agricultores. Eles devem desempenhar um papel cada vez mais importante no desenvolvimento agrícola e florestal na Amazônia (SERRÃO *et al.*, 1994). Na América Latina, os sistemas silvo-pastoris encontram-se ainda nas primeiras etapas de desenvolvimento, sendo encontrado em pequenas áreas, onde estão associadas árvores frutíferas e espécies florestais com gramíneas e/ou leguminosas forrageiras (VEIGA & SERRÃO, 1990).

A pesquisa e experimentação dos sistemas agroflorestais adquiriu tanta importância na amazônia brasileira, que em 1991 a EMBRAPA transformou suas seis unidades de pesquisas da região em "Centros de Pesquisas Agroflorestais".

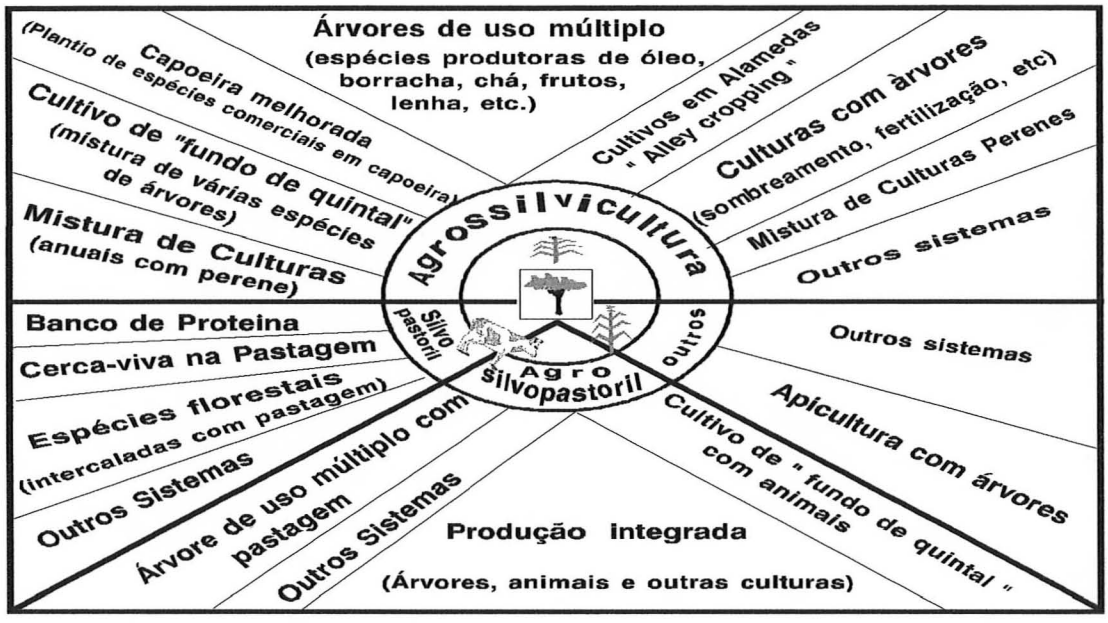

Figura 1: Classificação dos Sistemas Agroflorestais baseada no tipo de componente. FONTE: Adaptado de NAIR(1985)

Metodologias estatísticas têm sido exploradas e desenvolvidas para consórcios com culturas anuais, como citados por FEDERER (1993), MEAD & RILEY (1981), WILLEY & OSIRU (1972), KASS (1978), entre outros. Entretanto, para consórcios envolvendo cultuas perenes, há escassez de trabalhos quando se trata da análise quantitativa desses sistemas. Segundo BARNETT & RILEY (1995), é necessário testar a aplicabilidade das metodologias desenvolvidas para consórcios com culturas anuais em análise de sistemas agroflorestais, pois, esses sistemas envolvem maior complexidade, dada as diferentes mudanças nos modelos de desenvolvimento e ao fato de que as relações entre as espécies podem sofrer alterações no decorrer do tempo. Segundo esses autores, além das várias possibilidades de fontes de variações, deve-se considerar o fator tempo, pois, este representa um importante aspecto nos sistemas agroflorestais.

Nesses sistemas, além das associações biológicas, ainda é possível existirem interações econômicas, uma vez que pode haver diferenças entre os valores das produções dos referidos componentes. Segundo BARNETT & RILEY (1995),

a avaliação de uma mistura de duas diferentes espécies é complexa devido à possibilidade de diferentes tipos de variações de cada componente e da possibilidade de relações múltiplas entre ambas. Quando a mistura envolve culturas perenes, esta interrelação pode sofrer mudanças no decorrer do tempo e se há mais de duas espécies, tal complexidade pode tornar-se extrema.

Devido ao caráter multivariado dos dados, há inúmeras alternativas de modelagem e consequentemente várias técnicas para realizar as análises quantitativas dos sistemas agroflorestais. Segundo GONÇALVES (1982), não há consenso quanto à escolha da melhor forma de abordagem de uma dada situação, envolvendo os sistemas consorciados; o enfoque depende fundamentalmente do interesse do pesquisador e do nível de abordagem que se pretende dar aos resultados do experimento.

RAMALHO *et al.* (1983), relatam que a falta de informações sobre as teorias de análises, quase todas desenvolvidas para monocultivos, e a complexidade dos sistemas consorciados têm dificultado a análise dos dados e interpretação dos resultados. OLIVEIRA (1994), postula que a grande área demandada pelos sistemas agroflorestais impossibilita a utilização do número suficiente de repetições para detectar diferenças significativas a níveis de probabilidade rigorosos.

A seguir, serão descritas de forma sucinta, algumas alternativas encontradas na literatura para análise estatística de sistemas envolvendo associação de culturas. A abordagem aqui apresentada seguem as linhas básicas delineadas por FEDERER (1993) e MEAD & RILEY (1981).

2.2 Análise Com Base na Cultura Principal

Em alguns sistemas agroflorestais, o interesse centraliza-se sobre o desenvolvimento de uma cultura principal. Nesse tipo de consórcio é importante que o rendimento dessa cultura não seja prejudicado pela presença dos outros componentes ou que a redução seja pequena, dentro de limites pré-estabelecidos (FEDERER, 1993). Uma pressuposição básica deste tipo de consorcio é que a densidade da cultura principal seja a mesma usada no monocultivo e que os demais componentes mantenham-se constante em todo o sistema (MEAD & WILLEY,1981).

Esta situação é encontrada nos sistemas agroflorestais que envolvem uma cultura de importância econômica consorciada com árvores de uso múltiplos (adubação verde, sombreamento, quebra-vento, cerca-viva, etc.)

A análise desse sistema torna-se bastante simplificada, uma vez que a variável de interesse refere-se apenas à resposta da cultura principal. Dessa forma, a análise estatística é feita através de procedimentos univariados.

2.3 Análise Com Base na Produção Combinada

Os experimentos envolvendo sistemas consorciados apresentam o caráter multivariado dos dados, ou seja, cada parcela fornece variáveis-respostas relativas a dois ou mais componentes. Visando evitar a abordagem multidimensional, é comum a transformação dos dados em uma única unidade e a partir dessa variável comum, realizar a análise de forma univariada. Entretanto, a combinação desses rendimentos é difícil de ser obtida, uma vez que os mesmos diferem entre si, sobre vários aspectos, ou seja, em valores monetários, energético, ecológicos, etc.

O procedimento utilizado para fazer a análise estatística dos componentes simultaneamente, consiste em converter, por algum critério de equivalência, todos os rendimentos em um único valor, de maneira que se possa realizar a comparação dos tratamentos por uma abordagem univariada. Dentre os critérios de equivalência, os mais usados são:

2.3.1 Produção Equivalente (*Income Equivalent Ratio*)

É o método mais prático para realizar a análise estatística de ensaios em que existam dois ou mais componentes. O procedimento consite em analisar os componentes através de uma variável comum (PE), que pode ser a produção de matéria seca, produção de proteínas, etc., ou converter as produções observadas em um único padrão monetário (KASS, 1978). Esta variável comum é em geral uma função do tipo:

$$PE = \sum_{i=1}^{m} r_i \bar{y}_i$$

onde:

PE é a produção equivalente referente aos m componentes do sistema,

$\bar{y}_i$ é o rendimento médio do i-ésimo componente, $(i = 1, 2 \cdots m)$;

r_i é o valor relativo do i-ésimo componente, estabelecido em relação a um padrão definido arbitrariamente.

A análise estatística, neste caso, é feita da maneira usual, tendo como variável-resposta a variável aleatória PE. O emprego deste procedimento, para análise de ensaios envolvendo varias culturas pode ser encontrado em WILLEY & OSIRU (1972); WIJESINHA *et al* (1982); RAMALHO *et al.* (1983); CRUZ (1990); entre outros. Um exemplo da simplicidade deste método é apresentado por GOMES (1984), que é descrito como segue: "Seja um experimento de culturas consorciadas, envolvendo milho e feijão, cada parcela tem a produção de feijão (y_1) e a produção

de milho (y_2). Um quilograma de feijão vale tanto quanto 5 kg de milho, então a produção equivalente é $PE = \bar{y}_1 + 5\bar{y}_2$".

2.3.2 Razão da Área Equivalente - LER (*Land Equivalent Ratio*)

Originalmente proposto como uma medida de comparação entre a performance de uma espécie em consórcio e o desempenho da mesma em monocultivo (WILLEY E OSIRU 1972), a Razão da Área Equivalente tem sido o método mais empregado na análise de experimentos envolvendo culturas consorciadas. O LER recebeu essa denominação por ser um coeficiente que mede a área de terra requerida no monocultivo necessária para obtenção de uma mesma produção para o caso do cultivo consorciado.

NAIR (1990), mostra, através de um exemplo simples, como o LER é calculado. "Admita que em 1 *ha* de terra, cultivada de forma consorciada, seja possível obter a produção de 10 unidades de uma espécie arbórea e 50 unidades de uma lavoura. Quando cultivadas em monocultivos, necessita-se de $0,75$ *ha* para se produzirem as 10 unidades da espécie arbórea e $0,5$ *ha* para se produzirem as 50 unidades da lavoura. Assim, para se obterem as mesmas produções das espécies consorciadas é necessário $1,25$ *ha* (0,75 + 0,5) de terras com monocultivos. Então o LER é 1,25".

Quando o LER é maior que 1, como no exemplo acima, pode-se afirmar que os rendimentos dos cultivos consorciados supera os monocultivos. Se o LER for igual a 1, não há vantagens de um sobre o outro, mas se o LER for inferior a 1, indica que o cultivo consorciado é desfavorável.

Do ponto de vista estatístico, o LER trata-se de uma variável aleatória formada pela soma das razões entre a produções obtidas em consócio e as produções obtidas em monocultivos. Isto é,

$$LER = \sum_{i=1}^{m} \frac{\bar{y}_i}{\bar{z}_i} \tag{1}$$

onde:

$\bar{y}_i$ é a produção media observada do i-ésimo componente, quando cultivado de forma consorciada, em uma determinada unidade de área;

$\bar{z}_i$ é a produção média do mesmo i-ésimo componente na mesma unidade de área, quando cultivado em monocultivo.

A análise estatística é feita através de modelos univariados, sobre o valor obtido na Expressão (1) de cada parcela. O problema é que o LER trata-se de uma variável aleatória, formada pela razão de duas outras variáveis aleatórias Y e Z. Sendo assim, o LER apresenta uma distribuição complexa e desconhecida pois, segundo GONÇALVES (1982), as parcelas apresentam correlações entre a variáveis Y e Z, tornando duvidoso que as suposições de linearidade e normalidade se verifiquem. Consequentemente, a aplicação de testes de hipóteses na análise de variância deve ser feito com cautela.

OYEJOLA & MEAD (1982), estudaram o efeito de seis métodos de determinação do LER sobre a análise de variância e concluíram que quanto maior a padronização do denominador (produção observada em monocultivo) maior a normalidade do LER. Segundo estes autores, a melhor padronização ocorre quando é usado um denominador comum para todas os tratamentos e repetições de cada componente.

2.4 Análise Usando Modelos Multivariados

O uso de estatística multivariada na análise de experimentos com consorciação de cultura é uma prática bastante usada, principalmente nos ensaios envolvendo culturas anuais. DEAR & MEAD (1983), apresentam uma ampla discussão sobre essa análise. FEDERER (1993), cita vários exemplos em que a análise é feita através de técnicas multivariadas. Através de um experimento com culturas consorciadas de milho e feijão, realizado pela EMBRAPA, no `Centro de Pesquisa de Arroz e Feijão`, FEDERER (1993) demonstra a aplicação dos modelos multivariados na análise estatística de dados experimentais de culturas consorciadas. WIJESINHA *et al* (1982) utilizaram técnicas bivariadas na análise de um ensaio envolvendo consórcio de culturas e concluíram que os resultados dessa análise elevam as informações na identificação das combinações mais favoráveis.

A aplicação de técnicas multivariadas na análise de culturas consorciadas tem sido incentivada. MEAD & RILEY (1981), ressaltam a necessidade do uso mais intensivo das técnicas multivariadas e de uma abordagem mais completa envolvendo o LER e suas relações com os parâmetros da análise multivariada.

Quando se deseja obter uma interpretação para o sistema consorciado, como um todo, há sempre alguma dificuldade na aplicação das técnicas estatísticas e, segundo NAIR (1990), o uso de variáveis equivalente para análise estatística dos sistemas agroflorestais não tem sido aplicada com muita freqüencia, principalmente, porque a metodologia desenvolvida não retrata o atributo mais relevante que é sustentabilidade do sistema. Por outro lado, o uso de técnicas multivariadas requer delineamentos experimentais, com número suficientemente grande de repetições, o que pode torná-la inaplicável na maioria dos ensaios. Pois, sempre envolvem-se grande extensões de áreas quando se trabalha com sistemas agroflorestais.

2.5 Análise Usando os Métodos Computacionalmente Intensivos

Com os avanços da capacidade de processamento dos computadores, as análises estatísticas envolvendo métodos computacionalmente intensivos tornaram-se ferramentas atraentes e alternativas viáveis quando os modelos estatísticos não se aplicam ou possuem demasiada complicação analítica. Suas principais vantagens estão no fato de possibilitarem análises livres das suposições de modelos teóricos, para a distribuição dos dados e darem margem à exploração das propriedades amostrais de interesse, independentemente de suas formas analíticas. Os mais importantes são o método "Jackknife", introduzido por Quenouille e Tukey na década de 50 e o método "Bootstrap" sugerido por EFRON (1979).

GREGOIRE (1984), usou o método "Jackknife" em dados de populacões florestais para obtenção das estimativas do viés e da variância da razão entre a média do volume de madeira medido e a média do volume estimado visualmente; estimativas do viés e da variância da razão entre a média da área foliar e a média do peso das folhas; estimativas do viés e da variância da razão entre a média do peso seco de biomassa e a média do diâmetro do caule. O método "Jackknife" também foi empregado na obtenção da estimativa da variância da média quadrática do diâmetro do caule. Nesse trabalho o autor conclui que o método "Jackknife" é uma ferramenta segura para obtenção das estimativas das estatísticas descritas acima.

TRINCA (1988) aplicou o método "Bootstrap" em dados de espécies nativas da mata de Santa Genebra, no município de Campinas-SP, com o objetivo de estudar a biodiversidade; especialmente a curva do número de espécies com relação ao tamanho da amostra. Dentre outras conclusões, a autora informa que o padrão geográfico parece não afetar os resultados obtidos pelo método "Bootstrap" para a referida estatística.

3 Material e Métodos

3.1 Considerações Iniciais Sobre Inferência Estatística

A teoria estatística procura obter respostas para as três questões básicas:

i) Como obter as amostras de uma população;

ii) Como analisar os dados dessa amostra;

iii) Como acurar (verificar a exatidão) os valores obtidos dessa análise.

Esta última questão (ítem iii) constitui o processo conhecido como inferência estatística, pois a partir dos dados amostrados pretende-se inferir conclusões sobre a população da qual foi extraída a amostra. Para exemplificar, verifique-se o Exemplo 1, que se refere a um experimento simples envolvendo dois tratamentos.

Exemplo 1: Num experimento agroflorestal hipotético, foram implantados dois tratamentos. O tratamento A refere-se ao plantio de uma cultura perene consorciada com uma espécie de leguminosa fixadora de nitrogênio, enquanto o tratamento B refere-se ao plantio solteiro da cultura perene. O objetivo é verificar o efeito da leguminosa sobre a produção da cultura perene. As produções observadas são apresentadas na Tabela 1.

Tabela 1: Dados simulados de um experimento de agrosilvicultura com dois tratamentos

Tratamento	Amostras	Produções Observadas (kg/ha)				Média amostral	Variância amostral	Erro padrão da media
A*	y_i	485,61	519,63	413,93	535,01			
		502,49	480,84	412,26	-	478,54	2345,60	18,31
B**	z_j	464,58	339,25	457,00	433,47			
		439,86	432,04	430,89	529,89	440,87	2758,12	18,57
		diferença				37,67	–	–

$*$ Dados obtidos por simulação da distribuição normal ($\mu = 500$ e $\sigma^2 = 2500$)

$**$ Dados obtidos por simulação da distribuição normal ($\mu = 450$ e $\sigma^2 = 2500$)

A pergunta que se segue é a seguinte: A leguminosea aumentou a produção da cultura perene ? Com base nas médias amostrais dos tratamentos

$$\bar{y} = \frac{\sum_{i=1}^{7} y_i}{7} = 478,54 \qquad e \qquad \bar{z} = \frac{\sum_{j=1}^{8} z_j}{8} = 440,87 \ ,$$

pode-se, preliminarmente, pensar que sim, pois, a diferença das média $(37,68)$ sugere que a leguminosa promoveu o aumento na produção da cultura perene. Mas qual é a precisão das estimativas obtidas por $\bar{y}$ e $\bar{z}$? Afinal, elas foram estimadas com base em amostras muito pequenas, respectivamente, n=7 e m=8.

Para responder a estas questões é necessário verificar quão precisas são as médias estimadas $\bar{y}$ e $\bar{z}$. A medida mais comum da precisão de um estimador é dado pelo seu erro padrão, que é por definição a raiz quadrada da variância do mesmo (Efron & Tibshirani, 1993). Quando o estimador é a média amostral torna-se fácil obter uma medida de precisão pois, neste caso, existe uma fórmula simples que fornece o erro padrão da média.

A estimativa do erro padrão da média amostral $\bar{y}$ a partir de n amostras independentes, $y_1, y_2, \cdots y_n$ é dado por:

$$s_{\bar{y}} = \sqrt{\frac{s^2}{n}} \tag{2}$$

onde:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Para os dados do Exemplo 1, os erros padrões das médias dos tratamentos A e B são, respectivamente, $18,31$ e $18,57$. Se os erros padrões das médias amostrais fossem pequenos (por exemplo, menor que 1) saber-se-ia que os valores estimados por $\bar{y}$ e $\bar{z}$ estariam próximos de seus valores esperados e conseqüentemente, a diferença $37,67$ seria um bom estimador para o verdadeiro valor do efeito da leguminosa. Por outro lado, se os erros padrões fossem muito grandes (por exemplo, 50) a diferença $37,67$ seria um estimador bastante impreciso para avaliar o efeito da leguminosa sobre a produção da cultura perene.

A fundamentação teórica das considerações supra citadas é dada como segue: Admita que,

$$Y_1, Y_2, \cdots Y_n \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$$

e

$$Z_1, Z_2, \cdots Z_m \overset{iid}{\sim} N(\mu_Z, \sigma_Z^2).$$

Então, a diferença $(\bar{Y} - \bar{Z})$ será normalmente distribuida com média $(\mu_Y - \mu_Z)$ e variância $\left(\frac{\sigma_Y^2}{n} + \frac{\sigma_Z^2}{m}\right)$.

Logo,

$$\frac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_Z^2}{m}}} \overset{iid}{\sim} N(0,1).$$

Quando σ_Y^2 e σ_Z^2 são desconhecidas e os tamanhos das amostras são suficientemente grandes, podem-se substituir σ_Y^2 e σ_Z^2 por seus respectivos estimadores s_Y^2 e s_Z^2. Entretanto, se os tamanhos das amostras não são suficientemente grandes, usa-se a estatística dada na Expressão (4), que tem distribuição t-Student com $(n+m-2)$ graus de liberdade.

Admitindo-se a homogeneidade de variância $(\sigma_Y^2=\sigma_Z^2=\sigma^2)$, s^2 que fornece uma estimativa de variância comum para ambos os tratamentos, pode ser obtido através da Expressão (3).

$$s^2 = \frac{(n-1)s_Y^2 + (m-1)s_Z^2}{n+m-2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \sum_{j=1}^{m}(Z_j - \bar{Z})^2}{n+m-2}. \qquad (3)$$

Logo,

$$\frac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \overset{iid}{\sim} t \qquad (4)$$

E, com base na distribuição t-Student, pode-se testar a hipótese $H_0 : \mu_Y - \mu_Z = 0$ contra a hipótese $H_a : \mu_Y - \mu_Z \neq 0$ através da Expressão (5). Pois, sob H_0, a diferença $(\bar{Y} - \bar{Z})$ tem distribuição t-Student com $(n+m-2)$ graus de liberdade. Isto é:

$$t = \frac{(\bar{Y} - \bar{Z})}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \overset{iid}{\sim} t_{[n+m-2]} \, . \qquad (5)$$

Voltando aos dados do Exemplo 1, o valor para a estatística t é:

$$t = \frac{478,54 - 440,87}{50,67\sqrt{\frac{1}{7} + \frac{1}{8}}} = 1,44 \, .$$

Com base no teste t-Student com 13 gl e coeficiente de confiança de 95%, ($t_{[13\ ;\ 0,05]}$=2,160), não rejeita-se H_0, concluindo-se portanto, que não há efeito significativo da leguminosa sobre a produção da cultura perene. Isso contraria a idéia inicial, que se tinha, quando foi tomado o valor da diferença das médias amostrais (37,67).

Esse exemplo serviu para demonstrar que não se pode tomar decisões com base apenas no valor pontual de uma estatística. Deve-se sempre agregar à mesma uma medida de precisão. Agora, suponha por exemplo, que o interesse consista em comparar os dois tratamentos da Tabela 1, através de suas medianas e não mais através de suas médias. Para os tratamentos A e B têm-se, respectivamentes, as medianas 485,61 e 436,66, cuja diferença é 48,95. Por este critério poder-se-ia também imaginar que o tratamento com leguminosa é superior. Mas, como verificar a precisão dessa estatística agora ?

Foi visto que para o caso da média amostral, existe uma fórmula simples de obter o seu erro padrão, mas infelizmente, não existem fórmulas simples, como da Expressão (2), para outros estimadores que não seja a média amostral.

Em geral, as expressões analíticas que fornecem medidas de precisão para estimadores como mediana, coeficiente de correlação, estimadores de índices tipo razão, etc., são complexas e demandam-se muitos cálculos para obtenção de suas estimativas. Todavia, com o surgimento dos métodos computacionalmente intensivos é possível obter as estimativas dos referidos estimadores, sem o uso das complexas expressões analíticas.

3.2 O Método "Jackknife"

O método "Jackknife" é uma ferramenta que pode ser utilizada para estimar, de forma não-paramétrica, algumas medidas de variabilidade, como a variância e o viés de uma determinada estatística de interesse. A metodologia mais comum para implementar o método computacionalmente intensivo "Jackknife" é descrita como se segue.

A partir da amostra original de tamanho n, formam-se as pseudo-amostras de tamanho n-1 da seguinte forma: a primeira pseudo-amostra é formada com todas as observações originais com exceção do primeiro valor observado; a segunda pseudo-amostra é formada com todas as observações originais com exceção do segundo valor observado, e assim sucessivamente, até a n-ésima pseudo-amostra. No entanto, este é o caso particular, em que os p grupos de tamanho q são formados com os valores p=n e q=1. Todavia, o método "Jackknife" pode ser implementado com pseudo-amostras formadas a partir de quaisquer valores de p e q, tal que n=pq, conforme descrito abaixo.

Seja $\hat{\delta} = f(X_1, X_2 \cdots X_n)$ um estimador do parâmetro δ, baseado em n variáveis aleatórias, independentes e identicamente distribuídas (*iid*). Seja n divisível por p grupos de tamanho q, tal que n=pq e finalmente, seja $\hat{\delta}_{(-i)}$ um estimador de δ baseado em todos menos o i-ésimo grupo. O estimador "Jackknife" é definido como:

$$\hat{\delta}_j = p\hat{\delta} - \frac{p-1}{p}\sum_{i=1}^{p}\hat{\delta}_{(-i)}$$

$$\hat{\delta}_j = p\hat{\delta} - (p-1)\hat{\delta}_{(.)}$$

$$\hat{\delta}_j = \frac{\sum_{i=1}^{p}\hat{\delta}_i}{p}$$

onde $\hat{\delta}_{(.)}$ é a média das $\hat{\delta}_{(-i)}$ e $\hat{\delta}_i$=$p\delta - (p-1)\hat{\delta}_{(-i)}$. Aqui, o índice j refere-se a

notação "Jackknife" e o índice i refere-se ao i-ésimo grupo.

O nome "Jackknife" pretende dar a conotação de uma técnica de ampla utilidade. A denominação de "pseudo valores" para o termo $\hat{\delta}_i$, assim como o nome *jackknife* são creditados a John W. Tukey. Todavia, segundo MILLER (1974), muito mais importante que a rotulação destes dois termos, é de Tukey o resultado de que os pseudos valores $\hat{\delta}_i$ comportam-se, aproximadamente, como variáveis aleatórias independentes e identicamente distribuídas. Para o caso da estimação da variância através do método "Jackknife", temos que:

$$v(\hat{\delta}_i) = \frac{\sum_{i=1}^{p} (\hat{\delta}_i - \hat{\delta}_j)^2}{p-1},$$

e conseqüentemente, um estimador aproximado, não viesado, da variância de $\hat{\delta}_j$ é:

$$v(\hat{\delta}_j) = \frac{\sum_{i=1}^{p} (\hat{\delta}_i - \hat{\delta}_j)^2}{p(p-1)}$$

MOSTELLER & TUKEY[2], citados por GREGOIRE (1984), sugerem que a familia de distribuição t com $p-1$ graus de liberdade pode ser usada como uma referência contra a qual testa-se os valores da estatística

$$\frac{\hat{\delta}_j - \delta}{\sqrt{v(\hat{\delta}_j)}}, \tag{6}$$

que pode ser usada também para estimação por intervalo do parâmetro δ.

Devido aos pseudos-valores não serem verdadeiramente independentes, a expressão (6) tem somente distribuição t-Student aproximada. Apesar disso, ela tem sido considerada satisfatoria para a solução de uma grande variedade de problemas. HINKLEY (1977), mostra que a precisão desta aproximação aumenta com o aumento do tamanho de q (tamanho do grupo).

[2]MOSTELLER, F. and J. W. TUKEY. 1977. Data analysis and regression. Addison-Wesley Publishing Co.. Reading. MA.

O método "Jackknife" aplicado ao estudo do viés de estimadores e na inferência de parâmetros, quando as distribuições são obscuras ou desconhecidas, tem sido bastante discutido na literatura. No entando, neste trabalho não será abordado a aplicação deste método, pois o objetivo restringe-se ao método "Bootstrap". Em MILLER (1974), pode ser encontrada uma completa revisão da metodologia "Jackknife", com aplicações na redução de viés e inferências de parâmetros de distribuições complexas.

3.3 O Método "Bootstrap"

A idéia básica do método "Bootstrap" consiste em estimar estatísticas desejadas, quando os dados advêm de uma distribuição desconhecida ou complexa, reproduzindo-se o mecanismo probabilístico, gerador dos dados originais. A distribuição desconhecida é substituída por uma distribuição conhecida que aproxima-se à verdadeira distribuição dos dados originais. Assim, estatísticas que não poderiam ser avaliadas na estrutura original do problema são estimadas por estatísticas correspondentes, calculadas em uma pseudo estrutura de dados.

O método "Bootstrap" é mais versátil que o "Jackknife" e pode ser implementado facilmente, tanto na forma não-paramétrica como na forma paramétrica, para uma grande variedade de situações. No caso não-paramétrico, o método "Bootstrap" é implementado, extraindo-se amostras aleatórias com reposição de tamanho n, ao contrário do método "Jackknife", onde as amostras são extraídas sem reposição. Para o caso paramétrico, implementado quando existem informações suficientes sobre a forma da distribuição dos dados, a reamostragem é feita sob a referida distribuição, com os parâmetros desconhecidos sendo substituidos pelas respectivas estimativas paramétricas.

Segundo EFRON & TIBSHIRANI (1993), essa idéia é bastante antiga, mas, devido às dificuldades de obtenção da distribuição "Bootstrap" `Exata` de uma determinada estatística, só recentemente tem sido implementada, graças aos avanços da capacidade de processamento dos computadores. Usando o processo de simulação Monte Carlo pode-se sempre obter uma aproximação numérica para a distribuição "Bootstrap" `Exata`.

O termo "Bootstrap" surgiu da frase "*to pull oneself up one's bootstrap*", retirada de "Adventures of Boron Munchausen", de Rudolph Erich, que relata uma situação em que o Barão está afundando em um lago e vendo que tudo estava perdido, pensa que conseguirá emergir puxando os cadarços dos próprios sapatos (EFRON & TIBSHIRANI, 1993). Segundo SILVA (1995), o sentido básico do termo é passar a idéia de que em situações difícies deve-se tentar o impossível.

Na estatística, as "situações de dificuldades" podem ser vistas como os problemas de soluções analíticas complexas e o "impossível" seria a utilização de uma metodologia em que é necessário grande quantidade de cálculos, mesmo para analisar um pequeno conjunto de dados SILVA (1995). As soluções para esses casos, com o uso dos métodos computacionalmente intensivos, são obtidas substituindo-se o poder analítico das expressões teóricas pelo poder de processamento dos computadores.

De uma forma geral, o método "Bootstrap" pode ser usado para estimar estatísticas de interesse como: estimativas de variâncias, viés de um estimador, intervalos de confiança, etc. Testes de hipótese aproximados sobre parâmetros podem ser calculados conforme sugere HINKLEY (1988) e EFRON & TIBSHIRANI (1993). Será descrito a seguir a fundamentação teórica do procedimento "Bootstrap", bem como, algumas aplicações deste método, através da análise de um ensaio simulado e análise de um experimento com dados reais. Neste trabalho, todas as considerações e aplicações dizem respeito ao método "Bootstrap" não-paramétrico.

3.4 O Erro Padrão da Média

Seja X uma variável aleatória (iid) que possui uma distribuição de probabilidade F, com esperança μ_F e variância σ_F^2. Isto é,

$$X \overset{iid}{\sim} \left(\mu_F, \sigma_F^2\right) ,$$

onde:

$$\mu_F = E_F(X)$$

e

$$\sigma_F^2 = Var_F(X) = E_F[(X - \mu_F)^2] = E_F(X^2) - [E_F(X)]^2$$

Seja agora $\mathbf{x}=\{x_1, x_2, \cdots x_n\}$ uma amostra aleatória de tamanho n, da referida população, cuja distribuição é F. A média amostral ($\bar{x}$) terá esperança μ_F e variância $\left(\frac{\sigma_F^2}{n}\right)$. Isto é,

$$\bar{x} \overset{iid}{\sim} \left(\mu_F, \frac{\sigma_F^2}{n}\right) ,$$

onde:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

e

$$\sigma_F^2(\bar{x}) = Var_F(\bar{x}) = Var_F\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum^{n} Var_F(x_i)}{n^2} = \frac{n\sigma_F^2}{n^2} = \frac{\sigma_F^2}{n}$$

Como se pode observar, a média amostral tem a mesma esperança que X, no entanto, sua variância é $(1/n)$ da variância de X. Por esta razão diz-se que quanto maior o tamanho da amostra (n) menor será a variância da média e conseqüentemente, melhor a estimativa de μ_F.

O Erro-padrão da média, que é a medida de precisão mais comum, é obtido por:

$$\sigma_F(\bar{x}) = \sqrt{Var(\bar{x})} = \sqrt{\frac{\sigma_F^2}{n}} \tag{7}$$

Na maioria das vezes não se conhece o verdadeiro valor de σ_F^2, tendo-se apenas a sua estimativa ($\hat{\sigma}_F^2$). Logo, a estimativa não-viesada da variância da média é dado por:

$$\hat{\sigma}_F^2(\bar{x}) = \widehat{Var}(\bar{x}) = \frac{\sum^n (x_i - \bar{x})^2}{(n-1)n} = \frac{s^2}{n} \tag{8}$$

E, conseqüentemente, a estimativa do erro padão da média é obtida por:

$$\hat{\sigma}_F(\bar{x}) = \sqrt{\widehat{Var}(\bar{x})} = \sqrt{\frac{\sum^n (x_i - \bar{x})^2}{(n-1)n}} = \sqrt{\frac{s^2}{n}} \tag{9}$$

Quando n é suficientemente grande, independente da distribuição dos dados originais, a distribuição da média amostral aproxima-se de uma distribuição normal. Essa aproximação é garantida pelo teorema do limite central. Isto é:

$$\bar{x} \quad \dot{\sim} \quad N\left(\mu_F, \quad \frac{\sigma_F^2}{n}\right)$$

onde $\dot{\sim}$ indica distribuição aproximada e, quanto maior a normalidade dos dados originais maior a aproximação. Dessa forma, usando a tabela da distribuição normal, podem-se encontrar intervalos de confiança para μ_F ou realizar testes de hipóteses.

Não é objetivo demonstrar o teorema do limite central, no entanto, será ilustrado graficamente a distribuição amostral da média e da razão de dua variáveis aleatórias da população apresentada no Exemplo 2. Através das Figuras 2, 3 e 4 pode-se verificar que à medida que o tamanho da amostra aumenta a distribuição da média das variáveis aleatórias tende à normalidade, enquanto que a distribuição da razão de duas variáveis aleatórias não segue necessariamente a mesma regra.

Exemplo 2:

Seja uma população com as variáveis aleatórias: V={120, 50, 70, 80, 100, 20} e W={80, 20, 50, 50, 70, 30}. Dessa população extraíram-se amostras aleatórias, com reposição, de tamanhos n=1, n=2 e n=3. Os histogramas de freqüência são apresentados nas Figuras 2, 3 e 4.

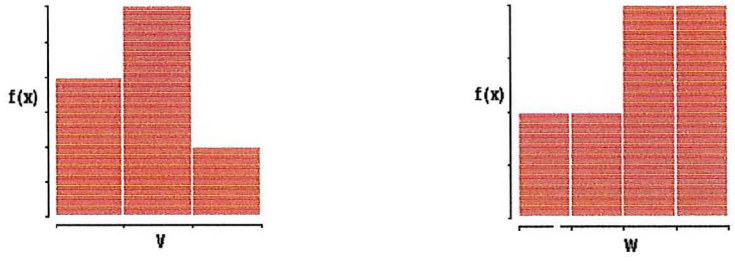

Figura 2: Histogramas das variáveis aleatórias V e W.

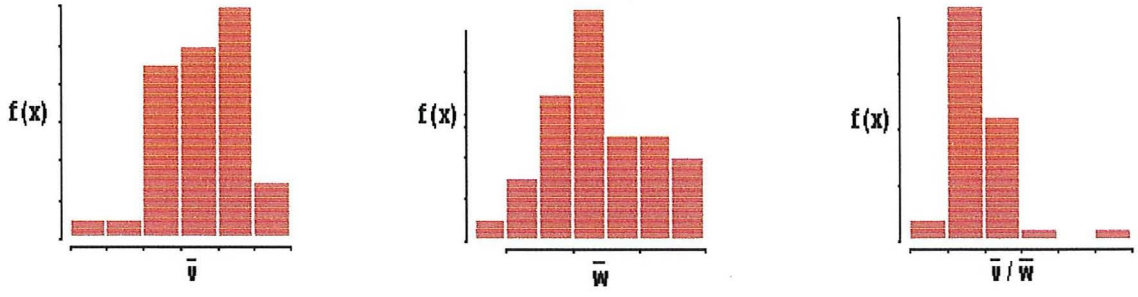

Figura 3: Histogramas das distribuições amostrais de $\bar{v}$, $\bar{w}$ e $\frac{\bar{v}}{\bar{w}}$ com n=2

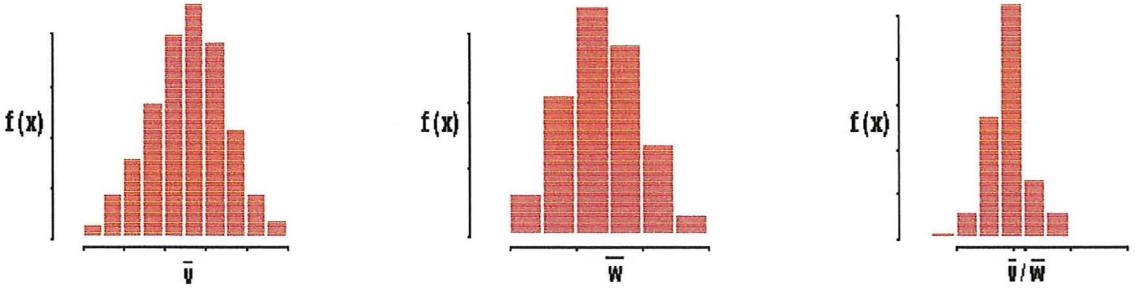

Figura 4: Histogramas das distribuições amostrais de $\bar{v}$, $\bar{w}$ e $\frac{\bar{v}}{\bar{w}}$ com n=3

Observando a expressão (9) pode-se verificar que $\hat{\sigma}_F(\bar{x})$ representa uma estimativa dos erros padrões das médias de todas as amostras de tamanho n, extraídas de uma população cuja distribuição é F. Isto quer dizer que, se houvesse possibilidades de se retirarem todas essas amostras, poder-se-ia calcular o valor de $\hat{\sigma}_F(\bar{x})$ diretamente pelos erros padrões das médias amostrais de cada uma dessas amostras.

Naturalmente, devido à simplicidade analítica das expressões (7) e (8) não faz sentido aplicar este princípio para obter a estimativa do erro padrão do estimador, quando este é a média amostral. Entretanto, em situações onde o interesse consiste em uma estatística que possui forma analítica mais complicada, como no caso da mediana, coeficiente de correlação, etc., a obtenção do erro padrão do estimador, diretamente pela variância dos valores amostrais, a partir de todas as amostras possíveis, é uma alternativa que pode ser adotada para evitar a complexidade que estes estimadores apresentam, quando se deseja obter uma medida de precisão para os mesmos.

O problema é que a distribuição F é desconhecida e portanto é impossível extrairem-se todas as amostras possíveis da população original. Todavia, é possível obterem-se repetidas amostras de outra população, cuja distribuição aproxima-se da verdadeira F. Esta é a idéia básica do método "Bootstrap", que consiste em se trocar a distribuição desconhecida F, que descreve uma população que não pode ser reamostrada, por uma distribuição empírica $\hat{F}$, que descreve uma população conhecida, que pode ser reamostrada exaustivamente.

3.5 Estimativa "Bootstrap" do Erro Padrão de um Estimador

Seja $\mathbf{x}=(x_1, x_2 \cdots x_n)$ uma amostra aleatória, de tamanho n, com distribuição de probabilidade F, desconhecida, isto é:

$$X_1, X_2 \cdots X_n \overset{iid}{\sim} F$$

Além disso, a esperança e variância de F são, respectivamente, μ_F e σ_F^2.

$$F(x) = Prob\{X_i \leq x_i\}$$

Suponha que se deseja estimar um parâmetro $\theta=t(F)$ da população F, com base nos dados da amostra observada, ou seja, $\hat{\theta}=s(\mathbf{x})$. Mas, como obter uma medida de precisão para $\hat{\theta}$ se a distribuição F da população que foi extraída a amostra é desconhecida ? Como foi visto anteriormente, em situações como esta, exceto para o caso da média, é complicado fazer inferências sobre o parâmetro θ. No entanto, através do método "Bootstrap", proposto por EFRON (1979), que substitui as soluções analíticas complexas ou inexistentes pelo poder de processamento dos computadores, podem-se fazer inferências sobre o parâmetro θ, de modo relativamente simples.

A seguir será descrito o procedimento "Bootstrap" para a obtenção da estimativa da variância da estatística $\bar{x}$. O uso da média amostral é para demonstrar a eficiência do método "Bootstrap", uma vez que a variância da mesma pode ser obtida através de expressões simples e dessa forma torna-se possível uma análise comparativa. Todavia, os conceitos aqui empregados poderão ser estendidos para outras estatísticas mais complexas.

3.5.1 Escolha de $\hat{F}$

O procedimento "Bootstrap" baseia-se no fato de se obter uma distribuição empírica $\hat{F}$, que reproduza o mecanismo probabilístico, gerador dos dados originais. $\hat{F}$ é um estimador de F que melhor se aproxima da distribuição verdadeira da população.

Dado um mínimo de suposições sobre a distribuição F, uma escolha para $\hat{F}$ é feita através do estimador de verossimilhança não-paramétrico, que coloca massa de probabilidade $\left(\frac{1}{n}\right)$ em cada uma das observações $x_1, x_2 \cdots x_n$, que representa a proporção amostral dos valores observados menores ou igual a x. Isto é:

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} 1_{(x_i \leq x)}}{n} = \frac{\#(x_i \leq x)}{n}$$

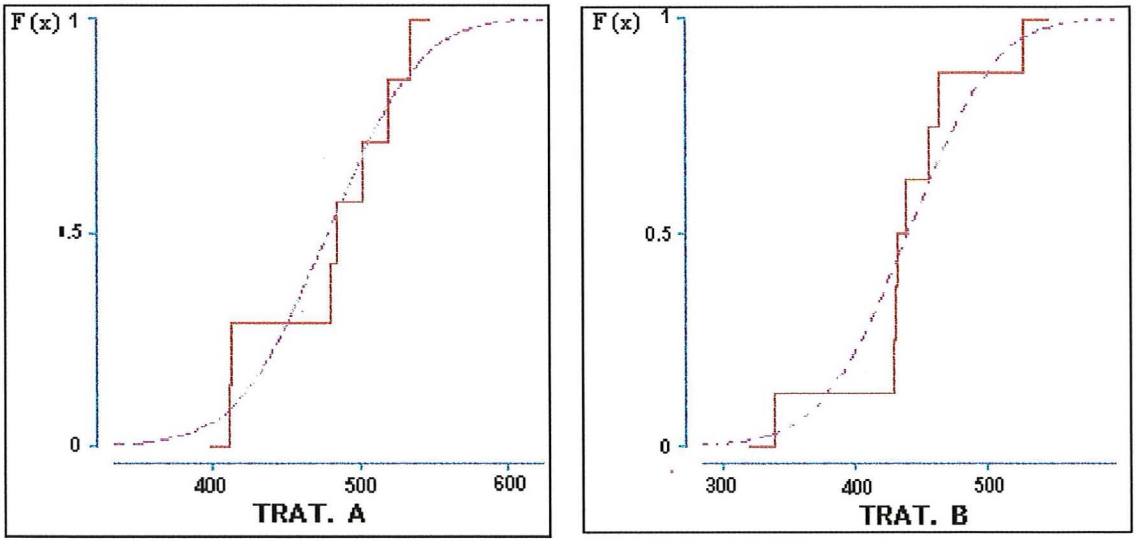

Figura 5: Dados do Exemplo 1. A linha cheia indica a distribuição empírica e a linha pontilhada, a distribuição acumulada dos dados $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

Os gráficos da Figura 5 mostram as distribuições empíricas $\hat{F}(x)$ para os dados dos tratamentos A e B (Exemplo 1), juntamente com as distribuições teóricas, que em ambos os tratamentos, trata-se da distribuição normal padronizada.

Através da Tabela 2, mostrar-se-á que as distribuições das médias amostrais seguem o princípio que coloca massa de probabilidade $\left(\frac{1}{n}\right)$ em cada observação. Nesta tabela encontram-se os valores observados para todas as possíveis amostras de tamanho n=2, da população apresentada no Exemplo 2, bem como, a distribuição de probabilidade das médias amostrais $\bar{v}$ e $\bar{w}$. Pode-se verificar que para a obtenção das referidas estatísticas, assim como a obtenção da esperança e variância de $\bar{V}$ e $\bar{W}$, considera-se o princípio que coloca massa de probabilidade $\left(\frac{1}{n}\right)$ em cada observação.

$$\bar{V} = \frac{\sum_{i=1}^{N} v_i}{N} = \frac{440}{6} = 73,33 \qquad e \qquad \bar{W} = \frac{\sum_{i=1}^{N} w_i}{N} = \frac{300}{6} = 50$$

$$E_C[\bar{V}] = \sum_{i=1}^{n} \bar{V}[pr(s)] = \frac{1}{36}(120) + \cdots + \frac{1}{36}(20) = 73,3 \qquad \therefore \qquad \bar{V} = E_C(\bar{V})$$

$$E_C[\bar{W}] = \sum_{i=1}^{n} \bar{W}[pr(s)] = \frac{1}{36}(80) + \cdots + \frac{1}{36}(30) = 50 \qquad \therefore \qquad \bar{W} = E_C(\bar{W})$$

$$Var[V] = \sum_{i=1}^{N} [\bar{V} - E_C(V)]^2 = 1055,5 \qquad e \qquad Var[W] = \sum_{i=1}^{N} [\bar{W} - E_C(W)]^2 = 433,33$$

$$Var_C(\bar{V}) = \sum_{i=1}^{n} [\bar{V} - E(\bar{V})]^2 [pr(s)] = (120 - 73,3)^2 \frac{1}{36} + \cdots + (20 - 73,3)^2 \frac{1}{36} = 527,78$$

$$\therefore \qquad Var_C(\bar{V}) = \frac{Var(\bar{V})}{n} = \frac{1055,5}{2} = 527,78$$

$$Var_C(\bar{W}) = \sum_{i=1}^{n} [\bar{W} - E(\bar{W})]^2 [pr(s)] = (80 - 50)^2 \frac{1}{36} + \cdots + (30 - 50)^2 \frac{1}{36} = 216,67$$

$$\therefore \qquad Var_C(\bar{W}) = \frac{Var(\bar{W})}{n} = \frac{433,33}{2} = 216,67$$

Tabela 2: Distribuição de probabilidades das possíveis amostras, de tamanho n=2, para as variáveis aleatórias V e W da população apresentada no Exemplo 2, extraídas sob um plano amostral de amostra aleatória com reposição (plano C)

Amostra $\{s\}$	Probabilidade de $\{s\}$	Variáveis $\{v_1$ e $v_2\}$	Valor obser. $\bar{v}$	Variáveis $\{w_1$ e $w_2\}$	Valor obser. $\bar{w}$
{1 e 1 }	1/36	{120 e 120 }	120	{80 e 80 }	80
{1 e 2 }	1/36	{120 e 50 }	85	{80 e 20 }	50
{1 e 3 }	1/36	{120 e 70 }	95	{80 e 50 }	65
{1 e 4 }	1/36	{120 e 80 }	100	{80 e 50 }	65
{1 e 5 }	1/36	{120 e 100 }	110	{80 e 70 }	75
{1 e 6 }	1/36	{120 e 20 }	70	{80 e 30 }	55
{2 e 1 }	1/36	{ 50 e 120 }	85	{20 e 80 }	50
{2 e 2 }	1/36	{ 50 e 50 }	50	{20 e 20 }	20
{2 e 3 }	1/36	{ 50 e 70 }	60	{20 e 50 }	35
{2 e 4 }	1/36	{ 50 e 80 }	65	{20 e 50 }	35
{2 e 5 }	1/36	{ 50 e 100 }	75	{20 e 70 }	45
{2 e 6 }	1/36	{ 50 e 20 }	85	{20 e 30 }	25
{3 e 1 }	1/36	{ 70 e 120 }	95	{50 e 80 }	65
{3 e 2 }	1/36	{ 70 e 50 }	60	{50 e 20 }	35
{3 e 3 }	1/36	{ 70 e 70 }	70	{50 e 50 }	50
{3 e 4 }	1/36	{ 70 e 80 }	75	{50 e 50 }	50
{3 e 5 }	1/36	{ 70 e 100 }	85	{50 e 70 }	75
{3 e 6 }	1/36	{ 70 e 20 }	45	{50 e 30 }	40
{4 e 1 }	1/36	{ 80 e 120 }	100	{50 e 80 }	65
{4 e 2 }	1/36	{ 80 e 50 }	65	{50 e 20 }	35
{4 e 3 }	1/36	{ 80 e 70 }	75	{50 e 50 }	50
{4 e 4 }	1/36	{ 80 e 80 }	80	{50 e 50 }	50
{4 e 5 }	1/36	{ 80 e 100 }	90	{50 e 70 }	60
{4 e 6 }	1/36	{ 80 e 20 }	50	{50 e 30 }	40
{5 e 1 }	1/36	{100 e 120 }	110	{70 e 80 }	75
{5 e 2 }	1/36	{100 e 50 }	75	{70 e 20 }	45
{5 e 3 }	1/36	{100 e 70 }	85	{70 e 50 }	60
{5 e 4 }	1/36	{100 e 80 }	90	{70 e 50 }	60
{5 e 5 }	1/36	{100 e 100 }	100	{70 e 70 }	70
{5 e 6 }	1/36	{100 e 20 }	60	{70 e 30 }	50
{6 e 1 }	1/36	{ 20 e 120 }	70	{30 e 80 }	55
{6 e 2 }	1/36	{ 20 e 50 }	35	{30 e 20 }	25
{6 e 3 }	1/36	{ 20 e 70 }	45	{30 e 50 }	40
{6 e 4 }	1/36	{ 20 e 80 }	50	{30 e 50 }	40
{6 e 5 }	1/36	{ 20 e 100 }	60	{30 e 70 }	50
{6 e 6 }	1/36	{ 20 e 20 }	20	{30 e 30 }	30

$\bar{v}$	20	35	45	20	60	65	70	75	80	85	90	95	100	110	120
$prob(\bar{V}=\bar{v})$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{5}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$\bar{w}$	20	25	30	35	40	45	50	55	60	65	70	75	80
$prob(\bar{W}=\bar{w})$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{8}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

3.5.2 Estimativas "Bootstrap" do Erro-padrão da média

A partir da escolha de $\hat{F}$ (estimador de verossimilhança não-paramétrico) e da amostra observada $\mathbf{x}=(x_1, x_2 \cdots x_n)$, inicia-se o processo "Bootstrap". Isto é, com $\hat{F}$ fixa, seleciona-se na amostra observada, uma amostra aleatória $\mathbf{x}^*=(x_1^*, x_2^* \cdots x_n^*)$, com reposição, de tamanho n. Essa amostra $\mathbf{x}^*$ é denominada **Amostra "Bootstrap"** e caracteriza-se por reuzar a própria amostra observada $\mathbf{x}$. Logo,

$$X_1^*, X_2^* \cdots X_n^* \overset{iid}{\sim} \hat{F}$$

De posse da amostra "Bootstrap", estimativas "Bootstrap" `Exata` das estatísticas de interesse serão obtidas, realizando-se nessa amostra sob distribuição $\hat{F}$, as mesmas operações matemáticas que seriam feitas na amostra original sob distribuição F. Considerando o caso em que o parâmetro de interesse $\theta = t(F)$ é a média populacional, estimada por $\hat{\theta} = s(\mathbf{x})$. Então,

$$\sigma_F^2(\bar{x}) = \frac{\sigma_F^2}{n} = \frac{E_F[x - E_F(x)]^2}{n} = \frac{E_F(x)^2 - [E_F(x)]^2}{n}$$

Como F é desconhecida torna-se impossível obter os valores dos momentos $E_F(x)^2$ e $[E_F(x)]^2$. Mas trocando-se F por $\hat{F}$ e $\mathbf{x}$ por $\mathbf{x}^*$ obtém-se a estimativa "Bootstrap" `Exata` da variância de $\bar{x}$, conforme descrito abaixo.

$$\hat{\sigma}^2_{Boot}(\bar{x}) = \sigma^2_{\hat{F}}(\bar{x}) = \frac{\sigma^2_{\hat{F}}}{n} = \frac{E_{\hat{F}}(x^*)^2 - [E_{\hat{F}}(x^*)]^2}{n} =$$

$$= \frac{\frac{\sum x_i^2}{n} - \bar{x}^2}{n} = \frac{\sum^n (x_i - \bar{x})^2}{n^2} = \frac{(n-1)}{n} \frac{s^2}{n}$$

Portanto,

$$\hat{\sigma}^2_{Boot}(\bar{x}) = \frac{(n-1)}{n} \frac{s^2}{n} = \left(\frac{n-1}{n}\right) \hat{\sigma}^2_F(\bar{x}) \tag{10}$$

E, consequentemente,

$$\hat{\sigma}_{Boot}(\bar{x}) = \sqrt{\frac{(n-1)\, s^2}{n^2}}$$

Comparando-se a expressão (10) com a expressão (8), pode-se verificar que a expressão da estimativa "Bootstrap" **exata** da variância da média é semelhante à expressão da estimativa **não viesada** da variância de $\bar{x}$, a menos do fator $\frac{(n-1)}{n}$.

A estimativa ""Bootstrap" **exata**, para o caso da média é fácil de ser obtida, mas o mesmo não ocorre quando a estatística de interesse possui uma fórmula analítica complicada. Nesses casos, pode-se aplicar o princípio visto anteriormente, ou seja, obter o erro padrão do estimador, diretamente pelos valores amostrais a partir das variâncias estimadas de todas as amostras possíveis. Todavia, este procedimento só poderá ser adotado quado se conhece a população, para que seja possível se extrairem todas as amostras possiveis.

No o caso em questão, a distribuição F é desconhecida e portanto não se pode extrairem todas as amostras possíveis da população original. Mas, adotando-se o método "Bootstrap", que troca a distribuição F, desconhecida, por seu estimador $\hat{F}$, que descreve uma população que pode ser reamostrada exaustivamente, pode-se obter a estimativa do erro padrão do estimador diretamente pelos valores das variâncias amostrais, como sugerido no parágrafo anterior.

Segundo EFRON & TIBSHIRANI (1993), para uma amostra observada de tamanho n tem-se $m = \binom{2n-1}{n}$ amostras "Bootstrap" distintas. Onde m representa o número de amostras não ordenadas de tamanho n extraídas com reposição. Por exemplo: sendo n=2, $\{x_1$ e $x_2\}$, então m=3, que são as amostras, $\{x_1$ e $x_1\}$, $\{x_1$ e $x_2\}$ e $\{x_2$ e $x_2\}$, já que as amostras $\{x_1$ e $x_2\}$ e $\{x_2$ e $x_1\}$ são iguais.

Dessa forma, obtendo-se a estatística de interesse de cada **amostra "Bootstrap"**, para todas as m amostras "Bootstrap" distintas tem-se a estimativa "Bootstrap" **Exata** para o parâmetro de interesse. Sejam as m amostras "Bootstrap" distintas denotadas por $\mathbf{z}^*{}_j = \{z_1^*, z_2^* \cdots z_m^*\}$, então a estimativa "Bootstrap" **exata**

da variância de uma estatística de interesse $\hat{\theta}=s(\mathbf{x})$ é obtida por:

$$Var_{\hat{F}}(\hat{\theta}^*) = \sigma^2_{\hat{F}}(\hat{\theta}^*) = \hat{\sigma}^2_{\hat{F}}(\hat{\theta}) = \sum_{j=1}^{m}[t(z_j^*) - t(\cdot)]^2 W_j$$

onde,

$$t(\cdot) = \sum_{j=1}^{m} W_j[t(z_j^*)]$$

W_j, com $j = 1, 2 \cdots \binom{2n-1}{n}$, é a probabilidade da j-ésima amostra "Bootstrap" distinta, com $W_j = \frac{n!}{j_1!,j_2!\cdots j_n!} \prod_{i=1}^{n} (\frac{1}{n})^{j_i}$ (EFRON & TIBSHIRANI, 1993)

O grande problema deste método é que m cresce muito rapidamente. Por exemplo, para n=20 têm-se 68.923.264.410 amostras "Bootstrap" distintas. Isto torna impraticável a obtenção da estimativa "Bootstrap" `exata`. Felizmente, é sempre possível obter, através de simulação de Monte Carlo, descrito no Algoritmo 1, da Figura 6, a estimativa "Bootstrap", que é uma aproximação numérica para a estimativa "Bootstrap" `exata`.

Na prática, usa-se sempre o algoritmo de Monte Carlo para obter a estimativa "Bootstrap" do erro padrão de uma estatística de interesse. Neste casos é importante tomar amostras "Bootstrap" do mesmo tamanho da amostra original, pois, segundo (EFRON & TIBSHIRANI, 1986), o algoritmo de Monte Carlo não converge para estimativa "Bootstrap" `Exata` se o tamanho da amostra "Bootstrap" difere da amostra original.

Quanto ao número de amostras "Bootstrap" (**B**), necessária para estimar o erro padrão, através do algorítimo de Monte Carlo, EFRON & TIBSHIRANI (1993), apresentam a seguinte regras práticas: **B**=25 é usualmente informativo; **B**=50 é suficiente para se obter a estimativa "Bootstrap" com boa precisão e raramente necessita-se $\mathbf{B} > 200$.

Figura 6: Algoritmo de Monte Carlo para obter a estimativa "Bootstrap" do erro padrão de uma estatística de interesse

Algoritmo 1

1) Através de um gerador de números aleatórios, seleciona-se, independentemente, um grande número **B** de amostras "Bootstrap", $\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)}, \cdots \mathbf{x}^*_{(B)}$. Isto é, **B** amostras aleatórias simples de tamanho n , extraídas com reposição da amostra original $\mathbf{x} = \{x_1, x_2, \cdots x_n\}$;

2) Para cada amostra "Bootstrap" $(\mathbf{x}^*_{(b)})$ calcula-se a estatística de interesse $\hat{\theta}^*_{(b)} = s(\mathbf{x}^*_{(b)})$ para $b = 1, 2, \cdots B$

3) Calcula-se a variância amostral dos **B** valores $\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)} \cdots \hat{\theta}^*_{(B)}$, isto é,

$$\hat{\sigma}^2_{\mathbf{B}}(\hat{\theta}) = \frac{\sum_{b=1}^{B} \{\hat{\theta}^*_{(b)} - \hat{\theta}^*(\cdot)\}^2}{\mathbf{B} - 1}$$

onde,

$$\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*_{(b)}}{\mathbf{B}}$$

4) A estimativa "Bootstrap" do erro padrão da estatística de interesse é obtido por:

$$\hat{\sigma}_{\mathbf{B}}(\hat{\theta}) = \sqrt{\hat{\sigma}^2_{\mathbf{B}}(\hat{\theta})}$$

Pode-se verificar que quando $\mathbf{B} \to \infty \Rightarrow \hat{\sigma}^2_{\mathbf{B}}(\hat{\theta}) \to \hat{\sigma}^2_{Boot}(\hat{\theta}) = \sigma^2(\hat{F})$.

Na Figura 7 encontram-se os resultados de um estudo de simulação, aplicando-se o Algoritmo 1 para observar a convergência das estimativas "Bootstrap" dos erros padrões das média $\bar{Y}$ e $\bar{Z}$, dadas no Exemplo 1, para as estimativas "Bootstrap" `exatas`. Neste estudo foram considerados valores de **B** de 25 até 600.

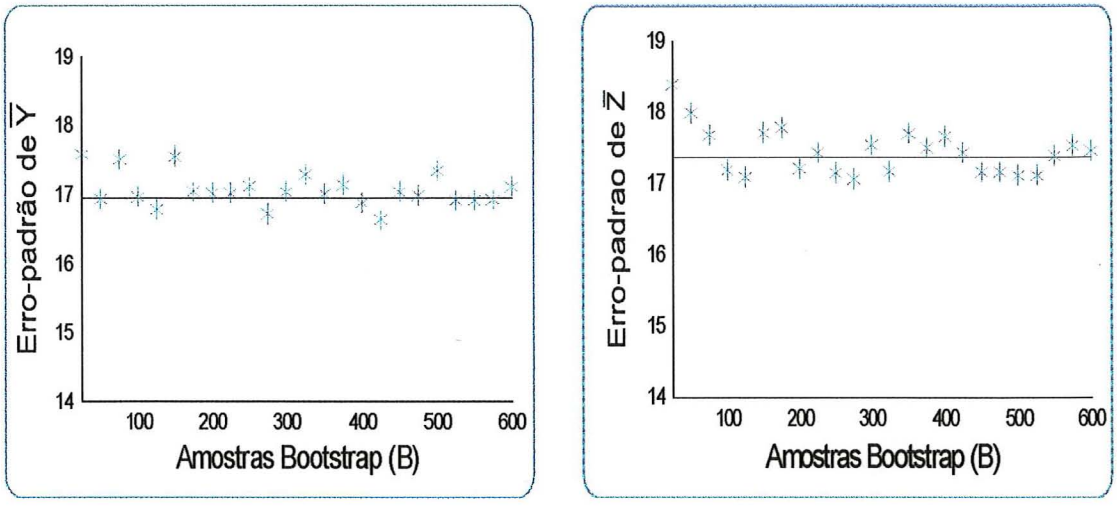

Figura 7: Estimativas "Bootstrap" do erro padrão de $\bar{Y}$ e $\bar{Z}$, em função do número de amostras "Bootstrap" (**B**). A linha horizontal da primeira e segunda figura representam as estimativa "Bootstrap" `exatas`, respectivamente, $16,95$ e $17,37$

Na Tabela 3 encontram-se os valores dos parâmetros das populações apresentadas no Exemplo 1, juntamente com as estimativas não viesadas, estimativas "Bootstrap" `exatas` e as estimativas "Bootstrap" obtidas através do Algoritmo 1 com **B**=200. Diante dos valores apresentados nesta tabela, pode-se verificar que as estimativas "Bootstrap" apresentam valores muito próximos às estimativas "Bootstrap" `exatas` e estimativas não viesadas.

Tabela 3: Resultados comparativos, obtidos a partir dos dados do Exemplo 1

Estatística de interesse	Parâmetro populacional	Estimativa não viesada	Estimativa `Bootstrap Exata`	Estimativa `Bootstrap`
$\hat{\theta}$	$\sigma_F(\hat{\theta}) = \sqrt{\frac{\sigma_F^2}{n}}$	$\hat{\sigma}_F(\hat{\theta}) = \sqrt{\frac{s^2}{n}}$	$\hat{\sigma}_{Boot}(\hat{\theta}) = \sqrt{\frac{(n-1)\ s^2}{n^2}}$	$\hat{\sigma}_{\mathbf{B}}(\hat{\theta})$
$\bar{y}$	18,90	18,31	16,95	17,02
$\bar{z}$	17,68	18,57	17,37	17,22

3.6 Estimativa "Bootstrap" do Erro Padrão do Coeficiente de Correlação

Considere o parâmetro de interesse $\theta=t(F)$ como sendo o coeficiente de correlação linear entre duas variáveis aleatórias Y e Z, que pode ser estimada a partir dos dados observados através da estatística $\hat{\theta}=s(\bar{x})$. De posse desta estatística, necessita-se agora obter uma medida de precisão para a mesma, afim de que se possa fazer inferências sobre as características da população. Mas como foi visto anteriormente, este é um caso em que há dificuldades na obtenção de uma medida de precisão para $\hat{\theta}$, uma vez que para o coeficiente de correlação linear não existe uma expressão simples que forneça o erro padrão de $\hat{\theta}$, como a expressão $\sqrt{\frac{s^2}{n}}$, que fornece o erro padrão da média amostral.

Todavia, é sempre possível obter o erro padrão de um estimador, diretamente pelos valores amostrais. E, aplicando-se o procedimento "Bootstrap", através do algoritmo de Monte Carlo, mostrado na Figura 8, pode-se obter uma estimativa "Bootstrap" do erro padrão do coeficiente de corelação linear entre duas variáveiss aleatórias $(\widehat{corr}(y,z))$.

Exemplo 3:

Este é um exemplo apresentado por EFRON & TIBSHIRANI (1993), que se refere ao estudo em 82 faculdades americanas de Direito. Nesse estudo, foram consideradas as variáveis aleatórias LSAT[3] e GPA [4]. Os dado na Tabela 4, referem-se aos valores observados das variáveis aleatórias Y e Z de uma amostra de 15 faculdades e a partir desses dados, obteve-se o valor $0,779$ para o coeficiente de correlação amostral.

$$\widehat{corr}(y,z) = \frac{\sum_{i=1}^{n} y_i z_i - n\bar{y}\bar{z}}{\sqrt{(\sum y_i^2 - n\bar{y}^2)(\sum z_i^2 - n\bar{z}^2)}} = 0,776$$

[3]LSAT - nota média no exame para ser admitido na faculdade de Direito

[4]GPA - nota média na escola anterior

Tabela 4: Dados observados de uma amostra (n=15) das faculdades americanas de Direito, referentes às variáveis `LSAT` e `GPA`

Faculdade	LSAT y_i	GPA z_i	Faculdade	LSAT y_i	GPA z_i
1	576	3,39	9	651	3,36
2	635	3,30	10	605	3,13
3	558	2,81	11	653	3,12
4	578	3,03	12	575	2,74
5	666	3,44	13	545	2,76
6	580	3,07	14	572	2,88
7	555	3,00	15	594	2,96
8	661	3,43	–	–	–

FONTE: Efron & Tibshirani,(1993)

Figura 8: Algoritmo de Monte Carlo para obter a estimativa "Bootstrap" do erro padrão do coeficiente de correlação linear

Algoritmo 2

1) Através de um gerador de números aleatórios, selecionam-se, independentemente, um grande número **B** de amostras "Bootstrap", $\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)}, \cdots \mathbf{x}^*_{(B)}$. Isto é, **B** amostras aleatórias simples de tamanho n , extraídas com reposição da amostra atual $\mathbf{x} = \{x_1, x_2, \cdots x_n\}$;

2) Para cada amostra "Bootstrap" $(\mathbf{x}^*_{(b)})$ calcula-se o respectivo coeficiente de correlação amostral $\widehat{corr}^*_1, \widehat{corr}^*_2, \cdots \widehat{corr}^*_B$

3) Calcula-se a variância amostral dos **B** valores $\widehat{corr}^*_1, \widehat{corr}^*_2, \cdots \widehat{corr}^*_B$, isto é:

$$\hat{\sigma}^2_{\mathbf{B}}(\widehat{corr}) = \frac{\sum_{b=1}^{B} \{\widehat{corr}^*_{(b)} - \hat{\theta}^*(\cdot)\}^2}{\mathbf{B}-1}$$

onde,
$$\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \widehat{corr}^*_{(b)}}{\mathbf{B}}$$

4) A Estimativa "Bootstrap" do Erro-padrão do coeficiente de correlação linear é obtido por:

$$\hat{\sigma}_{\mathbf{B}}(\widehat{corr}) = \sqrt{\hat{\sigma}^2_{\mathbf{B}}(\widehat{corr})}$$

Pode-se verificar que quando $\mathbf{B} \to \infty \Rightarrow \hat{\sigma}^2_{\mathbf{B}}(\widehat{corr}) \to \hat{\sigma}^2_{Boot}(\widehat{corr}) = \sigma^2(\hat{F})$.

Tabela 5: Estimativas "Bootstrap" do erro padrão do coeficiente de correlação linear para diversos tamanho de (**B**)

B	25	50	100	200	400	800	1600	3200
$\hat{\sigma}_{\mathbf{B}}(\widehat{corr})$	0,140	0,142	0,151	0,143	0,141	0,137	0,133	0,132

FONTE: Efron & Tibshirani,(1993)

Na Tabela 5 encontram-se as estimativas "Bootstrap" do erro padrão do coeficiente de correlação linear para os dados do Exemplo 3, com o número de amostras "Bootstrap" variando de 25 a 3200. Como a população da qual foi extraída a amostra é conhecida, é possível comparar os histogramas de distribuições amostrais, obtidas pelos dois procedimentos, ou seja, comparar o histograma da distribuição amostral dos dados originais para amostras de tamanho n=15, com o histograma da distribuição amostral para o caso em que foi usado o procedimento "Bootstrap" com 3200 amostras "Bootstrap", (**B**=3200). Como se pode observar, através da Figura 9, existe grande semelhança entre a distribuição "Bootstrap" e a distribuição amostral dos dados originais.

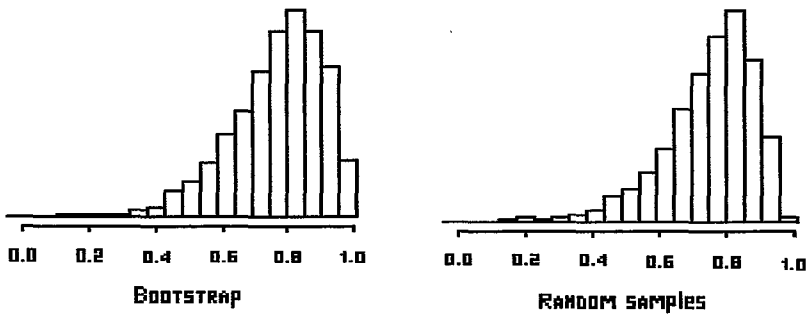

Figura 9: O painel esquerdo mostra o histograma da distribuição "Bootstrap" de $\widehat{corr}(\mathbf{x}*)$ com 3200 amostras "Bootstrap". O painel direito mostra o histograma da distribuição amostral de 3200 repetições do $\widehat{corr}(\mathbf{x})$, onde $\mathbf{x}$ é uma amostra aleatória da população de N=82 faculdades americanas. FONTE: Efron & Tibshirani,(1993).

3.7 Estimativa "Bootstrap" do Viés e Variância do Estimador Razão

3.7.1 Estimador do Tipo Razão de Média

Na inferência estatística é comum encontrar situações em que o parâmentro de interesse é uma razão (R), formada pelo quociente entre duas variáveis aleatória (por exemplo, V e W). O parâmetro R é estimado pela razão $\hat{R}$ estabelecida na amostra.

$$\hat{R} = \frac{\bar{v}}{\bar{w}} \quad \text{é um estimador de} \quad R = \frac{\mu_V}{\mu_W} = \frac{\bar{V}}{\bar{W}} \quad \text{Logo,} \quad \bar{V} = R\bar{W}$$

A distribuição de $\hat{R}$ é, em geral, bastante complexa de ser obtida, pois o denominador tabém é uma variável aleatória. Como conseqüencia deste fato, o estimador $\hat{R}$ possui viés e tem distribuição bastante assimétrica em pequenas amostras. O viés de $\hat{R}$ diminui à medida que se aumenta a amostra e, para grandes amostras, a distribuição aproxima-se da normal (BOLFARINE & BUSSAB, 1994).

Segundo COCHRAN (1977), não existem expressões exatas para o cálculo do viés e nem para o cálculo da variância amostral da estatística $\hat{R}$, apenas aproximações válidas para grandes amostras.

A expressão aproximada do viés de $\hat{R}$ pode ser obtida, segundo COCHRAN (1977), como se segue:

$$[\hat{R} - R] = \left[\frac{\bar{v}}{\bar{w}} - R\right] = \left[\frac{\bar{v} - R\bar{w}}{\bar{w}}\right]$$

Expandindo $\frac{1}{\bar{w}}$, conforme a expansão de Taylor, temos:

$$\frac{1}{\bar{w}} = \frac{1}{\bar{W}\left(1 + \frac{\bar{w}-\bar{W}}{\bar{W}}\right)} = \frac{1}{\bar{W}}\left(1 - \frac{\bar{w} - \bar{W}}{\bar{W}} + \frac{(\bar{w} - \bar{W})^2}{\bar{W}^2} - \cdots +\right)$$

Então,

$$[\hat{R} - R] \doteq \frac{\bar{v} - R\bar{w}}{\bar{W}}\left(1 - \frac{\bar{w} - \bar{W}}{\bar{W}} + \frac{(\bar{w} - \bar{W})^2}{\bar{W}^2} - \cdots\right)$$

Portanto:

$$[\hat{R} - R] \doteq \left(\frac{\bar{v} - R\bar{w}}{\bar{W}} - \frac{(\bar{v} - R\bar{w})(\bar{w} - \bar{W})}{\bar{W}^2} + \cdots \right) \tag{11}$$

onde, $\doteq$ indica aproximação. Para grandes amostras temos que $\bar{w} \cong \bar{W}$. Considerando $\bar{w} = \bar{W}$, o segundo termo da Expressão (11) desaparecerá e $E[\hat{R} - R] = 0$. Logo, para grandes amostras, $\hat{R}$ é um estimador assintoticamente não viesado para R.

$$E[\hat{R} - R] \doteq E\left[\frac{\bar{v} - R\bar{w}}{\bar{W}}\right] \doteq \frac{1}{\bar{W}} E[\bar{v} - R\bar{w}] \doteq \frac{1}{\bar{W}}(\bar{V} - R\bar{W}) \doteq 0$$

Considerando uma amostra pequena, onde $\bar{w} \neq \bar{W}$, a expressão aproximada para o viés de $\hat{R}$ é dado por:

$$E[\hat{R} - R] \doteq E\left[\frac{\bar{v} - R\bar{w}}{\bar{W}} - \frac{(\bar{v} - R\bar{w})(\bar{w} - \bar{W})}{\bar{W}^2}\right] \doteq 0 - E\left[\frac{(\bar{v} - R\bar{w})(\bar{w} - \bar{W})}{\bar{W}^2}\right]$$

$$\doteq -E\left[\frac{-R\bar{w}(\bar{w} - \bar{W}) + \bar{v}(\bar{w} - \bar{W})}{\bar{W}^2}\right] \doteq \frac{1}{\bar{W}^2} E\left[R\bar{w}(\bar{w} - \bar{W}) - \bar{v}(\bar{w} - \bar{W})\right]$$

$$\doteq \frac{1}{\bar{W}^2}\left(E\left[R\bar{w}(\bar{w} - \bar{W})\right] - E\left[\bar{v}(\bar{w} - \bar{W})\right]\right) \doteq \frac{1}{\bar{W}^2}\left(R\, E\left[\bar{w}(\bar{w} - \bar{W})\right] - E\left[\bar{v}(\bar{w} - \bar{W})\right]\right)$$

Como

$$E\left[\bar{w}(\bar{w} - \bar{W})\right] = E\left[\bar{w} - \bar{W}\right]^2 = Var(\bar{w}) \quad = \quad \frac{s_w^2}{n}$$

e

$$E\left[\bar{v}(\bar{w} - \bar{W})\right] = E\left[(\bar{v} - \bar{V})(\bar{w} - \bar{W})\right] = Cov(\bar{v}, \bar{w}) \quad = \quad \rho(\bar{v}, \bar{w}) s_{\bar{v}} s_{\bar{w}}$$

Portanto,

$$\boxed{E[\hat{R} - R] \doteq \frac{N - n}{N} \frac{1}{n\bar{W}^2}\left(R s_w^2 - \rho(v, w) s_v s_w\right)}$$

onde, $\frac{N-n}{N}$ é a correção de população finita e $\rho(v, w)$ é o coeficiente de correlação amostral entre as variáveis V e W.

A expressão aproximada para a variância de $\hat{R}$, segundo COCHRAN (1977), é obtida como se segue:

$$Var(\hat{R}) \doteq EQR(\hat{R}) = E[\hat{R} - R]^2$$

Quando n é suficientemente grande (segundo COCHRAN, $n \geq 30$) $\bar{w} \doteq \bar{W}$ e tem-se apenas o primeiro termo da Expressão (11). Logo,

$$[\hat{R} - R] \doteq \frac{\bar{v} - R\bar{w}}{\bar{w}} \doteq \frac{\bar{v} - R\bar{w}}{\bar{W}} = \frac{1}{\bar{W}}\left(\frac{\sum v_i}{n} - \frac{R\sum w_i}{n}\right)$$

$$\therefore \qquad [\hat{R} - R] \doteq \frac{1}{\bar{W}} \frac{\sum^{n}(v_i - Rw_i)}{n} = \frac{1}{\bar{W}} \frac{\sum^{n} d_i}{n} \doteq \frac{\bar{d}}{\bar{W}},$$

onde $d_i = (v_i - Rw_i)$. Para o caso de populações, temos que:

$$\bar{D} = \frac{\sum^{N} d_i}{N} = \frac{\sum\limits_{i=1}^{N}(v_i - Rw_i)}{N} = \frac{\sum^{N} v_i - R\sum^{N} w_i}{N} = \frac{\left(V - \frac{V}{W}W\right)}{N} = 0$$

Logo, a variância de $\hat{R}$ pode ser escrita da seguinte forma:

$$Var(\hat{R}) \doteq E\left[\frac{(\bar{d} - \bar{D})}{\bar{W}}\right]^2 \doteq \frac{1}{\bar{W}^2} E\left[(\bar{d} - \bar{D})\right]^2 \doteq \frac{1}{\bar{W}^2}\, Var(\bar{d}) \doteq \frac{1}{\bar{W}^2}\, \frac{Var(d_i)}{n}$$

$$\doteq \frac{1}{\bar{W}^2}\, \frac{\frac{\sum(d_i - \bar{D})^2}{N-1}}{n} \doteq \frac{1}{\bar{W}^2}\, \frac{\sum^{N}(d_i)^2}{n(N-1)} \doteq \frac{1}{\bar{W}^2}\, \frac{\sum^{N}(v_i - Rw_i)^2}{n(N-1)}$$

Sua estimativa amostral é obtida por:

$$\widehat{Var}(\hat{R}) \doteq \frac{N-n}{N}\frac{1}{\bar{w}^2}\, \frac{\sum^{n}\left(v_i - \hat{R}w_i\right)^2}{n(n-1)} \doteq$$

$$\doteq \frac{N-n}{N}\frac{1}{n\bar{w}^2}\, \frac{\sum^{n} v_i^2 - 2\hat{R}\sum^{n} w_i v_i + \hat{R}^2 \sum^{n} w_i^2}{n-1}$$

onde, $\frac{N-n}{N}$ é a correção de população finita

3.7.2 Estimativa "Bootstrap" do Viés de Um Estimador

Suponha a seguinte situação clássica de análise de dados: uma amostra aleatória $\mathbf{x}=\{x_1, x_2, \cdots x_n\}$ foi extraída de uma população, cuja distribuição de probabilidades F é desconhecida. Com base no dados da amostra observada $\mathbf{x}$, deseja-se estimar um parâmetro de interesse $\theta=t(F)$. Para tanto, calcula-se a estatística $\hat{\theta}=s(\mathbf{x})$. Agora necessita-se saber a precisão de $\hat{\theta}$.

Até aqui, havia-se empregado o erro padrão como medida de acuracidade de um estimador, entretanto, existem outras medidas usadas na estatística para avaliar a precisão de um estimador. A seguir abordar-se-ão as definições de `tendenciosidade de um estimador` como medida de acuracidade, bem como a aplicação do procedimento "Bootstrap" para obtenção do viés de um estimador do tipo razão.

Considere agora a estatística $\hat{\theta}=s(\mathbf{x})$ como sendo a estimativa do viés do estimador em relação a um parâmetro populacional $\theta=t(F)$. Por definição, o viés de $\hat{\theta}=s(\mathbf{x})$ é a diferença entre a esperança de $\hat{\theta}$ e o valor do parâmetro populacional θ.

$$\text{Viés}_F = \text{Viés}(\hat{\theta},\ \theta) = E_F[s(\mathbf{x})] - t(F) \tag{12}$$

Embora o parâmetro θ seja desconhecido, $t(F)$ pode ser calculado. Por exemplo: Quando θ é a média populacional $t(F)=E_F(X)=\int x\, dF(x)$. Portanto, se $E_F(\hat{\theta})=\theta$ então $\hat{\theta}$ é um estimador não viesado pora θ.

Seguindo o mesmo processo descutido nas seções anteriores, o método "Bootstrap" pode ser usado para estimar o viés de qualquer estimador $\hat{\theta}=s(\mathbf{x})$. A estimativa "Bootstrap" do viés é definida como sendo o estimador $\text{Viés}_{\hat{F}}$ obtido quando se substitui F por $\hat{F}$ e $\mathbf{x}$ por $\mathbf{x}^*$ na Expressão (12). Isto é:

$$\text{Viés}_{\hat{F}} = E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F}) \tag{13}$$

Onde $t(\hat{F})$ é um estimador "Bootstrap" de $\hat{\theta}$ e $E_{\hat{F}}[s(\mathbf{x}^*)]$ pode ser obtido por simulação de Monte Carlo, como no Algoritmo 3, apresentado na Figura 10. $\text{Viés}_{\hat{F}}$ é um estimador "Bootstrap" de Viés_F.

Figura 10: Algoritmo de Monte Carlo para se obter a estimativa "Bootstrap" do viés de um Estimador

Algoritmo 3

1) Através de um gerador de números aleatórios, selecionam-se, independentemente, um grande número **B** de amostras "Bootstrap" $\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)}, \cdots \mathbf{x}^*_{(B)}$. Isto é, **B** amostras aleatórias simples de tamanho n , extraídas com reposição da amostra atual $\mathbf{x} = \{x_1, x_2, \cdots x_n\}$;

2) Para cada amostra "Bootstrap" $(\mathbf{x}^*_{(b)})$ calcula-se a estatística $s(\mathbf{x}^*_{(b)})$

3) Calcula-se a média dos **B** valores $s(\mathbf{x}^*_{(1)}), s(\mathbf{x}^*_{(2)}) \cdots s(\mathbf{x}^*_{(B)})$ isto é:

$$\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} s(\mathbf{x}^*_{(b)})}{\mathbf{B}}$$

4) A `Estimativa ''Bootstrap"` do Viés é obtida por:

$$\widehat{\text{Viés}}_{\mathbf{B}} = \hat{\theta}^*(\cdot) - t(\hat{F})$$

A obtenção da estimativa Boostrap da variância do estimador Razão é feita seguindo os mesmos passos do Algoritmo 1.

Exemplo 4:

Suponha que foram extraídas as amostras $\mathbf{x}_{v_i}$={20, 100, 120} e $\mathbf{x}_{w_i}$={50, 20, 70} da população descrita no Exemplo 2. E, a partir das amostras observadas, desejam-se estudar as propriedades de $\hat{R}$ como um estimador da razão entre as variáveis aleatórias V e W.

$$R = \frac{\mu_V}{\mu_W} = \frac{\bar{V}}{\bar{W}} = \frac{73,33}{50} = 1,4667$$

$$\hat{R} = \frac{\bar{v}}{\bar{w}} = \frac{80}{46,67} = 1,714$$

$$\text{Viés}(\hat{R}) = E[\hat{R}] - R = 1,47225 - 1,4667 = 0,00558$$

$$\widehat{\text{Viés}}(\hat{R}) = \frac{6-3}{6}\frac{1}{3(50)^2}(1,467 * 633,33 - 0,075 * 52,915 * 25,166) = 0,055$$

$$Var(\hat{R}) = E[\hat{R} - R]^2 = 0,045$$

$$\widehat{Var}(\hat{R}) = \frac{6-3}{6}\frac{1}{3(46,67)^2}\frac{24800 - 2 * 1,714 * 11400 - (1,714)^2 * 7800}{2} = 0,3305$$

Tabela 6: Valor do parâmetro e suas respectivas estimativa obtidas pelas expressões aproximadas e as estimativas "Bootstrap" com 10 amostras "Bootstrap" (**B**=10)

Parâmetro	Estimativa aproximada	Estimativa "Bootstrap"	Parâmetro	Estimativa aproximada	Estimativa "Bootstrap"
Viés($\hat{R}$)	$\widehat{\text{Viés}}(\hat{R})$	$\widehat{\text{Viés}}_{\mathbf{B}}(\hat{R})$	$Var(\hat{R})$	$\widehat{Var}(\hat{R})$	$\widehat{Var}_{\mathbf{B}}(\hat{R})$
0,0056	0,055	0,014	0,045	0,3305	0,3078

Conforme pode ser observado na Tabela 6 as estimativas "Bootstrap" do viés de $\hat{R}$ e da variância de $\hat{R}$ apresentaram valores distantes dos valores dos parâmetros. Isso deve-se ao fato de que o tamanho da amostra é muito pequeno, n=3 e com isso o número máximo de amostra "Bootstrap" distintas é muito pequeno (**B**=10). Conseqüentemente o processo de reamostragem é ineficiente. Todavia, verifica-se que a obtenção da estimativa da variância e do viés, através das expressões aproximadas é muito mais complexo que através do procedimento "Bootstrap" e os valores encontrados pelos dois métodos são próximos.

3.8 Intervalo de Confiança "Bootstrap"

Como foi visto no capítulo sobre inferência estatística, geralmente não é suficiente avaliar determinadas características da população apenas com estimação pontual. Muitas vezes, é necessário recorrer a um mecanismo de estimação por intervalo que possibilite avaliar o erro que se comete na referida estimação pontual. A avaliação deste erro pode ser feita com a construção de um intervalo de confiança (IC) para o parâmetro de interesse, isto é, um intervalo do tipo:

$$[l_1(x)\ ;\ l_2(x)]$$

tal que, a probabilidade do intervalo conter o verdadeiro valor do parâmetro (θ) seja igual ao valor pré-fixado 1-2α, ou seja,

$$\underbrace{P\{l_1(x) \leq \theta \leq l_2(x)\}}_{\text{Prob. de cobertura}} = \underbrace{1-2\alpha}_{\text{Nível de confiança}} \tag{14}$$

Geralmente, nas aplicações práticas, procura-se construir IC com caudas iguais, ou seja:

$$P\{\theta < l_1(x)\} = \alpha \qquad e \qquad P\{\theta > l_2(x)\} = \alpha \tag{15}$$

Os intervalos de confiança que satisfazem a condição da Expressão (14) são ditos exatos. Mas, infelizmente, nem sempre é possível obter IC exatos, no sentido de que a condição da Expressão (15) sejam exatamente iguais a α. Freqüentemente, na prática, os IC são somente aproximados, isto é:

$$P\{\theta < l_1(x)\} \doteq \alpha \qquad e \qquad P\{\theta > l_2(x)\} \doteq \alpha \tag{16}$$

Nestes casos, além do IC, os limites l_1 e l_2 são ditos também aproximados. Por exemplo, o IC para μ de uma determinada população, com variância conhecida (σ^2) é dado por:

$$[l_1(x)\ ;\ l_2(x)] = \left[\bar{x} - z_{(1-\alpha)}\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} - z_{(\alpha)}\frac{\sigma}{\sqrt{n}}\right]$$

Onde $\bar{x}$ é a média amostral de n observações independentes, z_α e $z_{(1-\alpha)}$ são respectivamente o α-ésimo e o (1-α)-ésimo percentis da distribuição normal padrão.

Este IC é exato somente se as observações $x_1, x_2 \cdots x_n$ são normalmente distríbuidas com média μ e variância σ^2. Mesmo que esta pressuposição não seja satisfeita, ele é aproximado, uma vez que, pelo *teorema central do limite*

$$P\left\{\mu < \bar{X} - z_{(1-\alpha)}\frac{\sigma}{\sqrt{n}}\right\} = P\left\{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} > z_{(1-\alpha)}\right\} = P\left\{Z > z_{(\alpha)}\right\} \doteq \alpha$$

e

$$P\left\{\mu > \bar{X} - z_{(\alpha)}\frac{\sigma}{\sqrt{n}}\right\} = P\left\{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} < z_{(\alpha)}\right\} = P\left\{Z < z_{(\alpha)}\right\} \doteq \alpha$$

onde Z é uma variável aleatória que tem distribuição normal padrão.

Quando se trabalha com intervalos de confiança aproximados, o primeiro questionamento que se faz é sobre a proximidade das probabilidades em questão na Expressão (16), já que quanto mais próximas, mais exatos serão os intervalos. Outro questionamento que se pode fazer, é quão proximos estão os limites aproximados $l_1(x)$ e $l_2(x)$ dos limites exatos, como definidos na Expressão (15). Segundo SILVA (1995), estes dois questionamentos definem importantes propriedades para avaliação e comparação de IC aproximados, que são, respectivamente a `acurácia` e a `corretibilidade` do intervalo.

A noção de `acurácia` está relacionada com a proximidade entre as probabilidades de cobertura dos intervalos $(-\infty\ ,\ \hat{\theta}[\alpha])$ e $(\hat{\theta}[\alpha]\ ,\ \infty)$ e o nível desejado α. Portanto, quanto menores forem os $ERRO_I$ e $ERRO_{II}$, definidos em (17), mais acurado será o IC aproximado. O conceito de `corretibilidade` se refere à proximidade de um limite de confiança aproximado com relação ao limite de confiança exato.

Seja o intervalo de confiança aproximados $\left[\hat{\theta}[\alpha]\ ;\ \hat{\theta}[1-\alpha]\right]$, onde $\hat{\theta}[\alpha]$ e $\hat{\theta}[1-\alpha]$ representam, respectivamente, os limites inferior e superior de intervalo, tais que:

$$P\{\theta < \hat{\theta}[\alpha]\} \doteq \alpha \qquad e \qquad P\{\theta > \hat{\theta}[1-\alpha]\} \doteq \alpha$$

Considerem-se agora os seguintes tipos de erros

$$\begin{aligned} ERRO_I(\hat{\theta}[\alpha]) &= P\{\theta < \hat{\theta}[\alpha]\} = \alpha \\ ERRO_I(\hat{\theta}[1-\alpha]) &= P\{\theta < \hat{\theta}[1-\alpha]\} = \alpha \end{aligned} \tag{17}$$

e

$$ERRO_{II}(\hat{\theta}[\alpha]) = \hat{\theta}[\alpha] - \hat{\theta}_{exato}[\alpha]$$

ACURACIDADE DE ORDEM 1:

Um IC aproximado para o parâmetro θ, com probabilidade de cobertura aproximadamente iqual a 1-2α é denominado acurado de ordem 1, se (para duas constantes c_1 e c_2):

$$ERRO_I(\hat{\theta}[\alpha]) = \frac{c_1}{\sqrt{n}} \qquad e \qquad ERRO_I(\hat{\theta}[1-\alpha]) = \frac{c_2}{\sqrt{n}}$$

ACURACIDADE DE ORDEM 2:

Um IC aproximado para o parâmetro θ, com probabilidade de cobertura aproximadamente igual 1-2α é denominado acurado de ordem 2, se (para duas constantes c_1 e c_2):

$$ERRO_I(\hat{\theta}[\alpha]) = \frac{c_1}{n} \qquad e \qquad ERRO_I(\hat{\theta}[1-\alpha]) = \frac{c_2}{n}$$

CORRETIBILIDADE DE ORDEM 1:

Sendo σ uma estimativa de desvio-padrão de θ e $\hat{\theta}_{exato}[\alpha]$ um limite de confiança exato para θ, de nível α, que satisfaz $P\{\theta \leq \hat{\theta}_{exato}[\alpha]\} = \alpha$. Um limite de confiança aproximado $\hat{\theta}[\alpha]$ é denominado correto de ordem 1, se

$$ERRO_{II}(\hat{\theta}[\alpha]) = O(n^{-1}) \qquad ou \qquad ERRO_{II}(\hat{\theta}[\alpha]) = O(n^{-1/2})\hat{\sigma}$$

desde que $\hat{\sigma}$ seja usualmente de ordem $n^{-1/2}$.

CORRETIBILIDADE DE ORDEM 2:

Um limite de confiança aproximado $\hat{\theta}[\alpha]$ é denominado correto de ordem 2, se

$$ERRO_{II}(\hat{\theta}[\alpha]) = O(n^{-3/2}) \qquad \text{ou} \qquad ERRO_{II}(\hat{\theta}[\alpha]) = O(n^{-1})\hat{\sigma}$$

Os erros dos intervalos acurados de ordem 1 tendem a zero à medida que o tamanho da amostra tende ao infinito, a uma taxa (velocidade) de $1/\sqrt{n}$, enquanto que, os erros dos intervalos acurados de ordem 2 tendem a zero à medida que n tende ao infinito a uma taxa de $1/n$. A comparação destas duas taxas fornece, não somente uma 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 `corretibilidade` de uma ordem implica na `acurácia` da mesma ordem.

A seguir, será visto como o método "Bootstrap" pode ser usado para a construção de um IC de um parâmetro de interesse θ. Os intervalos de confiança Bootstrap poderão apresentar melhores aproximações, no que tange às propriedades descritas anteriormente, que outros intervalos baseados em métodos clássicos, além do que, podem ser uma alternativa implementável, quando existe dificuldade em construir os intervalos através de métodos clássicos. Serão apresentados os seguintes intervalos de confiança Bootstrap:

Bootstrap-padrão - é construído sob a suposição de normalidade da estatística $\frac{(\hat{\theta}-\theta)}{\sigma_{Boot}(\hat{\theta})}$;

t-Student Bootstrap - ao contrário do anterior, é construído como base na aproximação pela distribuição t-Student com $n-1$ gl, para a referida estatística;

t-Bootstrap - é construido sob a aproximação Bootstrap para a distribuição da estatística $\frac{(\hat{\theta}-\theta)}{\sigma_{Boot}(\hat{\theta})}$, de forma análoga ao método clássico que usa a aproximação t-Student para a distribuição da referida estatística.

3.8.1 Intervalo Bootstrap-Padrão

Seja $\sigma_{Boot}(\hat{\theta})$ a estimativa "Bootstrap" de $\sigma_F=\sqrt{Var_F(\hat{\theta})}$. O intervalo de confiança Bootstrap-Padrão ($BootPad$) para θ, com probabilidade de cobertura 1-2α é dado por:

$$\left[\hat{\theta}_{(BootPad)}[\alpha]\ ;\ \hat{\theta}_{(BootPad)}[1-\alpha]\right] = \left[\hat{\theta} - z_{(1-\alpha)}\sigma_{Boot}(\hat{\theta})\ ;\ \hat{\theta} - z_{(\alpha)}\sigma_{Boot}(\hat{\theta})\right] \qquad (18)$$

onde $z_{(\alpha)}$ e $z_{(1-\alpha)}$ são respectivamente o α-ésimo e o (1-α)-ésimo percentis da normal padrão, onde $z_\alpha=\phi^{-1}(\alpha)$.

A construção deste intervalo baseia-se na aproximação assintótica

$$T = \frac{\hat{\theta} - \theta}{\sigma_{Boot}(\hat{\theta})} \quad \sim \quad N(0\ ,\ 1)$$

Caso a distribuição de T seja perfeitamente normal padrão, então o intervalo definido na Expressão (18) é exato, caso contrário o intervalo é somente aproximado.

Considerando os dados do Exemplo 1, seja o interesse em obter o IC Bootstrap-Padrão para as média dos tratamentos A e B, com base nas médias amostrais $\bar{y}$ e $\bar{z}$, com coeficiente de confiança 1-2α=0,95 (α=0,025).

Para o tratamento A, temos $\hat{\theta}_1=\bar{y}=478,54$ e $\sigma_{Boot}(\hat{\theta}_1)=16,95$. Portanto, os limites do IC Bootstrap-Padrão serão dados por:

$$\hat{\theta}_{BootPad}[0,025] = \bar{y} - z_{(1-\alpha)}\sigma_{Boot}(\bar{y}) = 478,54 - 1,96(16,95) = 445,318$$

e

$$\hat{\theta}_{BootPad}[0,975] = \bar{y} - z_{(\alpha)}\sigma_{Boot}(\bar{y}) = 478,54 + 1,96(16,95) = 511,762$$

$\therefore$ o IC Bootstap-Padrão para $\bar{Y}$: [445,318 ; 511,762]

Para o tratamento B, temos $\hat{\theta}_2=\bar{z}=440,87$ e $\sigma_{Boot}(\hat{\theta}_2)=17,37$ e os limites do IC Bootstrap-Padrão serão dados por:

$$\hat{\theta}_{BootPad}[0,025] = 440,87 - 1,96(17,37) = 406,8248$$

e

$$\hat{\theta}_{BootPad}[0,975] = 440,87 + 1,96(17,37) = 474,9152$$

$\therefore$ o IC Bootstrap-Padrão para $\bar{Z}$: [406,8248 ; 474,9152]

Como os dados do Exemplo 1 foram obtidos por simulação e, portanto, conhecem-se as características das populações, é possível se obter os intervalos de confianças exatos e assim fazer uma comparação com o IC Bootstrap-Padrão.

O intervalo de confiança exato de θ é dado por

$$\left[\hat{\theta} - z_{(1-\alpha)}\left(\sqrt{\frac{\sigma^2}{n}}\right) \; ; \; \hat{\theta} - z_{\alpha}\left(\sqrt{\frac{\sigma^2}{n}}\right)\right].$$

Logo, para o tratamento A temos:

$$\hat{\theta}_1[0,025] = 478,54 - 1,96\left(\sqrt{\frac{2500}{7}}\right) = 441,4995$$

$$\hat{\theta}_1[0,975] = 478,54 + 1,96\left(\sqrt{\frac{2500}{7}}\right) = 515,5805$$

Para o tratamento B temos:

$$\hat{\theta}_2[0,025] = 440,85 - 1,96\left(\sqrt{\frac{2500}{8}}\right) = 406,2018$$

$$\hat{\theta}_2[0,975] = 440,85 + 1,96\left(\sqrt{\frac{2500}{8}}\right) = 475,4982$$

Conforme pode-se observar através da Tabela 7 os valores dos limites obtidos através do intervalo de confiança Boot-padrão estão muito próximos aos valores dos intervalos de confiança exatos.

Tabela 7: Intervalos de confiança para as média dos tratamentos A e B, do Exemplo 1, considerando um coeficiente de confiança de 95%

IC	Estatística	$\hat{\theta}$ (0,025)	$\hat{\theta}$ (0,975)	Parâmetro
Exato-Padrão	$\bar{Y}$	441,50	515,58	500
	$\bar{Z}$	406,20	475,50	450
Bootstrap-Padrão	$\bar{Y}$	445,32	511,76	500
	$\bar{Z}$	406,82	474,92	450

Considerando agora os dados do Exemplo 3, seja o interesse em obter o IC Bootstrap-Padrão para o coeficiente de correlação linear das variáveis aleatória Y e Z, como base nos dados amostrais, onde foi obtido $\widehat{corr}(y,z)$=0,776. Uma medida largamente aceita de precisão de um estimador $\widehat{corr}(y,z)$ é a largura do intervalo que abrange 68% da área central de sua distribuição de frequëncia (DIACONIS & EFRON, 1993). Essa largura é obtida quando o limite inferior é $\hat{\theta}(\alpha$=0,16) e o limite superior é $\hat{\theta}(\alpha$=0,84). Para obter estes limites, usando a tabela da distribuição normal é necessário conhecer o erro padrão da estatística $\widehat{corr}(y,z)$, e como foi visto nos capítulos anteriores não existe uma expressão simples para obtenção do erro padrão da estatística $\widehat{corr}$ e por consequência a estimativa "Bootstrap" exata. Por isso usar-se-á a estimativa "Bootstrap" obtida pelo algoritmo de Monte Carlo com B=3200 para construção do intervalo de confiança, conforme descrito na Tabela 8.

Tabela 8: Intervalos de confiança para o coeficiente de correlação para os dados do Exemplo 3, considerando-se um coeficiente de confiança de 68%

IC	Estatística	$\hat{\theta}$ (0,16)	$\hat{\theta}$ (0,84)	Parâmetro
Exato-Padrão	$Corr(Y,Z)^a$	0,606	0,876	0,761
Bootstrap-Padrão	$Corr(Y,Z)$	0,644	0,908	0,761

[a] Fonte: DIACONIS & EFRON (1993)

Como os dados populacionais são conhecidos (DIACONIS & EFRON, 1993), pode-se fazer uma comparação dos valores amostrais e valores "Bootstrap"

com os dados originais. Conforme pode-se se notar na Tabela 8, apesar dos limites não coincidirem, estão muito próximos.

3.8.2 Intervalo de Confiança t-Student Bootstrap

Anteriormente discutiu-se o IC Bootstrap-Padrão em que se tinha a suposição de qua a distribuição estatística T era aproximadamente normal padrão. No entanto, em pequenas amostras, pelo menos para $\hat{\theta}=\bar{x}$, uma melhor aproximação para a distribuição de T é dada pela distribuição de t-Student com (n-1) graus de liberdade. Isto é:

$$T = \frac{\hat{\theta} - \theta}{\sigma_{Boot}(\hat{\theta})} \quad \dot{\sim} \quad t_{n-1} \tag{19}$$

se as observações forem normalmente distribuídas então esta aproximação é exata.

Seja $t_{n-1}(\alpha)$ e $t_{n-1}(1\text{-}\alpha)$, respectivamente o α-ésimo e o (1-α)-ésimo percentis da distribuição t com n-1 gl, que satisfaz

$$P\{t_{n-1} \leq t_{n-1}(\alpha)\} = \alpha \quad e \quad P\{t_{n-1} \geq t_{n-1}(1-\alpha)\} = \alpha$$

Assim, sob a suposição dada em (19)

$$P\{t_{n-1}(\alpha) \leq T \leq t_{n-1}(1-\alpha)\} \doteq 1 - 2\alpha$$

Logo,

$$P\{\hat{\theta} - t_{n-1}(1-\alpha)\hat{\sigma}_{Boot}(\hat{\theta}) \leq \theta \leq \hat{\theta} - t_{n-1}(\alpha)\hat{\sigma}_{Boot}(\hat{\theta})\} \doteq 1 - 2\alpha$$

Portanto, o intervalo de confiança t-Student Bootstrap (IC t-SBoot) com a probabilidade de cobertura aproximadamente 1-2α para θ, será dado por:

$$\left[\hat{\theta} - t_{n-1}(1-\alpha)\hat{\sigma}_{Boot}(\hat{\theta}) \; ; \; \hat{\theta} - t_{n-1}(\alpha)\hat{\sigma}_{Boot}(\hat{\theta})\right]$$

Considerando os dados do Exemplo 1, seja o interesse em obter os Intervalos de Confiança t-SBoot, com a probabilidade de cobertura aproximadamente 1-2α (α=0,025), para as médias dos tratamentos A e B, com base nas médias amostrais $\hat{\theta}_1=\bar{y}=478,54$ e $\hat{\theta}_2=\bar{z}=478,54$.

Para o tratamento A, tem-se $\hat{\theta}_1=478,54$ e $\sigma_{Boot}(\hat{\theta}_1)=16,95$. Portanto, os limites do IC t-SBoot serão dados por:

$$\begin{aligned}\hat{\theta}_{t-SBOOT}[0,025] &= \bar{y} - t_{n-1}(1-\alpha)\sigma_{Boot}(\bar{y}) \\ &= 478,54 - 2,447(16,95) = 437,063\end{aligned}$$

$$\hat{\theta}_{t-SBOOT}[0,975] = 478,54 + 2,447(16,95) = 520,017$$

Para o tratamento B, tem-se $\hat{\theta}_2=\bar{z}_{Boot}=440,87$ e $\sigma_{Boot}(\hat{\theta}_2)=17,37$ e os limites do IC t-SBoot serão dados por:

$$\hat{\theta}_{t-SBOOT}[0,025] = 440,87 - 2,365(17,37) = 399,790$$

$$\hat{\theta}_{t-SBOOT}[0,975] = 440,87 + 2,365(17,37) = 481,950$$

Considerando-se que a estatística T tem distribuição t-Student com n-1 gl, o intervalo de confiança exato para θ, com o coeficiente de confiança 1-2α, é obtido por $\left[\hat{\theta} + t_{n-1}(\alpha)\left(\sqrt{\frac{s^2}{n}}\right) \; ; \; \hat{\theta} - t_{n-1}(\alpha)\left(\sqrt{\frac{s^2}{n}}\right)\right]$. Como os dados dos tratamentos A e B, do Exemplo 1, advêm de distribuições normais e portanto, os IC t-Exatos apresentam valores exatos, pode-se verificar a acuracidade do IC t-SBoot.

Tabela 9: Intervalos de confiança para as média dos tratamentos A e B, do Exemplo 1, considerando um coeficiente de confiança de 95%

IC	Estatística	$\hat{\theta}$ (0,025)	$\hat{\theta}$ (0,975)	Parâmetro
t-Exato	$\bar{Y}$	433,7354	523,344	500
	$\bar{Z}$	396,952	484,788	450
t-SBoot	$\bar{Y}$	437,063	520,017	500
	$\bar{Z}$	399,790	481,950	450

3.8.3 Intervalo de Confiança t-Bootstrap

Os intervalos de cofiança Bootstrap-Padrão e t-SBoot tinham as suposições da distribuição da estatística T baseadas, respectivamente na distribuição normal padrão e na distribuição t-Student. Estas suposições nem sempre são válidas e segundo EFRON & TIBSHIRANI (1993) a t-Student não ajusta o intervalo de acordo com a assimetria da distribuição dos dados e outros erros podem acontecer quando $\hat{\theta}$ não é a média amostral.

O intervalo de confiança t-Bootstrap para o parâmetro $\hat{\theta}=s(X)$ usa a aproximação "Bootstrap" para a distribuição da estatística T e de forma análoga aos casos anteriores obtêm-se os percentis da distribuição de T para formar o intervalo de confiança. Isto é:

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{Boot}(\hat{\theta})} = \frac{s(X) - t(F)}{\sqrt{Var_F[s(x)]}}$$

A aproximação "Bootstrap" para a distribuição de T é obtida por:

$$T^* = \frac{\hat{\theta}^* - t(\hat{F})}{\hat{\sigma}_{Boot}(\hat{\theta})} = \frac{s(X^*) - t(\hat{F})}{\sqrt{Var_{\hat{F}}[s(x^*)]}}$$

De acordo com EFRON E TIBSHIRANI (1993), a obtenção do intervalo de confiança é feita da seguinte forma: sejam $t^*(\alpha)$ e $t^*(1\text{-}\alpha)$, respectivamente o α-ésimo e o (1-α)-ésimo percentis da distribuição "Bootstrap" de T^* que satisfaz

$$P\{T^* \leq t^*(\alpha)\} = \alpha \quad e \quad P\{T^* \geq t^*(1-\alpha)\} = \alpha$$

Como as probabilidades são calculadas sob a distribuição "Bootstrap" de T^*, então

$$P\{t^*(\alpha) \leq T^* \leq t^*(1-\alpha)\} = 1 - 2\alpha$$

E, com base na distribuiçào de T^* tem-se uma aproximação para a distribuição de T, ou seja,

$$P_{\hat{F}}\{t^*(\alpha) \leq T^* \leq t^*(1-\alpha)\} \doteq 1 - 2\alpha$$
$$P_F\{t^*(\alpha) \leq T \leq t^*(1-\alpha)\} \doteq 1 - 2\alpha$$

Logo, o intervalo de confiança t-Bootstrap, com a probabilidade de cobertura aproximadamente 1-2α para o parâmetro θ, será dado por:

$$\left[\hat{\theta} - t^*(1-\alpha)\hat{\sigma}_{Boot}(\hat{\theta}) \;\; ; \;\; \hat{\theta} - t^*(\alpha)\hat{\sigma}_{Boot}(\hat{\theta})\right]$$

O procedimento t-Bootstrap estima a distribuição de T diretamente através dos dados amostrais e a partir daí constróem-se tabelas com os níveis de significância desejados. Dessa forma, os valores obtidos são usados para construir os intervalos de confiança da mesma forma que se usam os valores das tabelas das distribuições normal e t de Studnet.

Entretanto, como já foi mencionado, pode ser difícil calcular analiticamente a distribuição "Bootstrap" exata da estatística T^*. Neste caso a construção da tabela "Bootstrap" é feita através da aproximação Monte Carlo a partir da geração de **B** amostras "Bootstrap", coforme mostra o esquema da Figura 11.

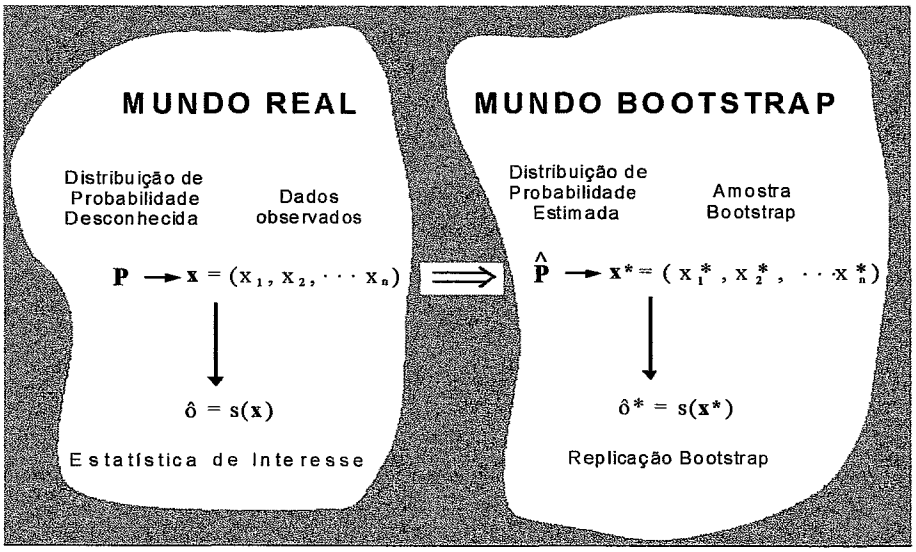

Figura 11: Diagrama esquemático da aplicação "Bootstrap" para problemas com estrutura de dados mais gerais. O passo crucial no processo "Bootstrap" é " $\Rightarrow$ ", a maneira pelo qual se constrói, a partir dos dados observados **x**, um estimador $\hat{P}$ do mecanismo probabilistico P. FONTE: EFRON & TIBSHIRANI,(1993)

Figura 12: Algoritmo de Monte Carlo para obter o intervalo de confiança t-Bootstrap de uma estatística de interesse

Algoritmo 4

1) Através de um gerador de números aleatórios, selecionam-se, independentemente **B** amostras "Bootstrap" $\mathbf{x}^*_{(1)}, \mathbf{x}^*_{(2)}, \cdots \mathbf{x}^*_{(B)}$. Isto é, **B** amostras aleatórias simples de tamanho n , extraídas com reposição da amostra atual $\mathbf{x} = \{x_1, x_2, \cdots x_n\}$;

2) Para cada amostra "Bootstrap" $(\mathbf{x}^*_{(b)})$ obtém-se o valor de $T^*_{(b)}$, que é dado por:

$$T^*_{(b)} = \frac{\hat{\theta}^*_{(b)} - \hat{\theta}}{\hat{\sigma}_{\mathbf{B}}(\hat{\theta}^*_{(b)})}$$

onde $\hat{\theta}^*_{(b)} = s(X^*_{(b)})$ é o valor de $\hat{\theta}$ para cada amostra "Bootstrap" $X^*_{(b)}$ e $\hat{\sigma}_{\mathbf{B}}(\hat{\theta}^*_{(b)})$ é a estimativa "Bootstrap" do erro-padrão de $\hat{\theta}^*$ para cada amostra "Bootstrap" $X^*_{(b)}$.

3) O α-ésimo percentil de $T^*_{(b)}$ é estimado pelo valor $\hat{t}(\alpha)$, tal que

$$\frac{\#\{T^*_{(b)} \leq \hat{t}^*(\alpha)\}}{\mathbf{B}} = \alpha$$

Para **B** amostras "Bootstrap", a estimativa do ponto $\hat{t}^*(\alpha)$ corresponde ao $[\mathbf{B} * (\alpha)]$ maior valor dos $T^*_{(b)}$ e a estimativa do ponto $\hat{t}^*(1-\alpha)$ corresponde ao $[\mathbf{B} * (1-\alpha)]$ maior valor dos $T^*_{(b)}$. Por exemplo, se $\mathbf{B} = 1000$, o percentil 5% da distribuição "Bootstrap" de T^* é estimado pelo $50^{\underline{o}}$ $(1000 * 0,05)$ maior valor entre os $T^*_{(b)}$ que numa forma ordenada corresponde ao valor $T^*_{(50)}$. o percentil 95% é estimado pelo $950^{\underline{o}}$ $(1000 * 0,95)$ maior valor entre os $T^*_{(b)}$ que numa forma ordenada corresponde ao valor $T^*_{(950)}$.

Se o valor obtido por $[\mathbf{B} * (\alpha)]$ não for um número inteiro, segundo EFRON & TIBSHIRANI(1993), deve-se adotar o seguinte procedimento: sendo $\alpha \leq 0,5$ toma-se $k = [(\mathbf{B}+1) * (\alpha)]$, o maior inteiro menor ou iqual a $[(\mathbf{B}+1) * (\alpha)]$ e os quantis empíricos (α) e $(1-\alpha)$ serão respectivamente o k-ésimo e $(\mathbf{B}+1-k)$-ésimo maior valor dos $T^*_{(b)}$, que numa forma ordenada correspondem, respectivamente, aos valores $T^*_{(k)}$ e $T^*_{(\mathbf{B}+1-k)}$.

Usando a notação das Figuras 11 e 12, apresentar-se-á o intervalo de confiança t-Bootstrap para o parâmetro μ=$t(F)$ a partir do estimador $\hat{F}$ de F, que é dado como segue:

$$t(\hat{F}) = \int x d_{\hat{F}}(x) = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

e

$$T^* = \frac{\bar{X}^* - \bar{x}}{\hat{\sigma}_{Boot}(\bar{x})} = \frac{\bar{X}^* - \bar{x}}{\sqrt{\frac{\sum^{n}(x_i - \bar{x})^2}{n^2}}}$$

Como a obtenção da distribuição de T^* pode ser complexa, usa-se a aproximação obtida a partir dos **B** de $T^*_{(b)}$, onde,

$$T^*_{(b)} = \frac{(\bar{x}^*_{(b)} - \bar{x}}{\hat{\sigma}_{\mathbf{B}}(\bar{x}^*_{(b)})} = \frac{(\bar{x}^*_{(b)} - \bar{x}}{\sqrt{\frac{\sum^{n}(x^*_{(b_i)} - \bar{x}^*_{(b)})^2}{n^2}}}$$

Dessa forma, o intervalo de confiança t-Bootstrap para θ é dado por:

$$\left[\hat{\theta} - \hat{t}(1-\alpha)\hat{\sigma}_{\mathbf{B}}(\hat{\theta}) \; ; \; \hat{\theta} - \hat{t}(\alpha)\hat{\sigma}_{\mathbf{B}}(\hat{\theta})\right] \tag{20}$$

Na Tabela 10 encontram-se os valores dos percentis de $T^*_{(b)}$ para os dados dos tratamentos A e B do Exemplo 1, juntamente com os percentis da distribuição Normal Padrão e t-Student com 6 e 7 graus de liberdades. Pode-se verificar que os percentis da distribuição "Bootstrap" dos $T^*_{(b)}$ são ajustados de acordo com a assimetria dos dados, pois para o tratamento B que tem pouca assimetria (coeficiente de assimetria -0,43) os percentis da distribuição t-Bootstrap são muito próximos aos percentis da distribuição t-Student. Já para os dados do tratamento A que têm maior assimetria (coeficiente de assimetria -0,61), os percentis diferem devido ao ajuste da distribuição t-Bootstrap à assimetria dos dados.

Tabela 10: Percentis da distribuição Normal Padrão, distribuição t-Student com 6 e 7 graus de liberdade e da distribuição "Bootstrap" de $T^*_{(b)}$, para os tratamentos A e B do Exemplo 1

Percentis	2,5%	5%	10%	50%	90%	95%	97,5%
Normal Padrão	-1,96	-1,65	-1,28	0,00	1,28	1,65	1,96
t-Student (6gl)	-2,45	-1,94	-1,44	0,00	1,44	1,94	2,45
t-Student (7gl)	-2,36	-1,89	-1,41	0,00	1,41	1,89	2,36
t-Bootstrap (A)	-2,39	-2,03	-1,45	0,17	3,63	5,16	6,24
t-Bootstrap (B)	-2,28	-2,07	-1,43	0,02	1,56	2,02	2,46

Na Tabela 11 encontram-se os intervalos de confiança para as medias populacionais dos tratamentos A e B do Exemplo 1, baseados na distribuição Normal Padrão, t-Student e t-Bootstrap. Pode-se verificar as proximidades dos valores obtidos pelo método "Bootstrap" aos valores da distribuição t-Student, principalmente para o tratamento de menor assimetria.

Tabela 11: Intervalos de confiança para as média dos tratamentos A e B, do Exemplo 1, com a probabilidade de cobertura de 95%, com base nas distribuições Normal Padrão, t-Studnet e t-Bootstrap

IC	Estatística	$\hat{\theta}$ (0,025)	$\hat{\theta}$ (0,975)	Parâmetro
Normal Padrão	$\bar{Y}$	442,651	513,512	500
	$\bar{Z}$	404,475	477,270	450
t-Student	$\bar{Y}$	433,679	523,398	500
	$\bar{Z}$	397,047	484,698	450
t-Bootstrap	$\bar{Y}$	372,295	519,294	500
	$\bar{Z}$	398,489	480,146	450

O intervalo de confiança t-Bootstrap apresentado em (20) é um intervalo acurado, obtido diretamente através dos dados amostrais, sem a suposição de normalidade dos dados. Segundo EFRON & TIBSHIRANI (1993), esse intervalo tem acuracidade de ordem 2 mas não apresenta a propriedade de *Invariância a transformações monótonas* nem a propriedade de *Preservação da amplitude.* Dessa forma, para estrutura de dados mais gerais, os autores apresentam outros intervalos de confiança baseados nos percentis da distribuição "Bootstrap", como o intervalo chamado *BCa* (*"Bias-Corrected and aceletated"*) que é um intervalo de confiança corrigido para tendência e aceleração e o intervalo chamado *ABC* (*"Approximate Bootstrap Confidence"*) que é uma aproximação do *BCa*, usando o processo de simulação Monte Carlo. Na Tabela 12 encontram-se todos os intervalos de confiança envolvendo o procedimento "Bootstrap", para as médias dos tratamentos A e B do Exemplo 1.

Tabela 12: Intervalos de confiança para as média dos tratamentos A e B, do Exemplo 1, com a probabilidade de cobertura de 95%.

Estatística	Parâmetro	I. Confiança	$\hat{\theta}$ (0,025)	$\hat{\theta}$ (0,975)
		Boot-Padrão	445,322	511,755
		t-SBoot	437,063	520,017
$\bar{Y}$	500	t-Bootstrap	372,295	519,294
		ABC	440,455	507,803
		BCa	439,831	507,173
		Boot-Padrão	406,820	474,914
		t-SBoot	399,790	481,950
$\bar{Z}$	450	t-Bootstrap	398,489	480,146
		ABC	403,516	472,042
		BCa	403,152	472,980

3.9 Teste de Hipótese "Bootstrap"

3.9.1 Introdução

A inferência estatística, além da estimação por ponto e estimação por intervalo, compreende também o problema de teste de hipóteses. Um teste de hipótese é uma regra de decisão para aceitar ou rejeitar uma hipótese estatística, com base nos elementos amostrais. Uma hipótese estatística é uma suposição quanto ao valor de um parâmetro populacional ou uma afirmação quanto à natureza da população.

Hipótese nula - (H_0), é uma hipótese estatística formulada pelo pesquisador e que será testada através do teste de hipótese, com o propósito de ser rejeitada. A hipótese alternativa - (H_a), é a suposição a qual o pesquisador deseja provar, que será aceita quando rejeita-se (H_0) ou será rejeitada quando aceita-se (H_0).

Um teste de hipótese começa com um teste estatístico $\hat{\theta}$, tal como a diferença de médias ($\hat{\theta}$=$\bar{y} - \bar{z}$=37,67) dada na Tabela 1. Por conveniência, para as deduções a seguir, assumir-se-á que hipótese alternativa trata-se de uma hipótese unicaudal à direita, ou seja, espera-se rejeitar (H_0) com valores maiores que o valor de (H_0).

Tendo observado $\hat{\theta}$, o nível crítico de significância do teste (NC) é definido como sendo a probabilidade de se observarem valores maiorer do que, ou iguais, ao valor estabelecido em (H_0), quando a hipótese nula é verdadeira, isto é,

$$NC = Prob_{H_0}\left(\hat{\theta}^* \geq \hat{\theta}\right) \tag{21}$$

onde $\hat{\theta}^*$ é a variável aleatória gerada sob (H_0) e $\hat{\theta}$ é o valor observado. O menor valor do NC é a mais forte evidência contra (H_0), em outras palavras, é o máximo

valor do teste para rejeitar (H_0).

O teste de hipótese de (H_0), consiste no cálculo do nível crítico de significância do teste (NC), verificando se ele é menor que o limiar convecionalmente estabelecido. Formalmente, escolhe-se uma probabilidade pequena α, como $0,05$ ou $0,01$, e rejeita-se (H_0) se o NC é menor do que α.

Para os dados do Exemplo 1, onde os dados observados do tratamento A, $\mathbf{y}=(y_1, y_2 \cdots y_n)$ e os dados observados do tratamento B, $\mathbf{z}=(z_1, z_2 \cdots z_m)$ são amostras aleatórias independentes extraídas, respectivamente, das populações F e G, isto é,

$$\begin{cases} F \rightarrow & \mathbf{y} = (y_1, y_2 \cdots y_n) \\ G \rightarrow & \mathbf{z} = (z_1, z_2 \cdots z_m) \end{cases}$$

F e G são independentes e um teste de hipótese tradicional pode ser feito a partir da suposição de que F e G são normalmente distribuídas com possíveis diferenças de médias, ou seja,

$$\begin{cases} F \sim & N(\mu_Y \ , \ \sigma^2) \\ G \sim & N(\mu_Z \ , \ \sigma^2) \end{cases}$$

A hipótese nula é $\mu_Y=\mu_Z$ que sob (H_0) $\hat{\theta}=(\bar{y}-\bar{z})$ tem distribuição Normal com os seguintes parâmetros:

$$\hat{\theta}_{H_0} \quad \sim \quad N\left(0 \ , \ \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right) \tag{22}$$

Tendo observado $\hat{\theta}$, o NC é a probabilidade de que a variável aleatória $\hat{\theta}^*$, que tem a distribuição descrita em (22), exceda $\hat{\theta}$.

Logo,

$$NC = Prob_{H_0}\left\{Z > \left(\frac{\hat{\theta}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)\right\} = 1 - \phi\left(\frac{\hat{\theta}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)$$

onde ϕ é a função de distribuição acumulada da variável Normal padrão Z e para o caso em que σ é desconhecido e o tamanho da amostra não é suficientemente grande,

uma estimativa para σ é dado por $\hat{\sigma} = \sqrt{\frac{\sum^{n}(y_i-\bar{y})^2+\sum^{m}(z_i-\bar{z})^2}{n+m-2}}$. Logo,

$$NC = Prob_{H_0}\left\{t_{(n+m-2)} > \left(\frac{\hat{\theta}}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)\right\}$$

onde $t_{(n+m-2)}$ indica uma variável aleatória com distribuição t-Student com $(n + m - 2)$ graus de liberdade.

Para os dados do Exemplo 1, um teste de hipótese para verificar se há diferença significativa entre os tratamentos A e B, com a suposição da normalidade dos dados, tem-se que:

$$NC = Prob_{H_0}\left\{t_{(13)} > \left(\frac{37,67}{50,67\sqrt{\frac{1}{7}+\frac{1}{8}}}\right)\right\} = 0,0872 \qquad (23)$$

Com base no valor do NC=0,0872, não se rejeita H_0, concluindo-se portanto, que os tratamentos A e B são iguais estatísticamente com um nível descritivo do teste de 0,0872 (*"P-value"*=0,0872).

Na prática, nem sempre é simples a obtenção do NC, pois, há casos em que não há condições seguras para a suposição de normalidade dos dados. Nesses casos, podem-se obter testes de hipótese aproximados através do método "Bootstrap". A seguir, será descrito um teste de hipótese "Bootstrap" para o caso de duas amostras.

3.9.2 Teste de Hipótese "Bootstrap" Para o Problema de Duas Amostras

Sejam as amostras **Y** e **Z** de duas possíveis distribuições de probabilidade diferentes e deseja-se testar a hipótese nula (H_0) : $F = G$. Um teste de hipótese "Bootstrap" é baseado em um teste estatístico θ=$t(x)$, que não é necessariamente uma estimativa de um parâmetro populacional, conforme foi visto anteriormente. Logo,

$$NC = Prob_{H_0}\{t(x^*) \geq t(x)\}$$

Como em (21), a quantidade $t(x)$ é fixada com base nos valores observados e a variável aleatória X^* tem uma distribuição especificada pela hipótese nula, denotada aqui por F_0, mas, se não foi feita a suposição de normalidade, quem é F_0 ?

Este problema é resolvido facilmente quando se usa o teste de hipótese "Bootstrap", o qual aplica o princípio que consiste em estimar F_0 por $\hat{F}_0$. Denotando a amostra combinada por **x**, seja a distribuição empírica $\hat{F}_0$ que coloca a probabilidade $\frac{1}{(n+m)}$ em cada elemento da amostra **x**. Então, sob H_0, $\hat{F}_0$ fornece uma estimativa não paramétrica da população que deu origem a **y** e **z**. O Algoritmo 5, mostrado na Figura 13 mostra como é obtido o Nível crítico de significância "Bootstrap" para $t(x)$.

Para o caso da diferença de média do Exemplo 1, onde $t(x) = 37,67$, foram geradas 1000 amostras "Bootstrap" e obteveram-se 75 valores em que $t(x^*) > 37,67$.

Logo,

$$\widehat{NC}_{Boot} = \frac{75}{1000} = 0,075 \qquad (24)$$

Figura 13: Algoritmo para obtenção do nível crítico de significância "Bootstrap" para testar $H_0 : \ F = G$

Algoritmo 5

1) Através de um gerador de números aleatórios, retiram-se **B** amostras "Bootstrap" de tamanho $(n+m)$ com reposição da amostra combinada **x**. As primeiras n observações são representadas por $y^*_{(b)}$ e as m observações restantes representam $z^*_{(b)}$. O processo é repetido para todas $b = 1, 2, \cdots \mathbf{B}$ amostras "Bootstrap";

2) Para cada amostra "Bootstrap" $(\mathbf{x}^*_{(b)})$ obtém-se o valor de $t(\cdot)$, que é dado por:

$$t(x^*_{(b)}) = \bar{y}^*_{(b)} - z^*_{(b)}$$

3) O Nível crítico de significância "Bootstrap" é aproximado por:

$$\widehat{NC}_{Boot} = \frac{\#\{t(x^*_{(b)}) \geq t_{obs}\}}{\mathbf{B}}$$

Onde, $t_{obs} = t(x)$ é o valor da estatística observada.

Com base no valor obtido em (24) não se rejeita a hipótese nula $H_0 : \ F{=}G$ e conclui-se, portanto, que não há diferença entre os tratamentos A e B, com um nível descritivo do teste de $0,075$ ($\text{"}P\text{-}value\text{"}_{Boot} = 0,075$). Pode-se verificar que o valor obtido para $\widehat{NC}_{Boot}{=}0,075$ está muito próximo do valor obtido quando foi aplicado o teste de hipótese usando a distribuição t-Student que forneceu $NC{=}0,0872$.

Teste de hipóteses mais exatos podem ser obtidos através do uso de uma estatística estudentizada, usando em vez de $t(x){=}(\bar{y} - \bar{z})$ a estatística estudentizada, descrita na Secção (**3.1**), que é dada como se segue:

$$t(x) = \frac{\bar{y} - \bar{z}}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{25}$$

onde,

$$\hat{\sigma} = \sqrt{\frac{\sum^{n}(y_i - \bar{y})^2 + \sum^{m}(z_i - \bar{z})^2}{(n+m-2)}}$$

No lugar de $t(x^*_{(b)})$, usa-se o valor estudentizado obtido de forma análoga à Expressão (25). Dessa forma, repetindo-se o Algoritmo 5, com as modificações descritas anteriormente, obtém-se um nível crítico de significância "Bootstrap" com base nas estatísticas estudentizadas, que é mais exato que o anterior.

Para os dados do Exemplo 1, usando-se as estatísticas estudentizada, onde $t(x)$=1,43, foram geradas 1000 amostras "Bootstrap" e obteveram-se 87 valores em que $t(x^*) > 1,43$.

Logo,

$$\widehat{NC}_{Boot} = \frac{87}{1000} = 0,087 \tag{26}$$

Embora, o nível crítico de significância "Bootstrap" obtido através de estatísticas estudentizada seja mais exato, pode-se verificar que os valores obtidos através das Expressões (24) e (26) são muito próximos, e por conseguinte são próximos ao valor obtido na Expressão (23) que é o nível descritivo do teste t-Student (*"P-value"*).

Como os dados dos tratamentos A e B foram obtidos através da simulacão da distribuição Normal e assim sendo, o valor obtido através do teste t-Student é exato, verifica-se desse modo a exatidão do nível crítico de significância obtido através da aproximação "Bootstrap", principalmente quando são usadas estatísticas estudentizadas.

4 Exemplo Ilustrativo

As características climáticas da região amazônica, associadas à baixa fertilidade dos solos, sugere que grande parte dessa região esteja vocacionada para as práticas de cultivos em sistemas consorciados com culturas de ciclo longo (perenes), numa tentativa de proporcionar um revestimento florístico mais próximo ao padrão natural da vegetação amazônica.

Diante da reconhecida vocação da região, o CPATU - Centro de Pesquisa Agroflorestal da Amazônia Oriental, unidade descentralizada da EMBRAPA, vem desenvolvendo desde 1977 pesquisas com sistemas de cultivos, envolvendo sistemas consorciados com culturas perenes tradicionais da região. O exemplo ilustrativo usado nessa dissertação foi cedido pelo CPATU e refere-se ao dados do projeto "Sistemas de produção com plantas perenes em consórcio", obtidos no experimento instalado no município de Capitão Poço - PA, localizado a $1^0 38'$ latitude sul e $47^0 01'$ longitude Oeste, que apresenta Latossolo Amarelo de textura Argilosa (EMBRAPA, 1993). As características climáticas da área experimental são apresentadas na Figura 14.

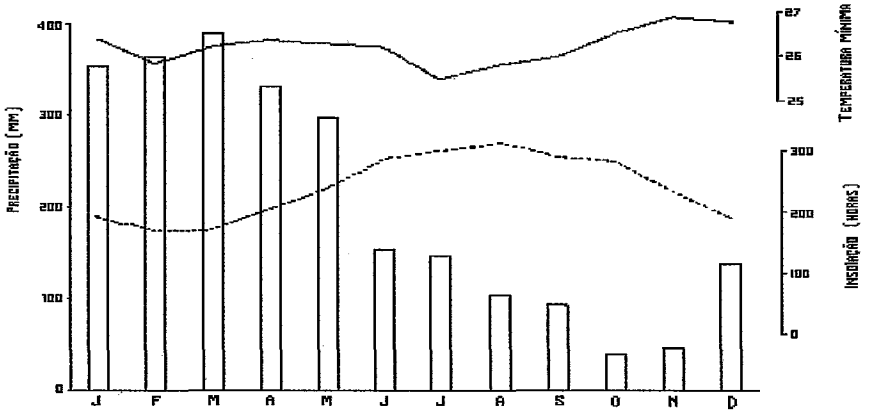

Figura 14: Condições climáticas do município de Capitão Poço-PA. (1981-1985)
Fonte: EMBRAPA, 1993

O consórcio 1 foi implantado com a cultura do cacau no espaçamento $2,5m$ X $2,5m$ e a cultura da pupunha, no espaçamento $10,0m$ x $10,0m$. No consórcio 2 a cultura do cacau foi plantada no espaçamento $2,5m$ x $2,5m$, enquanto a cultura da seringueira foi plantada no espaçamento $15,0m$ x $5,0m$. No ensaio de monocultivo as culturas do cacau, seringueira e pupunha foram plantadas, respectivamente, nos espaçamentos $2,5m$ x $2,5m$; $7,5m$ x $2,5m$ e $10,0m$ x $10,0m$.

Os dados apresentados na Tabela 13, referente ao consórcio 1, foram coletados nos anos de 1985 a 1989, enquanto os dados referentes ao consórcio 2, foram coletados nos anos de 1987 a 1991. Ambos os consórcios foram instalados em grandes quadras e, portanto não dispõem de delineamento experimental. Os dados da cultura do cacau, referem-se à produção de amêndoa seca em kg/ha; os dados de seringueira referem-se à produção de cernambi (borracha seca) em kg/ha e os dados da pupunha referem-se à produção de frutos em kg/ha.

Tabela 13: Dados de produção em kg/ha das culturas de Cacau, Seringueira e Pupunha em sistemas consorciados e de monocultivos

Anos	**Consórcio 1**		**Consórcio 2**		**Monocultivo**		
	Cacau[a]	Pupun.[a]	Cacau[a]	Serin.[a]	Cacau[b]	Sering.[a]	Pupun.
1985	1048	8400	-	-	-	-	-
1986	1788	6920	-	-	-	-	-
1987	1308	7060	1648	350	1216	505	4000
1988	1559	5809	1185	321	805	510	3193
1989	1691	6400	1819	237	1976	423	4986
1990	-	-	914	300	1034	420	–
1991	-	-	1524	414	1697	427	–

FONTES: [a] SERRÃO *et al*, 1993 e [b]EMBRAPA, 1993

Visando obter um estudo comparativo das produções obtidas nos sistemas consorciados em relação às produções obtidas nos monocultivos, a análise estatística do referido experimento foi realizada sobre a variável aleatória LER (*Land Equivalent Ratio*), definida no início deste trabalho através da Expressão (1), que representa uma medida de eficiência do uso da terra, com bastante apelo intuitivo e largamente empregada na literatura.

$$LER = \sum_{i=1}^{m} \frac{\bar{y}_i}{\bar{z}_i}$$

onde:

$\bar{y}_i$ é a produção media observada do i-ésimo componente, quando cultivado de forma consorciada, em uma determinada unidade de área;

$\bar{z}_i$ é a produção média do mesmo i-ésimo componente na mesma unidade de área, quando cultivado em monocultivo.

Logo,

$$LER_{\mathbf{C_1}_{(1985)}} = \frac{1048}{1216} + \frac{8400}{4000} = 2,962$$

$$LER_{\mathbf{C_1}_{(1986)}} = \frac{1788}{1976} + \frac{6920}{4986} = 2,293$$

$$LER_{\mathbf{C_1}_{(1987)}} = \frac{1308}{1216} + \frac{7060}{4000} = 2,841$$

$$LER_{\mathbf{C_1}_{(1988)}} = \frac{1559}{805} + \frac{5809}{3192} = 3,756$$

$$LER_{\mathbf{C_1}_{(1989)}} = \frac{1691}{1976} + \frac{6400}{4986} = 2,139$$

$$LER_{\mathbf{C_2}_{(1987)}} = \frac{1648}{1216} + \frac{350}{505} = 2,048$$

$$LER_{\mathbf{C_2}_{(1988)}} = \frac{1185}{805} + \frac{321}{510} = 2,101$$

$$LER_{\mathbf{C_2}_{(1989)}} = \frac{1819}{1976} + \frac{237}{423} = 1,481$$

$$LER_{\mathbf{C_2}_{(1990)}} = \frac{914}{1034} + \frac{300}{420} = 1,598$$

$$LER_{\mathbf{C_2}_{(1991)}} = \frac{1524}{1697} + \frac{414}{427} = 1,868$$

Tabela 14: Valores do LER para o consórcio 1 e o consórcio 2, obtidos a partir dos dados do exemplo ilustrativo.

Consórcio 1		**Consórcio 2**	
LER	2,962	LER	2,048
LER	2,293	LER	2,101
LER	2,841	LER	1,481
LER	3,756	LER	1,598
LER	2,139	LER	1,868
$\overline{LER}_{\mathbf{C_1}}$	2,79820	$\overline{LER}_{\mathbf{C_2}}$	1,81920
s^2_{C1}	0,40895	s^2_{C2}	0,07436

Como não há dados das produções dos monocultivos nos anos de 1985 e 1986, os denominadores dos $LER_{\mathbf{C_1}_{(1985)}}$ e $LER_{\mathbf{C_1}_{(1986)}}$, foram escolhidos aleatoriamente entre os valores existentes. E, com base nos valores apresentados na Tabela 14, pode-se verificar que ambos os sistemas consorciados superaram, em eficiência do uso da terra, os referidos monocultivos, pois, tanto o $\overline{LER}_{\mathbf{C_1}}$ como o $\overline{LER}_{\mathbf{C_2}}$ apresentaram valores superiores a 1.

Como foi discutido nas secções anteriores, os valores obtidos dos $\overline{LER}$ tratam-se de estimativas pontuais dos referidos parâmetros populacionais e portanto é necessário obterem-se estimativas por intervalos e aplicar testes estatísticos para a tomada de decisão a respeito de qual dos sistemas consorciados é o melhor.

A inferência sobre os parâmetros será feita, inicialmente, com o uso dos métodos estatísticos convencionais, neste caso, aplicar-se-á a distribuição t-Student.

$$T = \frac{\hat{\theta} - \theta}{\sqrt{s^2_{\hat{\theta}}}} \quad \sim \quad t_{n-1} \tag{27}$$

Considerando a suposição dada em (27) sobre a distribuição da estatística T, o Intervalo de Confiança t-Student, com a probabilidade de cobertura (coeficiente de confiança) de $(1-2\alpha)$ é dado por:

$$\left[\hat{\theta} - t_{[n-1\ ;\ (1-\alpha)]}\left(\sqrt{\frac{s^2}{n}}\right) ; \hat{\theta} - t_{[n-1\ ;\ (\alpha)]}\left(\sqrt{\frac{s^2}{n}}\right)\right]$$

Como a distribuição t-Student é simétrica, então

$$\text{IC para } \theta : \left[\hat{\theta} \pm t_{[n-1\ ;\ (1-\alpha)]}\left(\sqrt{\frac{s^2}{n}}\right)\right]$$

Para os dados da Tabela (14), os IC para as médias consórcio, com um coeficiente de confiança de $1-2\alpha$=95%, ou seja, α=0,025, são dados por:

$$\text{IC para } LER_{\mathbf{C_1}} : \left[\overline{LER}_{\mathbf{C_1}} \pm t_{[4\ ;\ (0{,}975)]}\left(\sqrt{\frac{0{,}40895}{5}}\right)\right] = [2{,}004\ \ ;\ \ 3{,}592]$$

$$\text{IC para } LER_{\mathbf{C_2}} : \left[\overline{LER}_{\mathbf{C_2}} \pm t_{[4\ ;\ (0{,}975)]}\left(\sqrt{\frac{0{,}07436}{5}}\right)\right] = [1{,}481\ \ ;\ \ 2{,}158]$$

O teste de hipótese, onde será testado $H_0 : LER_{\mathbf{C_1}} = LER_{\mathbf{C_2}}$ contra $H_a : LER_{\mathbf{C_1}} > LER_{\mathbf{C_2}}$, será obtido como se segue:

$$NC = Prob_{H_0}\left\{t_{[\ n_1+n_2-2\ ]} > \left(\frac{\overline{LER}_{\mathbf{C_1}} - \overline{LER}_{\mathbf{C_2}}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)\right\}$$

$$NC = Prob_{H_0}\left\{t_{[\ 8\ ]} > \left(\frac{2{,}7982 - 1{,}8192}{0{,}49159\sqrt{\frac{1}{5}+\frac{1}{5}}}\right)\right\} = 0{,}0068 \qquad (28)$$

Portanto, com base no Nível Crítico de significância (NC), conclui-se que o consórcio 1 é superior ao consórcio 2, com relação ao uso eficiente da terra (LER) com (*"P-value"*=0,0068), ou seja, a diferença entre os dois consórcios é altamente significativa. Encontram-se, respectivamente, nos Apêndices 7.1 e 7.8, os programas gerados no SAS para obtenção de intervalos de confiança e testes de hipóteses t-Student.

Todavia, como foi discutido na Secção (**2.3.2**), o LER representa uma variável aleatória, formada pela razão de variáveis aleatórias e sua distribuição de probabilidades não segue a distribuição normal. Além disso o LER pode apresentar valores correlacionados com os respectivos monocultivos e no caso específico deste exemplo ilustrativo, pode haver correlação com os anos de coleta das observações. Dessa forma, não se deve aplicar os métodos estatísticos convencionais, para se obterem intervalos de confiança e testes de hipóteses, quando a variável é um LER, obtido pela razão de duas variáveis aleatórias.

Visando minimizar os problemas que o LER apresenta sobre as correlações e não normalidade, MEAD & RILEY (1981), propõem outras formas de obtenção do LER. As alternativas são feitas a partir do uso de outros valores nos denominadores ao invés dos valores obtidos nas parcelas de monocultivos, de maneira que o denominador não seja mais uma variável aleatória e assim o LER passa a ter distribuição aproximadamente normal. O problema é que os melhores resultados, segundo esses autores, ocorrem quando são usados nos denominadores valores externos ao experimento, como a produção média da região ou valores ótimos pré-fixados. Todavia, nos ensaios de agroflorestas, esse procedimento pode promover uma descaracterização das informações intrínsecas ao experimento, visto que, esses ensaios envolvem culturas perenes que poderão apresentar possíveis correlações com o fator tempo.

Uma solução para os problemos apresentados pela variável aleatória LER, pode ser obtida com o uso dos métodos computacionalmente intensivos, como o método "Bootstrap".

Aplicando-se o Algoritmo 1, descrito na Figura 6 com 100 amostras "Bootstrap" (**B**=100), obteveram-se as seguintes estimativas "Bootstrap" do erro padrão das médias, $\hat{\sigma}_{\mathbf{B}}(\overline{LER}_{\mathbf{C_1}}) = 0,255798$ e $\hat{\sigma}_{\mathbf{B}}(\overline{LER}_{\mathbf{C_2}}) = 0,109078$, as quais serão usadas para obtenção dos intervalos de confiança Bootstrap-Padrão e t-SBoot.

Na Tabela 15 encontram-se os valores dos intervalos de confiança "Bootstrap" para os $\overline{LER}$ com uma probabilidade de cobertura de 95%, obtidos com 1000 amostras "Bootstrap". Os dois primeiros (Bootstrap-Padrão e t-SBoot) usam, respectivamente, os percentis da distribuições Normal Padrão e t-Student, enquanto os três últimos (t-Bootstrap, ABC e BCa) utilizam os percentis da própria distribuição "Bootstrap" dos dados. Encontram-se nos Apêndices 4, 5, e 6 os programas desenvolvidos no S-PLUS para obtenção dos referidos Intervalos de Confiança.

Tabela 15: Intervalos de Confiança "Bootstrap" para os $\overline{LER}$, com um coeficiente de confiança de 95%, obtidos com $\mathbf{B} = 1000$ amostras "Bootstrap"

Estatística	I. Confiança	$\hat{\theta}$ (0,025)	$\hat{\theta}$ (0,975)
	Bootstrap-Padrão	2,297	3,299
	t-SBoot	2,088	3,508
$\overline{LER}_{\mathbf{C_1}}$	t-Bootstrap	2,082	3,902
	ABC	2,370	3,395
	BCa	2,341	3,305
	Bootstrap-Padrão	1,605	2,033
	t-SBoot	1,516	2,122
$\overline{LER}_{\mathbf{C_2}}$	t-Bootstrap	1,054	2,122
	ABC	1,590	2,019
	BCa	1,605	2,033

O teste de hipótese "Bootstrap" foi obtido a partir da aplicação do Algoritmo 5, descrito na Figura 13, de forma estudentizada. Neste teste, foi testada a hipótese $H_0 : LER_{\mathbf{C_1}} = LER_{\mathbf{C_2}}$ contra $H_a : LER_{\mathbf{C_1}} > LER_{\mathbf{C_2}}$.

A implementação computacional do Algoritmo 5 foi realizada através do programa desenvolvido no *Software* S-PLUS (o programa encontra-se no Apêndice 7.7). Executando o referido programa, com **B**=1000 amostras "Bootstrap", obtiveram-se 7 valores em que $t(x^*) > t(x)$. Logo,

$$\widehat{NC}_{Boot} = \frac{7}{1000} = 0,007$$

Portanto, com base no Nível Crítico de significância "Bootstrap" (NC_{Boot}), pode-se concluir que o consórcio 1 é superior ao consórcio 2, com relação ao uso eficiente da terra (LER) com *"P-value"*=0,007, ou seja, a diferença entre os dois consórcios é altamente significativa.

Este teste de hipótese "Bootstrap", também pode ser facilmente implementado através do procedimento MULTTEST do SAS (o referido programa encontra-se no Apêndice 7.8). Executado o procedimento MULTTEST do SAS para os dados da Tabela 14, com 1000 amostras "Bootstrap" (**B**=1000), obtiveram-se os resultados que se encontram abaixo.

```
Copyright(c) 1989 by SAS Institute Inc., Cary, NC USA.

                    MULTTEST PROCEDURE

Test for continuous variables:          T-test of mean

Tails for continuous tests:             Upper-tailed

Strata adjustment?                      No

P-value adjustments:                    Bootstrap

Center continuous variables?            No

Number of resamples:                    1000

Seed:                                   12357
```

MULTTEST COEFFICIENTS

	Class	
Test	1	2
Y(1)	1	-1

MULTTEST TABLES

		Class	
Variable	Statistic	1	2
LER	Mean	2.7982	1.8192
	Std Dev	0.6395	0.2727
	N	5.0000	5.0000

Test	Raw_p	Adj_p
Y(1)	0.0068	0.008

Existem várias opções para implementação do PROC MULTTEST do SAS. Para obter o citado teste de hipótese "Bootstrap", através do Procedimento MULTTEST, com o objetivo de se testar as hipóteses $H_0 : LER_{\mathbf{C_1}} = LER_{\mathbf{C_2}}$ e $H_a : LER_{\mathbf{C_1}} > LER_{\mathbf{C_2}}$, determinamos o contraste `Y(1)=1 -1` e solicitamos que fosse utilizado apenas a cauda superior da distribuição (isso foi feito através da opção `uppertailed`). A opção `nocenter` foi feita para que o programa não centralizasse, pela média, os valores reamostrados.

Na última linha da saída acima, pode-se verificar que o SAS fornece o nível descritivo do teste t-Student, *"P-value"*=0,0068, (valor igualmente encontrado através da Expressão (28)) e o nivel descritivo ajustado pelo método "Bootstrap" (*"P-value"*=0,008). Esse valor foi obtido a partir de 1000 amostras "Bootstrap" (**B**=1000) e é muito próximo ao valor encontrado através do programa feito para o S-PLUS, que forneceu (*"P-value"*=0,007). A diferença entre os dois valores deve-se ao fato de que os mesmos são obtidos por reamostragem aleatória e consequentemente os valores reamostrados para cada *Software* são diferentes.

5 Conclusões

De acordo com a metodologia empregada e com base nos resultados obtidos neste trabalho pode-se chegar as seguintes conclusões:

- A aplicação dos procedimentos "Bootstrap", para obtenção de estimativas de parâmetros, quando as amostras são pequenas e advém de uma distribuição desconhecida ou complexa, oferecem resultados confiáveis. Nos dados da Tabela 12, onde os valores dos parâmetros populacionais são conhecidos, pode-se verificar que o método "Bootstrap" fornece estimativas bastante acuradas.

- A análise estatística de experimentos agroflorestais, através do LER (*Land Equivalent Ratio*), cuja distribuição não é facilmente conhecida, pode ser implementada a partir de métodos "Bootstrap".

- Para o caso específico dos sistemas agroflorestais, o LER não parece retratar o principal atributo considerado em agroflorestas, que é a sustentabilidade do sistema. Por isso faz-se necessário o estudo de outros índices que sejam mais adequados à análise estatística dos sistemas agroflorestais.

6 Referências Bibliográficas

BARNETT, V. & RILEY, J. Statistics for environmental change. **Experimental Agriculture**, London, **31**:131-149. 1995.

BOLFARINE, H. & BUSSAB, W.O.. **Elementos de Amostragem**. 11º SINAPE (Simpósio Nacional de Probabilidade e Estatística). Belo Horizonte-MG. 1994.

COCHRAN,W.G.. **Sampling techniques**. 3rd ed. John Wiley & Sons. New York. 1977.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas** Piracicaba-SP, ESALQ/USP, 187 p. (Tese de Doutorado) 1990.

DEAR, K.B.G. END MEAD, R. The use of bivariate analysis technics for the presentation, a analysis and interpretation of data. In **Statistics in Intercropping Technical Research** no 1. 18 p., 1983.

DIACONIS, P & EFRON, B.. Computer Intensive Methods in Statistics. **Scientific American 248**116-130, 1983

EFRON, B. Bootstrap methods: Another look at the Jackknife. **The Annals of Statistics 1**:1-20. 1979.

EFRON, B. & TIBSHIRANI, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). **Statistical Science. 1**:54-57, 1986

EFRON, B. & TIBSHIRANI, R. **An introduction to the Bootstrap**. New York, London, Chapman & Hall. 1993.

FEDERER, W. T. **Statistical Design and Analysis for Intercropping Experiment**. New York, Springer-Verlag. 1993. Volume I: Two Crops, 300p.

GOMES, F.P. **A estatística moderna na pesquisa agropecuária** POTAFOS, Piracicaba-SP. 1984.

GONÇALVES, S.R. **Consorciação de Culturas** - Técnicas de análise e estudo da distribuição do LER. Brasília, UNB, 1982. 218 p. (Dissertação de Mestrado).

GREGOIRE, T. G.. The jackknife: an introduction with aplications in forestry data analysis. **J. for. Res., 14**: 493-497. 1984.

KASS, D.L. **Polyculture cropping systems**: review and analysis. Ithaca, New York State College of Agriculture and Life Sciencies, 1978. 69 p.

HINKLEY, D.V.. Jacknife in unbalanced situations. **Technometrics, 19**: 285-292. 1977.

HINKLEY, D.V.. Bootstrap Methods **The Jounal of the Royal Statistical Society B**, London, **50**:321-337, 1988.

MADGWICK, H.A.I.. Estimating the above-ground weight of forest plot using the basal area ratio method. **N. Z. J. for. Sci., 11**: 278-286. 1981.

MEAD, R. END RILEY, J. A review of statistical ideas relevant to intercropping research. **The Jounal of the Royal Statistical Society A**, London, **144**:462-509, 1981.

MILLER, R.G.. The jackknife-a review. **Biometrika, 61**: 1-15. 1974.

NAIR, P.K.R. Classification of agroforetry systems. **Agroforestry Systems. 3**:97-128. 1985.

NAIR, P.K.R. **The prospects for Agroforestry in the tropics** Washington, U.S.A., 1990. 79 p. (World Bank technical paper, no 131) 79 p.

OLIVEIRA, E.B. Considerações sobre análise estatística na pesquisa de sistemas agroflorestais. **Anais, I Congresso Brasileiro sobre Sistemas Agroflorestais e I encontro sobre Sistemas Agroflorestais nos Países do Mercosul**, Porto Velho , RO, 1994.(CNPF-EMBRAPA - Curitiba, PR, v.1).

OYEJOLA, B.A.END MEAD, R. Statistical assessment of different ways of calculating land equivalent ratios (LER). **Experimental Agriculture**. London. **18**:125-138, 1982.

RAMALHO, M.A.P.; OLIVEIRA, A.C.; GARCIA, J.C. **Recomendações para o planejamento e análise de experimentos com as culturas de milho e feijão consorciadas**. Sete Lagoas, MG. CNPMSEMPRAPA. 1983. 74p (documento n. 2).

SERRÃO, E.A; TEIXEIRA, L.B.; OLIVEIRA, R.F.; BASTOS, J.B. Soil alterations in perennial pasture and agroforestry systems in the Brazilian Amazon. In:**Workshp on long-term soil management experiments in the tropics**, Columbus-Ohio, USA, 1993. *Proceedings*. Ohio State University, 1993.

SAS TECHNICAL REPORT 229. **SAS/STAT Software: Changes and Enhancements**. Release 6.07. 1993.

S-PLUS for Windows, User's Manual. Washington: StatSci, 1993. 1v. 420p.

S-PLUS for Windows, User's Manual. Washington: StatSci, 1993. 2v. 448p.

SILVA, D. N. O Método Bootstrap e Aplicações à Regressão Multipla. Campinas-SP, 1995. 160p. Tese (Mestrado) - IMECC/UNICAMP.

TRINCA, L. A. Métodos Computacionalmente Intensivos na Estimação do Número de Espécies. Campinas-SP, 1988. 90p. Tese (Mestrado) IMECC/UNICAMP.

VEIGA, J.B. & SERRÃO, E. A. S. Sistemas silvopastoris e produção animal nos trópicos úmidos: a experiência da Amazônia Brasileira. In: PASTAGENS. Sociedade Brasileira de Zootecnia. Piracicaba - SP. 1990. p. 37-68.

WIJESINHA, A.; FEDERER, W.T.; CARVALHO, J.R.P.; FORTES,T.A. Some statistical analysis for a maize and beans intercropping experiments **Crop Science**, **22**:660-8, 1982.

WILLEY, R.W. END OSIRU, D.J.O. Studies on mixtures of maize and beans (*Phaseolus vulgaris*) with particular reference to plant population. **Journal of Agricultural Science**, New York, **79**:519-29, 1972.

7 Apêndice

7.1 Programa SAS para obtenção do Intervalo de Confiança

```
data a; input trat ler;
cards;...
;
proc means data=a n mean var;
   var ler;
   class trat;
   output out=b n=n mean=mean  var=var;
proc print data=b;
     title 'Intervalos de confiança para diferença de média';
run;
data confia;
   set b;              where _type_=1;
   n=lag(n);           m=n;
   varia1=lag(var);    varia2=var;
   alpha=0.05;         gl1=n-1;
   t1=tinv(1-alpha/2,gl1);
   s1=sqrt(varia1/n);
   media1=lag(mean);
   limedia1=media1-t1*s1;
   lsmedia1=media1+t1*s1;
   gl2=m-1;
   t2=tinv(1-alpha/2,gl2);
   s2=sqrt(varia2/m);
   media2=mean;
   limedia2=media2-t2*s2;
   lsmedia2=media2+t2*s2;
   drop trat n mean var m gl1 gl2 varia1 varia2  _type_ _freq_;
run;
proc print data=confia(firstobs=2) label noobs;
   label alpha='Nivel de significancia'  t1='Valor critico do t para LER1'
         s1='Erro padrão LER1'             limedia1='Limite inferior LER1'
         media1='Media do LER1'            lsmedia1='Limite superior LER1'
         t2='Valor critico de t para LER2' s2='Erro padrão LER2'
         limedia2='Limite inferior LER2'   media2='Media do LER2'
         lsmedia2='Limite superior LER2';
run;
```

7.2 Programa SAS para obtenção do Teste de hipótese t de Student

```
data a;
   input trat ler;
cards; ...
;
proc means data=a n mean var;
   var ler;
   class trat;
   output out=b n=n mean=mean var=var;
   title 'Examinando o proc mean output';
proc print data=b;
     title 'Limites de confiança para diferença de média';
run;
data p_valor;
   set b;
   where _type_=1;
   nx1=lag(n);
   nx2=n;
   x1=lag(mean);
   x2=mean;
   var1=lag(var);
   var2=var;
   alpha=0.025;
   gl=nx1+nx2-2;
   dif=x1-x2;
   sconj=sqrt(((nx1-1)*var1+(nx2-1)*var2)/gl);
   epadrao=sconj*sqrt((1/nx1+1/nx2));
   tobs=dif/epadrao;
   t=probt(tobs,gl);
   NC=1-t;
   drop trat n mean var nx1 nx2 x1 x2 var1 var2 gl t _type_ _freq_;
   run;
proc print data=p_valor(firstobs=2) label noobs;
   label alpha='Nivel de significancia'
         dif='Diferença de Medias'
         sconj='Desvio padrão conjunto'
         epadrao='Erro padrão da diferença'
         tobs='t-obs'
         nc='P-Value';
  run;
```

7.3 Programa S-PLUS para obtenção da Estimativa "Bootstrap" do Erro Padrão de Um Estimador

Primeiro os dados são definido como um objeto e depois define-se uma função para o estimador. Como neste caso o estimador é a media amostral ($\hat{\theta} = \bar{x}$), temos:

```
ler1<- c(2.962, 2.293, 2.841, 3.756, 2.139)
theta<- function(x)
        {
        mean(x)
        }
```

A função que fornece a estimativa de variância Bootstrap (σ_B) é:

```
varia.boot <- function(x, nboot,theta)
      {
        call <- match.call()
        abootx <- matrix(sample(x, size = length(x) *
                  nboot, replace = T), nrow = nboot)
        mediax <- apply(abootx, 1, theta)
        sbootx <- sqrt(var(mediax))
        return(sbootx)
      }
```

Colocando-se essas funções no diretorio de trabalho do S-PLUS e Executando-se a função `varia.boot(ler1,100,theta)` obtém-se uma estimativa "Bootstrap" do erro padrão da média para os dados `ler1` com 100 amostras "Bootstrap".

7.4 Programa S-PLUS para obtenção Intervalo de Confiança Não-paramétrico t-Bootstrap

Funções obtidas no Carnegie-Mellon University através da INTERNET, no endereço www.stat.cmu.edu

Primeiro define-se os objetos e funções abaixo,

```
ler1<- c(2.962, 2.293, 2.841, 3.756, 2.139)
theta <- function(x){mean(x)}
nboot <- 1000
```

A função que fornece os Intervalos de Confiança t-Bootstrap é:

```
boott <- function(x, theta, ..., sdfun = sdfunboot, nbootsd = 25,
                  nboott=200, VS = F, v.nbootg = 100, v.nbootsd= 25,
                  v.nboott = 200,  perc = c(0.001, 0.01,0.025, 0.05,
                                0.1, 0.5, 0.9, 0.95, 0.975,0.99,0.999), ...)
{
        call <- match.call()
        sdfunboot <- function(x, nboot, theta, ...)
        {
                n <- length(x)
                junk <- matrix(sample(x, size = n * nboot, replace = T),
                                nrow =nboot)
                return(sqrt(var(apply(junk, 1, theta,...))))
        }
        thetahat <- theta(x, ...)
        n <- length(x)
        if(!VS) { sd <- sdfun(x, nbootsd, theta, ...)
        }
        else { sd <- 1
        }
        if(VS){ xstar <- matrix(sample(x, size = n * v.nbootg,
                                replace = T), nrow = v.nbootg)
               thetastar0 <- apply(xstar, 1, theta,...)
               sdstar0 <- apply(xstar, 1, sdfun, v.nbootsd, theta, ...)
               o <- order(thetastar0)
               thetastar0 <- thetastar0[o]
               sdstar0 <- sdstar0[o]
```

```
            temp <- lowess(thetastar0, log(sdstar0))$y
            sdstar0 <- exp(temp)
            invsdstar0 <- 1/sdstar0
            g <- ctsub(thetastar0, invsdstar0, thetastar0)
            g <- (g - mean(g))/sqrt(var(g))
            g <- g * sqrt(var(thetastar0)) + mean(thetastar0)
        }
        if(!VS) { thetastar0 <- NULL   g <- NULL }
        if(!VS) { xstar <- matrix(sample(x, n * nboott, replace = T),
                                  nrow = nboott)
        }
        else { xstar <- matrix(sample(x, n * v.nboott, replace = T),
                               nrow = v.nboott)
        }
        thetastar <- apply(xstar, 1, theta, ...)
        gthetastar <- rep(0, length(thetastar))
        if(VS) { gthetahat <- yinter(thetastar0, g, thetahat)
        }
        else { gthetahat <- thetahat
        }
        if(VS) { for(i in 1:length(thetastar)) {
                 gthetastar[i] <- yinter(thetastar0, g,thetastar[i])
                 }
        }
        else { gthetastar <- thetastar
        }
        if(!VS) { sdstar <- apply(xstar, 1, sdfun,nbootsd, theta, ...)
        }
        else { sdstar <- 1
        }
        tstar <- sort((gthetastar - gthetahat)/sdstar)[
                 length(gthetastar):1]
        ans <- gthetahat - sd * tstar
        if(VS) { for(i in 1:length(ans)) {
                 ans[i] <- xinter(thetastar0, g, ans[i])
                }
        }
        o <- trunc(length(ans) * perc) + 1
        ans1 <- matrix(ans[o], nrow = 1)
        dimnames(ans1) <- list(NULL, perc)

 return(confpoints = ans1, theta = thetastar0, g, call)
}
```

Executando-se `boott(ler1, theta)` obtém-se a saída

7.5 Programa S-PLUS para obtenção Intervalo de Confiança Não-paramétrico ABC

Funções obtidas no Carnegie-Mellon University através da INTERNET, no endereço www.stat.cmu.edu

Definem-se o objeto para os dados, a função do parâmetro $\hat{\theta}$ e o número de amostras "Bootstrap".

```
ler1<- c(2.962, 2.293, 2.841, 3.756, 2.139)
tt <- function(p,x)
       {
       sum(p*x)/sum(x)
       }
nboot <- 1000
```

A função que fornece os Intervalos de Confiança ABC é:

```
abcnon <- function(x, theta,  epsilon = 0.001,
                  alpha =c(.025,.05,.1,.16,.84,.9,.95,.975))
{
 call <- match.call()
#abc confidence intervals for nonparametric problems
#theta(P ,x) is statistic in resampling form, where P[i] is weight on x[i]
       if(is.matrix(x)) {n <- nrow(x)} else {n <- length(x)}
       ep <- epsilon/n; I<- diag(n); P0<- rep(1/n,n)
       t0 <- tt(P0,x)
#calculate t. and t.....................................................
       t. <- t.. <- numeric(n)
       for(i in 1:n) {  di <- I[i,  ] - P0
                        tp <- tt(P0 + ep * di,x)
                        tm <- tt(P0 - ep * di,x)
                        t.[i] <- (tp - tm)/(2 * ep)
                        t..[i] <- (tp - 2 * t0 + tm)/ep^2}

#calculate sighat,a,z0,and cq ..........................................
       sighat <- sqrt(sum(t.^2))/n
       a <- (sum(t.^3))/(6 * n^3 * sighat^3)
```

```
        delta <- t./(n^2 * sighat)
        cq <- (tt(P0+ep*delta,x) -2*t0 + tt(P0-ep*delta,x))/(2*sighat*ep^2)
        bhat <- sum(t..)/(2 * n^2)
        curv <- bhat/sighat - cq
        z0 <- qnorm(2 * pnorm(a) * pnorm( - curv))

#calculate interval endpoints.....................................
        Z <- z0 + qnorm(alpha)
        za <- Z/(1 - a * Z)^2
        stan <- t0 + sighat * qnorm(alpha)
        abc <- seq(alpha)
        pp_matrix(0,nrow=n,ncol=length(alpha))
            for(i in seq(alpha)) {abc[i] <- tt(P0 + za[i] * delta,x)
                      pp[,i]_P0 + za[i] * delta }
        limits <- cbind(alpha, abc, stan)
            dimnames(limits)[[2]]_c("alpha", "abc", "stan")
#output in list form.............................................


 return(limits=limits, stats=list(t0=t0,sighat=sighat,bhat=bhat),
         constants=list(a=a,z0=z0,cq=cq), tt.inf=t., pp=pp, call=call)
}
```

Executando-se `abcnon(ler1, tt)` obtém-se os Intervalos de confiança ABC e Bootstrap-Padrão para os dados do `ler1`

7.6 Programa S-PLUS para obtenção Intervalo de Confiança Não-paramétrico BCa

Funções obtidas no Carnegie-Mellon University através da INTERNET, no endereço www.stat.cmu.edu

Primeiro definem-se os objetos e as função abaixo

```
ler1<- c(2.962, 2.293, 2.841, 3.756, 2.139)
theta <- function(x){mean(x)}
perc95 <- function(x){quantile(x, .95)}
nboot <- 1000
```

A função que fornece os Intervalos de Confiança BCa é:

```
bcanon <- function(x, nboot, theta, ...,
            alpha = c(0.025, 0.05, 0.1, 0.16, 0.84, 0.9, 0.95, 0.975))
{
       call <- match.call()
       n <- length(x)
       thetahat <- theta(x, ...)
       bootsam <- matrix(sample(x, size = n * nboot, replace = T),
                         nrow = nboot)
       thetastar <- apply(bootsam, 1, theta, ...)
       z0 <- qnorm(sum(thetastar < thetahat)/nboot)
       u <- rep(0, n)
            for(i in 1:n) {u[i] <- theta(x[ - i], ...)}
       uu <- mean(u) - u
       acc <- sum(uu * uu * uu)/(6 * (sum(uu * uu))^1.5)
       zalpha <- qnorm(alpha)
       tt <- pnorm(z0 + (z0 + zalpha)/(1 - acc * (z0 + zalpha)))
       ooo <- trunc(tt * nboot)
       confpoints <- sort(thetastar)[ooo]
       confpoints <- cbind(alpha, confpoints)
       dimnames(confpoints)[[2]] <- c("alpha", "bca point")

   return(confpoints, z0, acc, u, call)
}
```

Executando-se `bcanon(ler1, nboot, theta)` obtém-se a saída

7.7 Programa S-PLUS para obtenção do Teste de Hipótese "Bootstrap"

Definem-se os objetos abaixo

```
ler1<- c(2.962, 2.293, 2.841, 3.756, 2.139)
ler2<- c(2.048, 2.101, 1.481, 1.598, 1.868)
nboot <- 1000
```

A função que fornece os O Teste de Hipótese Bootstrap é:

```
testhipot<-function(y, z, nboot)
{
        x <- c(y, z)
        n <- length(y)
        m <- length(z)
        vary <- var(y)
        varz <- var(z)
        dif <- mean(y) - mean(z)
        sconj <- sqrt(((n - 1) * vary + (m - 1) * varz)/
                 (n + m - 2))
        tobs <- (dif)/(sconj * sqrt((1/n) + (1/m)))
        booty <- matrix(sample(x, size = length(y) *
                 nboot, replace = T), nrow = nboot)
        bootz <- matrix(sample(x, size = length(z) *
                 nboot, replace = T), nrow = nboot)
        varbooty <- apply(booty, 1, var)
        varbootz <- apply(bootz, 1, var)
        sbootconj <- sqrt(((n - 1) * varbooty + (m - 1) *
                     varbootz)/(n + m - 2))
        difboot <- (apply(booty, 1, mean) - apply(bootz,1, mean))
        tboot <- (difboot)/(sbootconj * sqrt((1/n) + (1/m)))
        nc <- tboot[tboot >= tobs]
        ncboot <- length(nc)/nboot
        return(ncboot)
}
```

Executando-se a função `testhipot(ler1,ler2,1000)`

obtém-se o nível descritivo Bootstrap.

7.8 Programa SAS, através do PROC MULTTEST para obtenção do Teste de Hipótese "Bootstrap"

```
options ps=60 ls=64 nodate;
data a;
   input trat ler;
cards;
1 2.962
1 2.293
1 2.841
1 3.756
1 2.139
2 2.048
2 2.101
2 1.481
2 1.598
2 1.868
;


proc multtest data=a outsamp=resposta bootstrap nocenter
                    seed=12357 order=data nsample=1000;
   test mean(ler / uppertailed);
   class trat;
   contrast 'Y(1)' 1 -1;
 run;
```